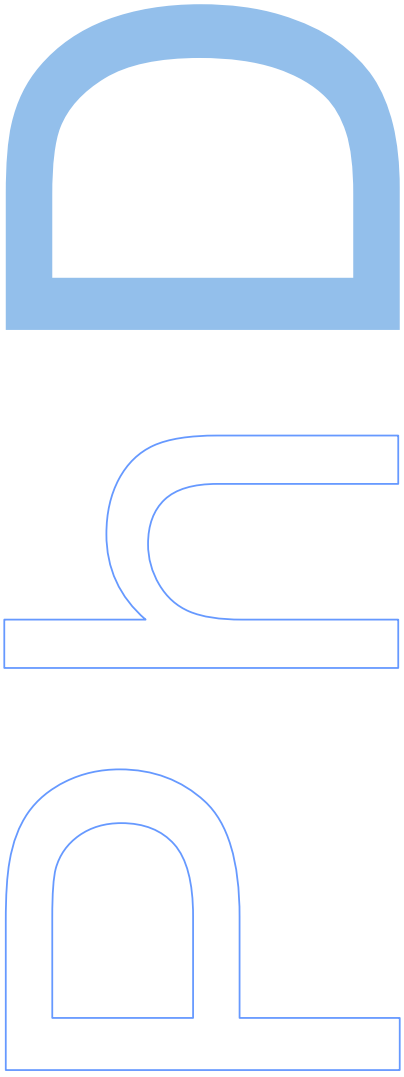
U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Prediction and Ranking of Highly Popular Web Content

Nuno Miguel Pereira Moniz
Programa Doutoral em Ciência de Computadores
Departamento de Ciência de Computadores
2017

Orientador
Luís Fernando Rainho Alves Torgo, Professor Associado,
Faculdade de Ciências da Universidade do Porto



Dedicated to
My parents, Mário and Maria.

Acknowledgments

This thesis is the results of four years of work for which many people contributed, directly and indirectly.

First and foremost, I would like to thank my supervisor Professor Luís Torgo for accompanying me in this journey and helping me throughout this process that now culminates with this thesis. Throughout this bumpy process, he was a key factor in keeping my motivation high and motivating me to go further and deeper into the areas that were studied, but also in providing important advices and warnings when concentration and focus were difficult to maintain. For that, he has my utmost respect and gratitude.

To all my colleagues in Room 1.70, António Gonçalves, Joana Côrte-Real, João Santos, Mariana Oliveira, Miguel Areias and Vítor Cerqueira, with whom I established good friendships, a big thank you, for enduring both the happy and the stressful times in the office.

Also, a special acknowledgment to Paula Branco and Rita Ribeiro for allowing me to "pick their brain" with multiple ideas, and their availability to discuss many of the doubts that I had throughout this process.

I would also like to express my gratitude to all my colleagues in LIAAD - INESC Tec, specially to Professor Alípio Jorge and Joana Dumas, and to Alexandra Ferreira.

A special word of appreciation to Allan Tucker and Stephen Swift from Brunel University (London, UK), and to Magdalini Eirinaki from San Jose State University (California, USA). By making it possible for me to spend time at your respective universities, you had a major impact on the work that I was able to do throughout my PhD. Also, a special word to Renato Araújo Soeiro, for the conversations and scientific discussions.

To Alex Gomes, Bárbara Silva, Hugo Monteiro, João Carlos, João Mineiro, João Paupério, Mariana Gomes, Pedro Rodrigues and Ricardo Sá Ferreira. For their patience and friendship; for providing many of the moments that kept me motivated, but also for providing much needed moments of distraction.

A special thank you to my family: my parents Mário and Maria, and my brother Bruno.

You are the reason I'm here today.

Finally, to Raquel. For always believing in me, and for providing all the affection and confidence that I needed.

The work presented in this thesis was funded by the Portuguese Science and Technology Foundation (FCT) with the PhD grant SFRH/BD/90180/2012, by the ERDF – European Regional Development Fund through the COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through FCT as part of project UID/EEA/50014/2013.

Abstract

This thesis addresses prediction and ranking tasks using web content data. The main objective is to improve the ability to accurately predict and rank recent and highly popular content, thus enabling a faster and more precise recommendation of such items. The main motivation relates to the profusion of online content, and the increasing demand of users concerning a fast and easy access to relevant content.

To fulfill these tasks an extensive review of previous work is carried out in order to define the state-of-the-art and to identify important research opportunities. As a result, three problems are identified and addressed in this thesis: *(1)* the lack of an interpretable and robust evaluation framework to correctly assess web content popularity prediction models focusing on highly popular web content; *(2)* issues concerning proposals of popularity prediction models and their ability to predict the rare cases of highly popular content; and *(3)* the need for recommendation frameworks concerning such items using multi-source data. For each of these problems novel solutions are proposed and extensively evaluated in comparison to existing work.

The first problem *(1)* concerns the evaluation methods commonly used in web content popularity prediction tasks. According to previous work, the popularity of web content is best described by a heavy-tail distribution. As such, at any given moment, most of the content under analysis has a low level of popularity, and a small set of cases has high levels of popularity. Standard evaluation metrics focus on the average behaviour of the data, assuming that each case is equally relevant. Given the predictive focus on highly popular content, it is argued that such assumption may lead to an over-estimation of the models' predictive accuracy. Therefore, an evaluation framework is proposed, allowing for a robust interpretation of the prediction models' ability to accurately forecast highly popular web content.

The second problem *(2)* is related to the fact that proposals concerning web content popularity prediction models are based on standard learning approaches. These are commonly biased towards capturing the dynamics of the majority of cases. Given the skewness of web content popularity data, this may lead to poor accuracy towards under-represented cases of highly

popular items. An evaluation with a diverse set of such proposals is carried out, confirming their issues when learning to predict such items. Also, it is additionally confirmed that the use of standard evaluation metrics often presents an over-estimated ability to accurately predict the most popular items. Novel approaches are proposed for the prediction of web content popularity focusing on accuracy towards highly popular items.

The third and final problem *(3)* concerns the task of ranking, but also evaluating, web content by its predicted popularity. Although the task of ranking may be trivial in most cases, when considering scenarios with multiple sources of data such task is considerably difficult. Notwithstanding, ranking tasks and their evaluation in single-source scenarios are not exempt of issues concerning the ability to account for highly popular content. The ability to rank web content based on models' predictions is discussed, given an extensive evaluation in both single-source and multi-source scenarios.

Each of these problems is evaluated using real-world data concerning online news feeds from both official and social media sources. This type of web content provides a difficult setting for the early and accurate prediction of highly popular items, given their short lifespan. Experimental evaluations show that the approaches proposed in this thesis concerning the prediction and ranking of highly popular content obtained encouraging results demonstrating a significant advantage in comparison to state-of-the-art work.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

The Internet has become an inescapable source of information for users, enabling access to a large and increasing amount of information. As information progressively shifts from a physical medium to online content, making sense of this information is crucial to provide the best set of information resources to users.

The increasing ability to access information was accompanied by the added speed and reach in information sharing. This development poses a significant impact in social, economic and political contexts [136], to which the advent of social media platforms played and continues to play a key role. The creation and proliferation of such platforms allow users to post information in real-time and interact with others. This provoked an increase of available information known as social media data, and a growing demand concerning the computational capability to analyse, interpret and act upon such information.

Consider the case of search engines: search engines essentially gather, process and store documents. Then, either factoring or not the influence of a given social media platform or other mediums, these systems provide a ranking of resources classified by itself as most relevant given a user query. These suggestions are commonly based on data that ranges from a given point in the past to the present. This opens the issue dealt in this thesis: with the profusion of online content and related feedback from users via social media platforms, an important task is to anticipate the relevance of very recent web content, for which user feedback is nonexistent or scarce.

This thesis addresses the problem of predicting and ranking highly popular web content. Our main problem is the prediction of numeric values of a continuous variable (*i.e.* popularity) in order to provide a quicker and more accurate ranking of relevant and recent web content.

In this first chapter the main problem is contextualized and defined, and the main motivation for this work and its contributions are described.

1.1 Scope

The evolution of data mining provided tools to understand the relation between variables, providing valuable insights in many domains spanning from finance to meteorology. However, data mining allows for more than modelling interplay between variables such as the prediction of future values in the same domain. Consider the example of the meteorology domain. Given a set of temperature observations, data mining enables the learning of models. By using this model and a set of future values' indicators, it is possible to attempt the prediction of future target values. Target values differ concerning the type of data mining technique and approach employed. For example, in classification tasks the target value is nominal; in regression and time series forecasting tasks the target value is numeric.

An important issue is that the prediction of each future value in a given domain may not have the same relevance for the users. In many cases the user is focused on predicting what is anomalous or rare instead of what is common or standard. This is a well known problem known as imbalanced data. Data imbalance is defined by the existence of an over-representation of a given class(es) or numeric value interval(s), over another. Adding to this, in many cases, the under-represented class or numeric value interval is the most relevant for the user and a wrongful prediction may be costly. The combination of these two factors (skewed distribution and focus on under-represented items) is the basis of imbalanced domain learning tasks. Solving such tasks is an interesting and open issue, considered to be one of the most important problems in machine learning and data mining [247].

1.2 Context and Problem Definition

This thesis addresses a specific type of data mining applications: tasks where the main objective is to maximize predictive accuracy concerning an interval of a given numeric target variable, using social media data. Additionally, a second objective is using the outcome of such prediction tasks to significantly improve the accuracy and timeliness of web content suggestions.

Social media data surged with the advent of social media platforms and the effort of some proprietaries to share some of its user-generated content. This type of data is considerably different from conventional types due to its distinct characteristics such as size, noise and bias, but also due to it being heterogeneous, multi-source, partial and asymmetrical [152]. These characteristics have challenged the data mining community over recent years and the contributions using such data in predictive tasks have significantly increased.

An important characteristic in many social media data sets is that the popularity of items is described by a heavy-tail distribution [216]. As such, we may infer that only a small set

of cases are in fact highly relevant to users and that most of the cases are non-relevant. This is the case for some of the most known types of social media data, such as YouTube video visualizations or users' feedback in social media platforms concerning online news. In predictive tasks using social media data from either of these examples, one major challenge is to accurately predict the popularity of highly relevant cases, so that they may be promptly suggested to users. On the other hand, failing to do so comes at a cost. This implies the existence of non-uniform preferences as to predictive accuracy concerning web content considered by users as non-relevant or highly relevant.

A great amount of interest surrounding the task of accurately predicting relevant web content using social media data is related to the ability to anticipate the impact of such content in order to properly suggest it to users. Therefore, in addition to successfully tackling the predictive task, a second component also plays a significant role: time. Upon the publication of a given web content, the interest in anticipating its popularity decreases over time. After users start to generate related social feedback, and given a period of time, it will become obvious which content is highly relevant or not. As such, the problem addressed in this thesis does not only focus on predictive accuracy towards the most relevant web content, but also concerning the ability to suggest it as soon as possible.

In this thesis, experimental evaluation efforts are focused on the online news type of social media data. Online news is one of the most researched types of social media data in popularity prediction tasks. This type of data is very interesting due to two reasons. First, it is massively diffused over social media platforms, but its life-span is relatively short, raising greater interest in its early and accurate prediction of the highly popular items. Secondly, it is very heterogeneous in terms of data considering that it is not only described by related social feedback from online platforms, but it also has official meta-data descriptors from news outlets (*i.e.* title, media source, publication date).

1.3 Motivation and Main Contributions

Previous work concerning the task of predicting the popularity of web content has determined the success of proposals based on standard error metrics. Overall, these metrics quantify the magnitude of prediction errors. However, given the heavy-tail distribution and the importance of accurately predicting rare cases of highly popular web content, these metrics may lead to over-estimated results and misleading conclusions. As such, serious issues may be raised in understanding if the evolution of web content popularity prediction approaches has provided an increase in predictive ability concerning the cases that are the most relevant.

Although the problem of imbalanced domain learning has an extensive record of research in classification tasks, only recently a framework was proposed in order to provide regression

tasks with an adaption of the concept: utility-based regression [196].

In this thesis such framework is leveraged. It is an adaptation of cost-sensitive learning for regression tasks. It is used as a platform to analyse approaches to the task of predicting web content popularity and for the development of new approaches that are more attuned to the problem of imbalanced domain learning. However, some issues arise concerning the concept of utility-based regression. The proposal is originally focused on actionable forecasting tasks. In such tasks, the predicted target value is used in order to make a certain decision. However, the specifics of such tasks are not entirely adaptable to the scope of this thesis, and therefore require an extension concerning the evaluation of the prediction tasks.

Given that the objective of this thesis includes the ability for timely suggestions of web content, the matter of ranking the results of prediction models is also crucial. Previous work in predicting the popularity of web content and ranking such predictions have focused on single-source scenarios. However, a significant problem arises when considering the need to rank a given web content as to its popularity in scenarios with multiple social media data sources.

The work carried out and described in this thesis lead to the following main contributions:

i) an extensive review of previous work is presented, including a thorough discussion on several open issues;

ii) two new data sets of online news feed data are presented, including an exploratory analysis of the data;

iii) unlike previous work, the task of web content popularity prediction is framed as an imbalanced domain learning task;

iv) a new approach for deriving utility surfaces is proposed, based on rules knowledge;

v) a new evaluation metric is proposed, and a proper evaluation framework for popularity predictions tasks is presented;

vi) based on the concept of utility-based regression, a set of diversified prediction approaches are proposed;

vii) an extensive evaluation as to predictive and ranking accuracy of the prediction approaches is presented, for both single-source and multi-source scenarios.

1.4 Organization of the Thesis

This thesis is organized in seven chapters, outlined as follows.

Introduction
In the present chapter, the context and the problem definition of this thesis is described and the motivations and contributions of this work are presented.

Literature Review
In the second chapter a review of previous work is presented. A discussion is provided on key aspects of existing work, including the type of predictive modelling tasks used to formalize the problem, objectives, the types of features used as well as evaluation metrics employed to assess the accuracy of the models. A thorough discussion is provided, pointing out some of the main caveats raised by previous work, motivating the work presented in this thesis

The Case of Online News Feeds
In the third chapter two new data sets are presented concerning online news feeds data. Upon the description of the methods used to create such data sets, an exploratory analysis is provided. This analysis is based on 10 research questions, providing significant insights.

Learning with Imbalanced Domains
The fourth chapter raises the issues related to the focus of this thesis on highly popular web content. The task of utility-based regression is presented and detailed on its key aspects. A discussion is provided on the adequacy of existing evaluation metrics for the task tackled in this thesis. Previous work and new proposals are combined to present a new evaluation framework. This framework is detailed and finally, an experimental analysis is provided and results are discussed.

Popularity Prediction Models
The fifth chapter describes the approaches proposed to tackle the prediction of web content popularity when focusing on highly popular content. It thoroughly describes a representative group of existing approaches. It further describes each of the proposed approaches in this thesis. Finally, an extensive experimental evaluation is carried out.

Single- and Multi-Source Ranking
The sixth chapter introduces the problem of ranking web content in two scenarios: single- and multi-source data. The use of popularity prediction models to derive rankings is evaluated in an extensive experimental evaluation, and a discussion of results is provided.

Conclusions
The seventh and final chapter concludes the thesis. It provides a summary of the motivations and contributions provided by this work, as well as insights on future directions of research in the topic of web content popularity prediction.

1.5 Bibliographic Note

Some of the work described in this thesis has already been published. The following list provides a reference to such publications and the chapter to which the contribution is related.

- Nuno Moniz, Luís Torgo, Improvement of News Ranking through Importance Prediction, *NewsKDD Workshop - Data Science for News Publishing* (co-located with KDD'2014), New York, United States of America, 2014. **(Chapter 5, 6)**
- Nuno Moniz, Luís Torgo, Fátima Rodrigues, Resampling approaches to improve news importance prediction, *IDA'2014 - Thirteenth International Symposium on Intelligent Data Analysis*, Leuven, Belgium, 2014. **(Chapter 4, 5)**
- Nuno Moniz, Luís Torgo, Combining Social and Official Media in News Recommender Systems, *ECML PKDD 2015 - Doctoral Consortium*, Porto, Portugal, 2015. **(Chapter 4, 5)**
- Nuno Moniz, Paula Branco, Luís Torgo, Resampling Strategies for Imbalanced Time Series, *3rd IEEE International Conference on Data Science and Advanced Analytics (DSAA 2016)*, Montreal, Canada, 2016. **(Chapter 5)**
- Nuno Moniz, Luís Torgo, Magdalini Eirinaki Time-Based Ensembles for Prediction of Rare Events In News Streams, *2nd International Workshop on Data Science for Social Media and Risk* (co-located with ICDM'2016), Barcelona, Spain, 2016. **(Chapter 5)**
- Nuno Moniz, Paula Branco, Luís Torgo, Resampling Strategies for Imbalanced Time Series Forecasting, *International Journal of Data Science and Analytics*, Springer, 2017. **(Chapter 5)**
- Nuno Moniz, Luís Torgo, Magdalini Eirinaki, Paula Branco A Framework for Recommendation of Highly Popular News Lacking Social Feedback, *New Generation Computing Journal* (EPIA'2017 Journal Track), Springer, 2017. **(Chapter 5, 6)**

Chapter 2

Literature Review

In this chapter the work described in this thesis is framed, by presenting a thorough review of related work on web content popularity prediction and ranking. Such tasks has been extensively studied and analysed throughout the last decade, presenting many distinct approaches. Such approaches can be framed within two objectives: the prediction and ranking of web content popularity before or after the items are published. A detailed description of such approaches is presented concerning their various aspects such as the formalization of prediction tasks, features used and evaluation methods employed. A discussion on their strengths and shortcomings is introduced, motivating the work developed in this thesis.

2.1 Introduction

The exponential growth of online users and content developed rapidly throughout the course of the last decade. Many justifications for such development may be presented, such as: *i)* increased connectivity, *ii)* multi-modal access, and *iii)* the rise of Social Media.

Increased Connectivity. The age of information in which we currently live has been defined by the appearance and exponential growth of Internet users, as observed by Internet World Stats[1]. The results of this observation state that from the year 2000 until 2012 the number of Internet users grew 566.4%, and that in 2012 the number of users was set around 360 million worldwide, with the highest concentration registered in Asia (44.8%), followed by Europe (21.5%) and North America (11.4%).

Multi-Modal Access. The number of Internet users has risen at an accelerated pace, accompanied by the evolution of devices. Today, these provide near full-time access to the Internet. The main contributor to this new reality is the rise of mobile technology, both concerning the devices themselves and mobile data connectivity. This allows an increasing

[1]Internet World Stats: `www.internetworldstats.com`

number of Internet users to access the Internet from virtually anywhere and at anytime.

The Rise of Social Media. Social media platforms such as Facebook [2] and Twitter [3] have greatly influenced the Internet due to their ability to not only connect users, but mainly for allowing users to generate content of various types. These platforms continue to expand at a great pace, increasing their number of users, and the amount of content generated by them. Additionally, with the growing connectivity between users, these platforms form powerful mechanisms that are highly important for both information dissemination and information search.

The combination of these three factors caused a severe shortening of distance between each and every other person in the world. Consider the concept of "six degrees of separation": initially set forth by poet and journalist Frigyes Karinthy in his short story Chains [120], it claims that everyone in the planet may be reached by at most six connections to other people, i.e. six "friend-of-a-friend" iterations. This idea has been the focus point of much multi-disciplinary research, namely the famous experiment by Milgram [169] and the more recent contributions in computer science of Barabasi [25] and Watts [234]. Recent studies provide important insights concerning the impact of social media. For example, Ugander et al. [229] provided a study of Facebook's social graph concluding that the network approaches full connectivity in one large connected component. Also, Backstrom et al. [22] has shown that in the year 2012 the degree of separation in Facebook was in fact 4.74, therefore less than the original idea of "six degrees of separation". This shows not only the impact of social media platforms today, but also the potential impact for content spreading: if the world is "smaller", it is easier and faster for information to travel to massive amounts of users.

The increasing amount of available user-generated content presents crucial opportunities in order to understand and improve user experience. One of the most challenging tasks involving web content concerns accurately forecasting its popularity, in order to promote agile and accurate suggestions to users. This is the main focus of this thesis, and in this chapter a thorough review of related work concerning this problem is presented.

In the remainder of this chapter the following subjects are addressed:

- motivation of the use of social media and web content in data mining and machine learning tasks;
- description of previous work on prediction tasks using such data;
- presentation of evaluation methods and formalization of the metrics used in previous work;

[2] Facebook: `http://www.facebook.com`
[3] Twitter: `http://www.twitter.com`

- discussion concerning the issues raised by web content popularity prediction and evaluation.

2.2 Social Media and Web Content

Social media is a broad term encompassing forums, blogs, video sharing websites, collaborative coding platforms and social networking platforms (e.g. Facebook, Twitter) [223]. Their common denominator is the possibility for users to generate content, i.e. web content. Web content generated by social media platforms (also referred to as social media data) can be generically defined as any type of information on a given web site. A more elaborate definition was introduced by Tatar [214], described as follows.

Definition 2.2.1. Web Content: any individual item, publicly available on a web site and which contains a measure that reflects a certain interest shown by an online community.

This content is diverse, including text, images, audio and videos, and the impact of social media platforms is enormous when considering the volume of web traffic and content it generates. Consider the following statistics from three of the most popular platforms.

- Youtube has over a billion users (almost a third of Internet users) who watch several hundreds of millions of hours in videos and generate billions of views, where more than half come from mobile devices [250]. Youtube is owned by Google, which generates 6-10% of all Internet traffic, and most of it comes from Youtube [96];
- Facebook counted 1.18 billion daily active users on average in September 2016, where most of those (1.06 billion) accessed the social media platform with mobile devices [77];
- Twitter has 313 million monthly active users, where 82% are active on mobile devices, generating over a billion visits to sites with embedded tweets [228].

The overwhelming amount of traffic and data generated by social media also provides interesting opportunities for researchers. For example, the conditions provided by these platforms allow the study of the inter-play in social networks and the dynamics of web content concerning both its generation and spreading.

However, the context in which social media data is generated presents several factors that may facilitate or hinder the ability to use such data [152]. On one hand, social media data is massive and linked. On the other hand, it is commonly noisy, sparse, informal and biased. Additionally, it is heterogeneous, partial, asymmetrical and multi-sourced. This combination of factors illustrates well the difficulty of using web content and their impact in reducing the ability to successfully disclose its full potential.

To illustrate the opportunities opened by web content generated in social media platforms, or social media data, consider the case of Twitter. Twitter is a micro-blogging network where communication is done through small communication packets called tweets that must not contain more than 140 characters. This network facilitates access to information published on it, making it a popular data resource in research. Social media data from Twitter has been used in tasks related to many research topics, such as information retrieval, topic modelling, summarization and others [143, 257].

Concerning the field of information retrieval, Dong et al. [63] use real-time data from Twitter to crawl and detect fresh URLs and also to compute features for ranking those URLs. Massoudi et al. [162] propose a retrieval model for searching data from Twitter according to a given topic of interest, by introducing quality indicators to enhance the proposed model, and a dynamic query expansion model for posts retrieval. On this same path, other proposals were made, such as the intersection of event detection with query expansion methods [202], the analysis of hashtags [108], amongst others [59, 156]. Efron et al. [70] present a review that may be of interest for a further and deeper analysis on the state-of-the-art concerning information retrieval in micro-blogs and several other sub-areas.

In the field of topic detection and tracking, Cheong et al. [48] use a combination of visualization techniques and data mining to classify messages related to a determined topic, by accounting for the demographics of the users. A technique to detect emergent topics in real-time is proposed by Cataldi et al. [42] by modeling terms' life cycle and applying an aging theory which leverages user authority, in order to study its usage in a specific time frame. O'Connor et al. [178] present TweetMotif, an exploratory search engine for Twitter based on a message clustering technique for similar terms. The topic extraction system used involves syntactic filtering, language modeling, near-duplicate detection and set cover heuristics.

Focusing on text summarization, Xu et al. [245] propose an event graph-based method to create summaries of variable length for different topics. The authors use an extended version of the PageRank algorithm proposed by Page et al. [182], in order to partition event graphs and detect the fine-grained aspects to be summarized. A participant-based approach for event summarization is proposed by Shen et al. [205], using a novel mixture model that combines the "burstiness" and cohesiveness properties of event tweets. Based on user interaction information on Twitter, Chang et al. [44] propose an approach to summarization that leverages those signals, creating Twitter context trees for that process within a supervised learning framework.

As made clear by the previous descriptions, the potential of social media data is immense. In addition, its broad definition concerns a very diverse set of data types beside text, where each type contains particular properties. For example, online news are a very dynamic type of web content, due to the information it conveys and its impact on everyday life. As such, new events and stories are constantly being published as well as updates on old stories. This

causes news items to have a very short life span [60, 246]. In contrast, online videos have a much longer alive-time, enduring for weeks or months [212].

2.3 Popularity Prediction

The concept of web content popularity has known different definitions, mainly due to its subjectivity [135]. For instance, popularity can be defined as the number of views on an online news site or Youtube; the number of likes, comments or shares in Facebook; the number of "diggs" in Digg; or the number of retweets in Twitter. In order to proceed, instead of defining popularity in relation to a given data source, we provide the following definition.

Definition 2.3.1. Web Content Popularity: Given a web content item, available in a single or multiple data sources, the popularity associated to it is given by a metric, or combination of metrics, capable of reflecting the magnitude of attention the item received by users, in a given period of time.

A large portion of research using social media data has been focused on attempting to predict the popularity of web content. The underlying interest of solving this task is mainly related to improving user experience, by providing more timely and accurate suggestions of web content or anticipating information needs, amongst other aspects of user activity. Accounting for this potential, researchers have addressed this problem differently, regarding various dimensions. These include the objective of the work, the type of features used to build the prediction models, the data mining task used to formalize the problem, or even the evaluation framework employed.

Objectives. A singularity of the work related to the scope of our problem is the objective of the proposed approaches. The objectives presented are mainly: *i)* to predict web contents' popularity before they are published or upon publication, and *ii)* to predict the popularity after the items are published. The main distinction between these objectives are the features used for modelling. The prediction of popularity prior or upon publication relies solely on descriptors of the items, as social feedback from users is still not available or scarce. Therefore, it is not possible to peek into the early signs of the evolution of popularity. Oppositely, when the objective is to predict the popularity of web content items after they are published, approaches commonly resort to the use of early social feedback in social media sources, in order to boost performance. The objectives mentioned have been previously referred to as "ex-ante" or "ex-post" prediction [161]. Throughout our work, we will refer to these objectives as *a priori* prediction and *a posteriori* prediction, regarding approaches where the aim is to predict the popularity of web content before their publication or after publication, respectively.

Features. An important component of related work concerns feature selection. A diverse set of features and combination of features has been used in previous approaches. We categorize such features in six classes: *i)* behavioural, *ii)* social network, *iii)* content, *iv)* temporal, *v)* meta-data, *vi)* external sources. These are described as follows, providing some examples:

i. Behavioural: features describing the evolution of popularity w.r.t. a given item, e.g. number of comments [118, 109], votes [140, 207] or number of early adopters [239, 47];

ii. Social Network: descriptors of given users' social network including, for example, its structure [109, 226] the number of followers and/or followees [140, 185, 95], times listed in user groups [185, 133];

iii. Content: includes a broad spectrum of features, ranging from bag-of-words models [114, 164] where features are derived from the presence of given words [174], to natural language processing and text mining tools such as sentiment analysis [29], subjectivity in language [24] or both [18], length of text [15] or presence of hashtags and emoticons [174];

iv. Temporal: description of the temporal aspect of the items, such as the time difference between the publishing time and first page view [121], the month, day and/or hour of publication [224];

v. Meta-data: features providing indicators concerning the items or the users, such as if the item contains a summary or the number of authors [224], the usual users' topics of interest, and the interaction of topics [161], or the past influence of publishing entities [23];

vi. External Sources: describes features obtained from external sources to the actual source of the data, such as temperature (in Celsius) at publication time [224], the contents' URL category according to the OpenDirectoryProject[4], or the popularity of recognized entities in Wikipedia, Twitter and in web search [15]. In some cases [179, 200, 41] behavioural features are extracted from external sources[5].

This comprehensive list allows a thorough analysis of the feature selection process carried out by the various contributions to tackling the problem of web content popularity prediction. Also, it should also be noted that behavioural features are only used in *a posteriori* prediction approaches.

Data Mining Tasks. Concerning the data mining tasks used, previous work has mainly focused on classification, regression and time series forecasting tasks. A notable distinction between these tasks is the type of target variable. The objective of the first task, classification, is to predict a nominal value, e.g. an item being classified as "high" or "low" regarding its popularity (e.g. [140, 224, 121]). As for regression, its objective is to predict

[4]OpenDirectoryProject: `https://www.dmoz.org/`

[5]In such cases, we consider that the approaches use both behavioural and external sources' features.

a numeric value, such as the exact amount of an items' popularity (e.g. [24, 157, 217]). Finally, in the case of time series forecasting tasks, it enables the use of either nominal or numeric values. In addition to formalizing the problem according to one of the previously mentioned prediction tasks, other contributions have formalized this problem differently, such as time series clustering, learn to rank, survival analysis and matrix factorization tasks. Time series clustering is a technique for classifying similar temporal sequences, by placing them in homogeneous groups without previous knowledge of the groups' definition [4]; in learn to rank tasks the objective is to predict the order of a given set of examples according to a given criteria; survival analysis refers to a statistical approach that analyzes and predicts the expected time duration for a given event (or events) to happen; finally, in matrix factorization tasks, the objective is to decompose relational matrices into (commonly) two factors, allowing for easier inspection and reasoning.

2.3.1 Classification Tasks

Classification tasks assume the existence of an unknown function that maps predictor variables to a nominal target variable. This function can be defined as $Y = f(X_1, X_2, \cdots, X_p)$, where Y is the nominal target variable, $X_1, X_2, \cdots, X_p$ are features describing the items and $f()$ is the unknown function we want to approximate. In order to obtain an approximation (a model) of this unknown function we use a data set with examples of the function mapping (known as a training set), i.e. $D = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n}$.

A Posteriori Approaches

Early work by Lerman and Galstyan [140] studies the relation between social networks' settings and the promotion of content in the news platform Digg[6], which allows users to share news items. The authors approach the problem of predicting the popularity of news in the platform by discretizing the amount of votes obtained by items into two classes: interesting and non-interesting, using an ad-hoc threshold of 520 votes. The prediction approach uses a decision tree classifier [240], and a data set with behavioural and social network features. The authors show that, in a small set of cases, it is possible to predict how interesting a story will become, using statistics of its social network spread.

Focusing on virality, Weng et al. [239] propose an approach to predict meme (i.e. web content) virality soon after they are published using data from Twitter. The proposal is solely based on leveraging information concerning social network and behavioural features. The authors frame the problem as a binary classification problem, where memes are denoted as viral or not using a percentile threshold (experiments include thresholds of 70%, 80%

[6]Digg: `http://digg.com/`

and 90%) regarding the number of tweets obtained by memes. The proposed approach uses Random Forest [36] classifiers and results show that it provides better results than approaches solely based on memes' content features.

Using data from news items and the news platform Digg, Jamali and Rangwala [109] propose to predict a popularity index defined as the difference between the up-votes (digg) and down-votes (bury). Based on the work of co-authorship [145] and citation networks [153], the authors propose a prediction approach using behavioural, social network and content features. Prediction models are built using decision trees [235], k-nearest neighbor classifiers [5], and support vector machines (SVM) [230], for three different scenarios (2-, 6- and 14-class), with different intervals of discretization. The authors show that the two most discriminating features were the number of comments per story, and the sum of up-votes until the time of prediction. Also, results show that SVM models obtained the best overall performance and that it decreases as the size of the training set is reduced.

Gupta et al. [95] approach the problem of predicting future popularity trends of events in Twitter. The objective is to predict the popularity of events in the next time interval. The authors defined the problem as a 5-class classification problem, by discretizing the target variable as nominal, where three of the classes describe different levels of decrease in popularity, one describes an increase and the final class describes a non-significant change in popularity. In order to train the prediction models, the authors decided on a feature space of behavioural, social network and content features. Experiments include classification modelling tools such as SVM, k-nearest neighbors, naïve bayes and decision trees classifiers. Results confirm that the SVM models obtained the best results in comparison to other modelling tools. With a similar objective, Cheng et al. [47] formalize the problem as a probabilistic classification task (logistic regression), and attempt to predict if the popularity of given tweets will grow higher than the median of known cases. To build these models, the authors propose the use of behavioural, social network, content and temporal features.

Keneshloo et al. [121] approach the problem of predicting the dynamics of online news views in order to anticipate when the item will reach its popularity peak. Using data from the well-known news outlet Washington Post, the authors use a diverse set of features (behavioural, social network, content, temporal and metadata) to build predictive models using Random Forest classifiers. The authors frame the problem as a two-step classification problem. First, they address the prediction of the shape of the visualisations' evolution as a 4-class classification problem, corresponding to four clusters of possible views' evolution shapes. Afterwards, for each cluster and using this information, the prediction problem is formulated as a binary classification problem, in order to predict a peak in news views for several points in time of the near future.

A cross-domain algorithm (see [214]), *Social Transfer*, is proposed by Roy et al. [200] with the objective of predicting popularity bursts of Youtube videos, using related data from Twitter.

The approach consists of extracting topics from Twitter and associate them with the topics of Youtube videos. Using this information, the popularity in both sources is compared and used as an indicator of future activity. Experimental results using SVM classifiers and behavioural, meta-data and external sources features show that the use of social feedback from other social media sources is capable of improving the prediction of which videos will have sudden bursts of popularity. Authors report an increased performance of up to 60% when compared to solely using data from Youtube.

The work of Schulman et al. [207] questions the ability of prediction models in anticipating item's popularity when peeking into its early behaviour. The authors compose the problem as a binary classification task where the objective is to predict if a given item will be more popular than a certain percentage of other items, using data from several sources. The proposal includes behavioural, social network, temporal and meta-data features, which are used to build predictive models with logistic regression, random forests and SVM classifiers. Results show that the best predictive accuracy is reported when using logistic regression. More importantly, authors conclude that in addition to obtaining fairly good results in terms of predictive accuracy, these models also generalize well, performing with comparable accuracy in data sets of other social media sources.

A Priori Approaches

Unlike previously described approaches, Petrovic et al. [185] tackle the problem of anticipating if a given tweet will be retweeted (shared), i.e. *a priori* prediction. The authors propose a machine learning approach to time-sensitive Passive-Aggressive classifiers [53], where the overall model is combined with hourly models, using social network and content features. Results show that this approach improves the results obtained by naive baselines (random guessing, predicting all positive) and a standard Passive-Aggressive classifier model. Additionally, an analysis of the contribution of features shows that the performance is dominated by social network features, but content features add a considerable improvement.

Tsagkias et al. [224] tackle the problem of *a priori* prediction by attempting to anticipate the number of comments that news items will receive, as a two-step classification problem. First, if the news items will obtain any comments, or not. The second problem is also presented as a binary classification where the objective is to predict if the items will obtain a low or high volume of comments. The discretization of the number of comments is done by using the inverse cumulative log-normal distribution function at 0.5. The authors proposal uses a broad type of features including social network, content, temporal, meta-data and external sources features, and models are built using Random Forest classifiers. Results reported by the authors show that although it is possible to obtain good results concerning the first task, the performance of the models degrades considerably in the second task.

Lakkaraju and Ajmera [135] study the problem of anticipating the popularity of Facebook Pages' posts before they are published. In a setting of closed communities the authors propose a 5-class classification problem in order to predict if a given item will obtain "very less attention", "less attention", "mediocre attention", "high attention" or "very high attention". Using SVM classifiers and social network, content and temporal features (denoted by the authors as an Attention Prediction Framework), results report a considerable improvement of predictive accuracy over some baseline approaches [194, 248].

Addressing this problem as a logistic regression task, Hong et al. [104] approach the task of web content popularity prediction, focusing on the popularity of tweets. Using social network, content, temporal and meta-data features, the authors frame the problem similarly to Tsagkias et al. [224], by using a two-step procedure. The first task is focused on predicting if a given tweet will be retweeted in the future. The objective of the second task is to predict the volume range of future retweets (discretized as in the work of Khabiri et al. [122]). Similarly to the first task of the previously mentioned work, Naveed et al. [174] propose to predict the likelihood of a given tweet being retweeted in the future in an *a priori* context. In this case, the authors use content and meta-data features.

Concerning the prediction of news popularity, Bandari and Huberman [24] propose the use of content and meta-data features on data from Feedzilla[7] (news) and Twitter (related tweets). The problem is framed as a 3-class classification problem, where the amount of tweets related to the respective news are discretized according to the following rules: class A news have between 1 and 20 tweets, class B news from 20 to 100 tweets, and class C news have more than 100. The authors use support vector machine (SVM), decision tree, and Bagging [97] classifiers to build predictive models, reporting that the best results were obtained with the Bagging approach.

Finally, Arapakis et al. [15] study the feasibility of *a priori* approaches, by using news items from Yahoo News[8] and related tweets. The authors review the work of Bandari and Huberman [24] and analyse the results w.r.t. several dimensions of the approach: experimental methodology, evaluation metrics, and features used. A careful analysis of the prediction results shows that the distribution of classes is skewed, and therefore, the results reported may be over-estimated when concerning its ability to predict highly popular news. As such, the authors conclude that the classifiers were biased to learn unpopular news and that *a priori* prediction of news popularity is still an unsolved problem.

[7] This site was discontinued.

[8] Yahoo! News: `https://www.yahoo.com/news/`.

2.3.2 Regression Tasks

Regression tasks are similar to classification tasks in terms of their definition. Given a set of features $X_1, X_2, \cdots, X_p$ describing the items, the objective of the task is to approximate an unknown function $f()$ in order to predict a target variable Y. However, in regression tasks, Y is a continuous variable, as opposed to the nominal target variables in classification tasks.

In some of the previously described work concerning classification tasks [109, 95, 135, 24, 15] the authors also modelled their data as a regression task. Using the same feature spaces, the authors modelled the problem using a numerical value of popularity instead of its discretization as the target variable. Concerning the modelling tools used in such approaches, for *a posteriori* approaches, Jamali and Rangwala [109] used SVM models and Gupta et al. [95] used linear regression models. As for *a priori* prediction approaches, Lakkaraju and Ajmera [135] used SVM models, and Bandari and Huberman [24] and Arapakis et al. [15] experimented with linear regression, k-nearest neighbors and SVM models.

A Posteriori Approaches

Seminal work by Kaltenbrunner et al. [118] approach the problem of predicting the amount of comments that news items will receive in the popular news site Slashdot[9], solely using data concerning the first moments of their respective activity, i.e. behavioural features. The authors report experiments with different prediction time windows (30 minutes, 1, 8 and 24 hours and 14 days). The proposed approach is based on fitting the data in a PCI-distribution, using such fit to predict the future amount of comments that news will obtain. Results show that, despite the high levels of prediction error, this approach is capable of predicting the magnitude of users' reaction to posts using a small amount of the comments' total.

With a similar objective, Szabo and Huberman [212] propose two modelling approaches for the *a posteriori* prediction of web content popularity: the constant scale model and the log-linear model. The authors validate their proposal with news items from Digg and Youtube videos. The objective of the authors is to predict the future amount of votes in the former, and the future number of views of the latter. As in the work of Kaltenbrunner et al. [118], these models are solely based on information extracted from the early behaviour of the respective items. Results show that the proposed models are capable of achieving interesting results in both data sets.

Furthermore, the authors study the time series of popularity evolution, concluding that most news achieve their top level of popularity within one day. In contrast, videos keep attracting views for a longer period of time. These conclusions on the dynamics of popularity are confirmed by the work of Tsagkias et al. [225]. In this work, the authors use the log-linear

[9]Slashdot: `https://slashdot.org/`

model proposed by Szabo and Huberman [212] to predict the number of comments that news from several news outlets (e.g. Algemeen Dagblad, De Pers, Financieel Dagblad, Telegraaf) and a collaborative news platform (NUjij) will obtain in the future.

Kim et al. [123] propose the prediction of the popularity of an article using blog posts from a political discussion blog in South Korea, SEOPRISE. The objective set by the authors is to predict the number of hits the articles receive. The authors use an analogy between virtual temperature and the concept of popularity, defining four levels on a temperature scale: cold ($hits < 100$ comments), warm ($100 \leqslant hits < 500$), hot ($500 \leqslant hits < 1500$) and explosive ($hits \geqslant 1500$). As in the previously described approaches for regression tasks, Kim et al. solely resort to data concerning the evolution of the posts' popularity. The proposed approach for prediction shares the same idea as the log-linear model proposed by Szabo and Huberman [212], by transforming the data with logarithms and studying its linear correlation. Results show that the approach is able to achieve a good performance in articles with the most common discrete temperatures (cold, warm and hot) using data of the first 30 minutes of activity. The authors denote the difficulty of predicting the highly popular cases ("explosive").

Tatar et al. [217] address the problem of popularity prediction by proposing a simple linear regression model using news items' data from a news web site (20Minutes). Although also based solely on behavioural features, the methodology shows efficiency in evaluating the expected popularity of news articles after their publication. This work was extended [215, 216] in order to compare the proposal to the work of Szabo and Huberman [212] (the constant scale model and the log-linear model) and to provide an analysis of the efficiency of such prediction models in accurately ranking news. The authors claim that a simple linear model out-performs such proposals.

Proposing a two-step prediction approach, Ahmed et al. [6] tackle the problem of predicting the popularity of news and video items from three different sources: Digg, Vimeo and Youtube. The authors propose an approach that is conceptually similar to that of Keneshloo et al. [121], although for regression tasks. Firstly, the patterns of popularity evolution are clustered, and secondly, the future popularity of the content is predicted using maximum likelihood path prediction. The clustering process is carried out using the Affinity Propagation algorithm [84]. Evaluation using linear models as baselines, shows that the proposed approach significantly improves predictive accuracy.

Shen et al. [206] propose a generative probabilistic framework using a reinforced Poisson process [184] in order to model the evolution of items' popularity, building on the bayesian concept of conjugate priors. Solely based on behavioural features, the authors evaluate this proposal with well-known approaches (e.g. [212]) using a data set containing all papers and

citations published by American Physical Society[10] between 1893 and 2009. Building on this, Gao et al. [87] propose an approach to predict the future popularity of messages in Weibo[11], based on an extended reinforced Poisson process model. The proposal includes a time mapping method and a power-law temporal relaxation function, in order to better capture retweeting dynamics. Also based solely on behavioural features, experimental results report that the proposed approach is able to provide better results than other approaches [191, 206, 212] both in terms of predictive and ranking accuracy.

Lerman and Hogg [142] propose an approach for popularity prediction based on stochastic models of user behaviour. Unlike previously described regression approaches that solely explore the correlation between popularity in different points in time, the authors specify a mechanism to explain the evolution of popularity by combining behavioural and social network features to build predictive models. The proposed approach is based on a mathematical model of the dynamics of social voting [139, 103]. Experimental results show that the proposed approach is capable of accurately predicting the success of a story by tracking its spread through the social network and by using their first 10 votes.

Also based on behavioural and social network features, Yu et al. [251] tackle this prediction problem differently, by framing it as a cascading process prediction. This addresses the problem of predicting the cumulative cascade size within a given social network. Cascades can be loosely defined as the behaviour of re-sharing content from user to user [47]. The authors propose a novel Networked Weibull Regression model and a novel method for predicting cascading processes through the aggregation of behavioural dynamics. An experimental evaluation was carried out using publications from Weibo, which are similar to tweets, using three baselines: the Cox Proportional Hazard Regression model [52], the Exponential/Rayleigh Proportional Hazard Regression (special models of the former) model and the log-linear model [212]. Results show that the approach proposed by the authors obtained the best results, but also that the popular log-linear model did not perform well, which is justified by it focusing on the average behaviour of the data (i.e. small-sized cascades), a conclusion also corroborated by the work of Cheng et al. [47].

Accounting for the circadian nature of user activity and aging of information, Kobayashi and Lambiotte [124] approach the popularity prediction problem by modelling the system by a Time-Dependent Hawkes process, using data from Twitter. Specifically, the authors use a feature space based on the temporal patterns of retweet activity from the original tweets and the underlying social networks. Experimental evaluation results on a set of popular tweets (with over 2000 retweets) show that the proposed approach is able to out-perform several baselines [212, 206, 87, 256].

A proposal to predict the popularity of hashtags in Twitter is proposed by Tsur and Rap-

[10] American Physical Society (APS): `https://www.aps.org/`.
[11] Weibo: `https://www.weibo.com`

poport [226] in order to study the propagation of information in social networks. The authors propose a linear regression approach to the problem. To optimize parametrization the proposal applies Stochastic Gradient Descent [32] and the Nelder-Mead method [175]. Results show that by combining behavioural, social network, content and temporal features, the proposal was able to minimize prediction error.

Kupavskii et al. [133] study the dissemination of information in Twitter and address two predictive tasks, differing on the data used and the target variable: the first predicts the number of publications using only information available at the moment the item is originally shared (*a priori* prediction), while the second predicts the audience size using additional information on the cascade growth up until the moment of analysis (*a posteriori* prediction). The authors use behavioural, social network, content, temporal and meta-data features to build the data sets, and train all models using gradient boosted decision [99]. The authors show that predicting the audience of a tweet is more precise than predicting the amount of retweets.

Kong et al. [126] study the dynamics of hashtags in Twitter, based on formal definitions of popularity evolution states: *i)* active, *ii)* bursting, *iii)* off-burst, and *iv)* inactive. Based on these definitions, the authors propose to tackle two distinct research questions: *i)* will an active hashtag burst in the near future, and *ii)* what will be the popularity of a bursting hashtag. Using behavioural, social network, content, temporal and meta-data features, results report that weighted SVM models provide the best results concerning the first task, and that Gaussian Process Regression achieves the best performance in the second task. Additionally, by studying the contribution of the various features used, authors report that the biggest negative impact in predictive accuracy occurred when removing temporal features.

Using data from Twitter, Asur and Huberman [18] propose an approach to forecast movies box-office revenue. Although not focused on web content popularity prediction, it provides insights into the evolution of the movies' attention and popularity. The authors propose a linear regression approach using behavioural and content features and evaluate their work against another similar proposal by Zhang and Skiena [255] which uses behavioural, content and external sources features. Results show the authors' approach was able to significantly improve forecasting accuracy in comparison to previous work. In addition, results also show the positive impact of using sentiment analysis to build features in the models' evaluation.

Based on the theory of self-exciting point processes [171], Zhao et al. [256] propose a statistical model to predict the popularity of tweets. Self-exciting point processes assume that all previous points of knowledge influence the evolution of the data thereafter. Based on behavioural and social network features, the proposed approach does not require training. Experimental results show that this approach is capable of providing better results than previously proposed approaches [212, 87, 2, 54] in both prediction and ranking.

Recently, Lymperopoulos [157] proposed a three-stage approach prediction model for anticipating the popularity of web content. The proposed approach encompasses the analysis of the data, the identification of popularity evolution patterns and the use of such information in building models and prediction tasks. The approach distinguishes the popularity evolution patterns as linear and non-linear phases. Using such patterns in combination with temporal and meta-data features, the authors evaluated the proposal with data from several sources (Twitter, Memetracker and Flickr) using popular prediction approaches as baselines [191, 212, 126, 256]. Results show that the approach proposed by the authors obtained significant improvements w.r.t. all of the baselines.

Oghina et al. [179] explore the possibility of using the behaviour of users in Twitter and Youtube in order to predict ratings of movies in IMDB[12]. Using data from external sources, the authors propose a cross-domain approach resorting to features obtained from Twitter (e.g. sentiment words from tweets) and Youtube (e.g. ratio of likes/dislikes from Youtube). Using a linear regression model, the authors were able to demonstrate that the combination of data from multiple social sources is capable of boosting predictive performance. In a similar approach, Castillo et al. [41] propose the use of the behaviour of users in Facebook and Twitter in order to predict the popularity of news in the Al Jazeera website[13]. Also based on external sources features, and using multiple linear regression models, the authors show that this approach is capable of obtaining a predictive accuracy in 10 minutes of data, similar to the accuracy obtained by models solely based on behavioural features in the first 3 hours.

A Priori Approaches

In contrast with to the previously described regression approaches, Bakshy et al. [23] and Martin et al. [161] approach this problem as an *a priori* prediction task. The former addresses the task of predicting the influence of Twitter users. Using social network and meta-data features, the authors build linear regression models and confirm that the most popular items tend to be generated by the most influential users, in addition to discovering that links that have a more positive sentiment associated with them are more likely to spread. As for the latter work by Martin et al. [161], the authors provide a deep discussion and experimental evaluation on the limitations of *a priori* prediction of web content popularity. The main research question is to assess the feasibility of the popularity prediction task using simple regression models. The authors build several regression models using linear regression, regression trees and Random Forests, based on social network, content, meta-data and external sources features. Experimental results lead the authors to conclude that the best performing models were unable to explain more than half of the variance in popularity,

[12] International Movie Database (IMDB): `http://www.imdb.com/`.
[13] Al Jazeera: `http://www.aljazeera.com/`.

and that their performance is far from achieving deterministic accuracy.

2.3.3 Time Series Forecasting Tasks

In this section previous work on popularity prediction focusing on time series forecasting tasks is described. A time-series is a time-ordered set of observations of a given continuous variable $y_1, y_2, \ldots, y_t \in Y$, where y_t is the value measured at time t. The overall assumption is that an unknown function correlates the past and future values of Y, i.e. $y_{t+h} = f(\langle y_{t-k}, \ldots, y_{t-1}, y_t \rangle)$. The objective of time series forecasting tasks is to apply a learning process that provides an approximation of this unknown function, i.e. a model, in order to forecast future values of variable Y. Previous work in time series forecasting tasks are all concerning *a posteriori* prediction, since one of its requirements is data on the items' popularity evolution.

Pinto et al. [191] propose two new models, solely based on behavioural features, for time series forecasting tasks. The proposed models build on the work of Szabo and Huberman [212]. However, instead of using data on a single point in time concerning the popularity of the items, the authors propose to use historical data and sample the total number of views in regular intervals. The first proposed model consists of a multivariate linear model where weights are attributed to the intervals depicting the evolution of popularity. The second model proposed by the authors builds on the former, adding a measure of similarity to other patterns of popularity using Radial Basis Functions [99]. The models were validated using data from Youtube videos. Results show that both proposed models are able to provide better results in comparison to the work of Szabo and Huberman [212]. Additionally, the authors address an issue left unresolved by Szabo and Huberman, concerning the specialization of models. The authors studied this issue by building models for each of the available video categories which, in the vast majority of cases, had a minimal impact on the prediction error.

Using a bayesian model, Zaman et al. [254] propose to describe the evolution of tweets in terms of their respective retweets. The objective defined by the authors is to enable the prediction of the popularity of tweets as early as possible. Using behavioural and social network features, the experimental evaluation of the time series forecasting task shows the ability to predict items with under 40% of error (relative to the metric Mean Average Percentage Error, defined in Section 2.4) in the first 5 minutes after the tweet is posted.

Gursun et al. [96] provide an analysis of the popularity evolution of Youtube videos, concluding that there are two types of patterns: those with rapid changes in popularity (rarely-accessed) and those that are consistently popular in long time periods (frequently-accessed). Based on these conclusions, the authors propose two approaches for each of the types of patterns, solely based on behavioural features, i.e. the evolution of popularity. Concerning the rarely-accessed videos, the authors propose a time series clustering approach based on

hierarchical clustering. As for the frequently-accessed videos, the authors use Autoregressive Moving Average models (ARMA) [38]. The authors evaluate their proposal both in terms of predictive and ranking ability, concluding that it is capable of minimizing the prediction error and of improving the quality of the rankings, in comparison to a set of standard learning tools.

2.3.4 Other Tasks

Several proposed approaches have formalized the problem of web content popularity prediction differently from classification, regression or time series forecasting and tasks. In this section we provide an overview of such work including time series clustering, learn to rank, survival analysis and matrix factorization tasks.

Time Series Clustering

Clustering techniques are commonly employed in order to organize unlabelled data into similar groups. Clusters are formed by grouping cases that show a maximum similarity with other cases in the group and minimum similarity with other groups. Time series clustering is a special type of clustering techniques where cases concern temporal sequences. Its application is advantageous as it leads to discovering relevant patterns in time series data [4].

Concerning time series clustering approaches and the task of web content popularity prediction, Yang and Leskovec [246] propose the K-Spectral Centroid (K-SC) clustering algorithm. This time series clustering approach uses a similarity metric which is invariant to scaling and shifting. Additionally, in order to allow for its use in large data sets, it is coupled with an adaptive wavelet-based incremental approach [43] for scaling purposes. The authors' proposal is validated using news, blog posts and Twitter data. The time series clustering approach proposed in the work by Keneshloo et al. [121], and described in Section 2.3.1, uses this approach for clustering in combination with SpikeM [163], a time-series detection method.

The work of Figueiredo et al. [81] presents a thorough study of the patterns and trends of popularity evolution in groups of Youtube videos from three different data sets (most popular videos, copyrighted videos, and videos from random queries). The authors provide important insight into some of the main issues concerning the problem of popularity prediction. Results show that the most popular and the copyrighted videos achieve the majority of the observed views very early on and that this behaviour is concentrated in bursts, in contrast with videos from random queries. Also, using the work of Yang and Leskovec [246], authors denote that the most popular and the videos from random queries share the same type of popularity

trends. Finally, results show that the internal recommendation mechanisms of Youtube are the main propellers of popularity, and that videos with similar content often share the same type of trends.

Learn to Rank

These tasks are quite similar to regression and classification tasks in *a posteriori* prediction settings, where instead of predicting popularity (e.g. number of comments, popularity class) the objective of the task is to predict the position of the items in a ranking. The target variables in learn to rank tasks are ordinal, commonly described as integers, and ordered by decreasing order of relevance.

Yin et al. [249] propose the Conformer-Maverick model based on two latent personalities of items' quality voting users: *i)* conformer, those with a high degree of conforming behaviour; and *ii)* maverick, those casting opposite votes in relation to social groups. Based solely on behavioural features, the authors evaluate their proposal with posts data from Jokebox[14]. Results show that the proposed approach is capable of accurately predicting and ranking the news articles clicked by users. Also using only behavioural features, McCreadie et al. [165] propose a voting model combined with the weighting model DPH of the Divergence from Randomness framework [14]. Results show that it improves performance over other popular approaches when validated with a data set from TREC 2009 [158].

Hsu et al. [106] propose the use of SVM models with a feature space of behavioural, social network and content features using news items' data from Digg. Results show that this approach could be useful in filtering comments in order to promote those of higher quality (according to other users) and remove the low quality ones. Building on the work of Joachims [115], an approach for personalized recommendation of news articles is proposed by Morales et al. [60] by leveraging behavioural, social network and meta-data features from both social media (Twitter) and official media (Yahoo! News).

Survival Analysis

Assuming a hazard distribution, these tasks are focused on predicting the likelihood of an item's lifetime, i.e. how long will users keep sharing a web content. Lee et al. [136], building on previous work [137], propose a generalized framework for such type of task. Using behavioural and temporal features, the authors resort to the Cox proportional hazard regression model [52]. Results show that the approach is capable of predicting the lifetime of threads based on 5-6 days of observation, and also capable of predicting the number of comments the threads will receive based on its first 2-3 days of behavioural data.

[14] Jokebox: `http://www.jibjab.com/jokebox`.

Matrix Factorization

This task consists of the factorization of a given matrix into a product of its sub-matrices. For example, in the work of Wu et al. [241], the authors propose the Multi-scale Temporal Decomposition (MTD) framework, in order to decompose a popularity matrix into latent spaces based on contextual associations. In the proposed framework the authors factorize popularity of web content into user-item contexts and time-sensitive contexts and build the prediction models using SVR [65]. This proposal is validated using a data set of photos' popularity from Flickr. The performance of the proposed approach was evaluated using several baselines, and results confirm its effectiveness in popularity prediction tasks.

In order to provide a concise and summarized overview of the previous work described, Table 2.1 is presented, focusing on distinguishing the approaches by type of task, scenario (*a priori* or *a posteriori*, whether ranking evaluation was conducted and features used.

2.4 Evaluation

To evaluate the performance of the proposed approaches to the problem of popularity prediction, a diverse set of metrics has been used. Their diversity is greatly related to the task used to formalize the prediction problem. As such, the description of evaluation approaches is divided according to the type of target variable: *i)* nominal, *ii)* numeric, or *iii)* ranking.

Evaluation metrics regarding the first type of target variables, nominal, are associated to classification and time series clustering tasks. However, in some proposals concerning numeric target variables, the authors discretize the prediction outcome in order to evaluate results as a nominal variable [95]. The second type of target variables, numeric, are mainly related to regression and time series forecasting tasks. Nonetheless, we should note that they have been used on a few occasions in other tasks such as ranking [249]. The third type of target variables, rankings, are ordinal variables reporting to a type of tasks where the results are ordered by importance. This importance is commonly denoted by numeric values, ordered in decreasing fashion, i.e. the most important item is ranked first.

Nominal Target Variables

In a considerable number of previous works, approaches concerning nominal target variables frame the problem as a two-class scenario, i.e. relevant or non-relevant items. Results are commonly presented with a confusion matrix. An example of such matrix is illustrated in Table 2.2. This matrix allows the analysis of correct and incorrect prediction of classes. Cases

Table 2.1: Summary of the reviewed web content popularity prediction approaches as to the type of task and setting. It provides an indication on whether the work provides a ranking evaluation and the type of features used, including behavioural (**B**), social network (**SN**), content (**C**), temporal (**T**), meta-data (**MD**) and external sources (**ES**) features.

Approach			**Features**						
Task	**Setting**	**Ranking**	**B**	**SN**	**C**	**T**	**MD**	**ES**	**Proposals**
Classification	A Posteriori	No	✓	✓					[140, 239]
			✓	✓	✓				[109, 95]
			✓	✓	✓	✓			[47]
			✓	✓	✓	✓	✓		[121]
			✓	✓		✓	✓		[207]
			✓				✓	✓	[200]
	A Priori	No		✓	✓				[185]
				✓	✓	✓			[135]
				✓	✓	✓	✓		[104]
				✓	✓	✓	✓	✓	[224]
					✓		✓		[174, 24]
					✓	✓	✓	✓	[15]
Regression	A Posteriori	No	✓						[118, 212, 225, 123, 217] [215, 6, 206, 87]
			✓	✓					[142, 251, 124]
			✓	✓	✓				[109, 95]
			✓	✓	✓	✓			[226]
			✓	✓	✓	✓	✓		[133, 126]
			✓		✓				[18]
			✓		✓			✓	[255]
			✓			✓	✓		[157]
			✓					✓	[179, 200]
		Yes	✓	✓					[256]
	A Priori	No		✓	✓	✓	✓		[133]
				✓	✓	✓			[135]
				✓	✓		✓	✓	[161]
				✓			✓		[23]
					✓		✓		[24]
		Yes			✓	✓	✓	✓	[15]
Time Series Forecasting	A Posteriori	No	✓						[191]
			✓	✓					[254]
			✓	✓	✓				[95]
		Yes	✓						[96]
Time Series Clustering	A Posteriori	No	✓						[246]
		Yes	✓						[96]
			✓			✓	✓		[81]
Learn to Rank	A Posteriori	Yes	✓						[249, 165]
			✓	✓			✓		[60]
			✓	✓	✓				[106]
Survival Analysis	A Posteriori	No	✓			✓			[137, 136]
Matrix Factorization	A Posteriori	No	✓			✓			[241]

that are correctly predicted as being positive or negative are considered to be true positives (TP) and true negatives (TN), respectively. Conversely, instances wrongly predicted as positive or negative are respectively considered as false positive (FP) or false positive (FN).

Table 2.2: General confusion matrix for a two-class scenario.

		Predicted	
		Positive	Negative
True	Positive	TP	FN
	Negative	FP	TN

Evaluation metrics concerning nominal target variables are mostly derived from the notions presented by confusion matrices. The most common and general-purpose evaluation metric is Accuracy (Equation 2.1), which concerns the overall predictive performance of the models by denoting the rate of correct predictions in both classes.

$$Accuracy = \frac{TP + TN}{FP + TP + FN + TN} \tag{2.1}$$

It is broadly used in previous work for classification ([207, 185, 95, 15]), logistic regression ([47, 104]) and time series clustering tasks ([246]). Roy et al. [200] evaluated their work using the error rate, which is the complement of accuracy, i.e. $1 - Accuracy$. The Accuracy metric may also be applied to multi-class scenarios, as in the work of Jamali et al. [109], where the authors refer to the metric as k-way classification accuracy. This can be achieved by framing the problem as a 1-n or n-n class scenario. This could also be accomplished by applying the Balanced Accuracy metric, which measures the predictive quality for each class independently, averaged by the number of classes. This metric uses the diagonal elements of the confusion matrix (Table 2.2) and its row sums:

$$BalancedAccuracy = \frac{\frac{c_{11}}{c_{11}+c12} + \frac{c_{22}}{c_{21}+c_{22}}}{2} \tag{2.2}$$

Although widely used, the accuracy metric has raised various issues, such as its inability to handle imbalanced distributions [35], i.e. a set of cases in a binary classification scenario where the positive and negative cases are not divided equally and are not equally important. This issue, known as the accuracy paradox [260], expresses the contradiction that a model with high accuracy may not be the best prediction model, and vice-versa. Taking into account these issues, previous work has also used the F-Score metric, based on the work of Rijsbergen [198], and the receiver operating characteristic (ROC) curves [71] in order to evaluate web content popularity prediction tasks.

The F-Score metric is a combination of two other metrics (*Precision* and *Recall*). These may be loosely described as follows: *Precision* provides an indication on how accurate the model is when predicting the positive cases, and *Recall* describes how frequently the positive cases are identified as such by the model. The F-Score (F_β) combines these two metrics using the β coefficient, which defines the importance of recall w.r.t. precision. As such, when $\beta = 1$, both precision and recall have the same weight; when $\beta > 1$, the weight of recall over precision increases; and, when $\beta < 1$, precision will weigh more than recall regarding the F_β outcome. These metrics are formally described in Equations 2.3, 2.4 and 2.5. In previous work, these metrics are widely used (e.g. [239, 224, 140, 121]), which is explained by their ability to convey information on the predictive accuracy of the models on the target cases [75].

$$Precision = \frac{TP}{TP + FP} \quad (2.3) \qquad Recall = \frac{TP}{TP + FN} \quad (2.4)$$

$$F_\beta = \frac{(1 + \beta) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (2.5)$$

Finally, ROC curves are described as graphical representations of the classifiers' performance considering several levels of discrimination. They illustrate the trade-offs between the true positive rate (TPR) and the false positive rate (FPR), also named as sensitivity and fall-out, respectively, and defined in Equations 2.6 and 2.7.

$$TPR = \frac{TP}{TP + FN} \quad (2.6) \qquad FPR = \frac{FP}{FP + TN} \quad (2.7)$$

As such, ROC curves are essentially representations of TPR as a function of FPR. However, the main caveat of this approach is that given a large amount of curves, a decision on the best model may be difficult [192]. This motivates the use of the metric Area under the ROC curve (AUC) [166]. This metric, defined in Equation 2.8, allows for the evaluation of the best model regarding its average performance. Both the ROC curves and AUC have been previously proposed as alternatives to accuracy [193].

$$AUC = \frac{1 + TPR - FPR}{2} \quad (2.8)$$

In previous work addressing the problem of web content popularity prediction both the ROC curves and AUC were used in different types of data mining tasks, such as classification [109], logistic regression [47, 174] and time series clustering [96].

Numeric Target Variables

Unlike data mining tasks such as classification, where the evaluation objective is to assess the ability to correctly predict the class of a given example, the objective in numeric target tasks is to reduce the distance (i.e. metric error) between the predicted value $\hat{y}$ and the true value y of a given item.

In order to capture different aspects in terms of evaluation, the metrics proposed and used in the problem of web content popularity prediction are mostly related to the notion of absolute, relative and squared error, or a combination of such notions.

Given a set of n predicted items, the mean absolute (MAE), and squared (MSE) errors are defined as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \quad (2.9) \qquad MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \quad (2.10)$$

The main difference between these two evaluation metrics is the scale in which the prediction errors are calculated. In the case of the MAE (used in [124, 179]), the scale used is the same as the data, averaging all the pair-wise differences between predicted and true values, i.e. prediction error. As for the MSE (used in [226, 133, 24]), this metric averages the prediction error, which in turn is scaled to the power of two. The scaling property of these well known metrics entail some caveats as to the impact of under- or over-prediction. In metrics accounting for absolute errors, each pair-wise comparison contributes equally to the final score. However, this is not the case for the MSE evaluation metric: larger errors weigh more when comparing to smaller errors.

Based on the notions of absolute, relative and squared error, several other metrics have been used in previous work. For example, solely concerning the absolute error, previous approaches have evaluated their work using the Median Absolute Error ([124]), Absolute Percentage Error ([256, 157]), Mean Absolute Percentage Error ([123, 254]) and Alternative Mean Absolute Error ([18]) metrics. These evaluation metrics are defined in Equations 2.11, 2.12, 2.13 and 2.14, respectively.

$$mAE = \text{median}(|\hat{y}_i - y_i|, \forall i \in n) \quad (2.11)$$

$$APE = \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (2.12)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (2.13)$$

$$AMAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{\bar{y}} \right| \quad (2.14)$$

The most common evaluation metric in previous work is the Root Mean Squared Error (used in [15, 249, 6, 225, 179] and others), defined in Equation 2.15. It is generally conceived as a good measure of error magnitude, mainly due to its two main characteristics: as in MSE, *i)* it penalizes exponentially large prediction errors, but *ii)* it rescales the results to the scale of the target variable making the outcome more comprehensible.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \quad (2.15)$$

Following the same idea of the use of the square root, the metrics Normalized Mean Squared Error ($NMSE$) and Root Mean Squared Logarithmic Error ($RMSLE$) are also used in previous work [6, 251]. The former normalizes the squared error using the square of the true value in the respective pair-wise comparisons. The latter is defined similarly to $RMSE$ but uses the logarithms of the predicted and true values. These metrics are defined in Equations 2.16 and 2.17.

$$NMSE = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{y_i^2} \quad (2.16)$$

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(log\left(\hat{y}+1\right) - log\left(y+1\right)\right)^2} \tag{2.17}$$

Additionally, some metrics used in previous work ([216, 157, 191]) combine the different types of errors. These include the Absolute Squared Error (QSE) and the Relative Squared Error (RSE), described in Equations 2.18 and 2.19.

$$QSE = \sum_{i=1}^{n}\left(|\hat{y}_i - y_i|\right)^2 \tag{2.18}$$

$$RSE = \sum_{i=1}^{n}\left(\frac{\hat{y}_i}{y_i} - 1\right)^2 \tag{2.19}$$

Apart from the aforementioned evaluation metrics, two of the most standard tools for evaluation of models are the Correlation Coefficient (R) and the Determination Coefficient (R^2). The former (used in [15, 18, 161, 135, 23, 179]) is a statistical correlation coefficient (defined in Equation 2.20) also known as the Pearson product moment correlation coefficient [183], which measures the linear dependency between two variables, i.e. the predicted and true values.

$$R = \frac{n\sum \hat{y}y - (\sum \hat{y})(\sum y)}{\sqrt{n(\sum \hat{y}^2) - (\sum \hat{y})^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \tag{2.20}$$

The latter is a metric based on the variance proportion of the target variable, indicating the models' ability to predict such target. It is defined as the square of the correlation coefficient, R^2, and is commonly used in previous work to explain the contribution of specific sets of features [41].

Finally, in the work of Yu et al. [251], the authors used the metric Precision with δ-Tolerance. This metric is conceptually similar to the aforementioned metric *Precision*, used in tasks with nominal target variables. However, since it is not possible to apply it in the same manner, this metric considers a prediction as correct when the prediction error is smaller than δ. This enables the metric to account for cases where the prediction is considered to be accurate or inaccurate, thus allowing a procedure as denoted in Equation 2.3.

Rankings

In the previous sections the evaluation metrics described pertained to tasks where the objective is either to predict a nominal or to predict a numeric value. In the case of rankings, which are ordinal variables commonly denoted as integers, the objective is to predict and evaluate the correct order of a set of items. In terms of evaluation, given a set of queries Q where each query q has n items, which are ordered by their relevance, the objective is to evaluate the correctness of the items' order in the predicted ranking ($\hat{R}$) in comparison to that of the true ranking (R). In previous work concerning regression tasks, some authors used

the outcome of their numeric prediction tasks to derive rankings and evaluate the outcome as a ranking (e.g. [256, 216, 106, 15]).

The Spearman Correlation (ρ) and Kendall Correlation Tau (τ) are amongst the most well-known evaluation metrics for rankings. The former (used in [241]) assesses two rankings as to their deviation (i.e., error). The latter (used in [15, 215, 256]) is based on the number of pairs with the same value (concordant) and those with different values (discordant), and as such is insensitive to error. They are defined in Equations 2.21 and 2.22. Both metrics are bounded by $[-1, 1]$, and the predicted and true rank correspond to an exact monotone function when ρ or τ are equal to -1 (i.e. exact opposite) or 1 (identical).

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} \left(\hat{R}_i - R_i\right)^2}{n \cdot (n^2 - 1)} \quad (2.21) \qquad \tau = \frac{|concordant| - |discordant|}{\frac{n \cdot (n-1)}{2}} \quad (2.22)$$

Precision at k ($P@k$) and Recall at k ($R@k$) are two standard ranking evaluation metrics used in previous work concerning the scope of our problem [15, 60, 165]. Similarly to the metrics of Precision and Recall in classification, the former measures the number of relevant items on the top-k ranking positions, and the latter expresses the fraction of relevant items in the top-k positions that were correctly ranked as such. These are described in Equations 2.23 and 2.24.

$$P@k = \frac{|relevant_k \cap retrieved_k|}{|retrieved_k|} \quad (2.23) \qquad R@k = \frac{|relevant_k \cap retrieved_k|}{|relevant_k|} \quad (2.24)$$

The metric Average Precision (AP) computes the average precision for all values of i where i is the rank position, n is the number of retrieved items and $Rel(q, i)$ is a binary function evaluating the relevance of the i-th ranked item of a given query q. This binary function attributes 1 to relevant items at rank i and 0 otherwise. Furthermore, the Mean Average Precision (MAP) is computed to determine the effectiveness of the ranking mechanism over all queries, where $|Q|$ is the number of queries. These metrics (used in [165, 215]) are described in Equations 2.25 and 2.26.

$$AP(q) = \frac{\sum_{i=1}^{n} P@i(q) \times Rel(q, i)}{|relevant_i|}, \quad (2.25) \qquad MAP = \frac{\sum_{q=1}^{|Q|} AP(q)}{|Q|} \quad (2.26)$$

An important commonality of the previously described metrics is that they do not account for the position of the item itself. Conversely, the Reciprocal Rank ensures such accountability, and is described as the inverse of the rank at which the first relevant document is retrieved. As in MAP and MRP, the Mean Reciprocal Rank (used in [60]) is defined as the average of the reciprocal ranks over all queries where $|Q|$ is the total number of queries and $rank_q$ is the rank position where the first relevant item was found, for each query q. This metric is defined in Equation 2.27.

$$MRR = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\frac{1}{rank_q} \tag{2.27}$$

The previously presented metrics, although different regarding the consideration of the items' position in the ranking, are defined with the same binary concept of relevance, i.e. an item is either relevant or non-relevant. Contrary to this binary notion, the Normalized Discounted Cumulative Gain [111] ($nDCG$) allows users to associate multi-graded relevance judgments to items in a ranking. It also uses discount factors to simulate user experience by decreasing the impact of a given item's evaluation, as one goes through the ranking in a descending fashion, while most metrics weigh the positions uniformly. The normalization of Discounted Cumulative Gain (DCG) to a value between 0 and 1 is done by dividing the DCG score by the score corresponding to the optimal ordering of ranking ($idealDCG$). These metrics are described in Equations 2.28 and 2.29.

$$DCG@k(q) = \sum_{i=1}^{k}\frac{2^{Rel_{q,i}}-1}{log_2(1+i)} \tag{2.28}$$

$$NDCG@k = \frac{\sum_{q=1}^{Q}\frac{DCG@k(q)}{idealDCG@k(q)}}{Q} \tag{2.29}$$

2.5 Discussion

So far, this chapter presented a thorough and extensive review of previous work. The review addressed the problem of web content popularity prediction in several of its dimensions. These include the previous proposals' objectives, features, formalized data mining tasks and evaluation metrics used.

Martin et al. [161] performed an extensive analysis of the predictive ability of popularity prediction models. Concerning *a posteriori* prediction approaches, the authors conclude that the "peeking" strategy enables the prediction models to have a much better performance than *a priori* approaches. This is justified as using the early behaviour of the items enables the prediction models to take advantage of the cumulative advantage dynamics [207]. However, the *a priori* task is difficult to solve since it solely relies on descriptors of the items and the known behaviour of past items. Given the fast changing context of the interplay between content and user preferences, this results in more difficult prediction tasks.

The limitations of previous *a priori* prediction proposals are outlined by the experimental evaluation also carried out by Martin et al. [161], evaluating the use of a diverse set of features (content, social network, meta-data and external sources). The authors conclude the following: *i)* models solely based on content and/or meta-data features provide a poor performance; *ii)* the use of social network features is more useful, but *iii)* the best results are obtained when combined with user's past behaviour; finally, *iv)* using meta-data or external

sources' features with the best performing models does not provide significant additional performance.

Despite these conclusions, it should be noted that the authors used the Determination Coefficient (R^2) metric in order to evaluate the models. This metric is insensitive to outliers (i.e. rare cases), and therefore could cause the introduction of some bias in the analysis. In any case, the use of social network features raises serious issues in at least two aspects. First, using social network features in a broad spectrum prediction-based system raises the issue of access to data. Although it is possible to relax this issue, by obtaining a sample of data which is used to derive and test prediction models, its extensive use would virtually require full access to real-time data. Unfortunately, it is highly unlikely that this option is available to anyone other than the data owners. Secondly, handling data on the connections that users establish with other users, as well as harnessing and using their interactions, raises issues related to privacy [91].

It is possible that the distinction between *a priori* and *a posteriori* prediction scenarios made in previous formalizations of the popularity prediction task could undermine the potential of both types of approaches. As stated before, the *a priori* prediction models are focused on predicting the popularity of web content items before their publication, when no social feedback from users is available. It also includes situations when the access to the evolution of popularity is not available. As for *a posteriori* prediction models, these are essentially focused on modelling the behaviour of popularity in order to forecast the future popularity. However, no previous approach has studied the combination of these two approaches to the problem of popularity prediction, specially when focusing on the first moments after the publication of web content items.

One of web content popularity's most distinct characteristics relates to its distribution. Barabasi [26] claims that increasing evidence shows that human behaviour is characterized by short bursts of high activity followed by long periods of inactivity, as a result of decision-based queuing processes, leading to power-law distributions. It should be noted that some work [209, 146, 118] has presented evidence contrary to this claim, by proposing that the data is better fitted by a log-normal distribution or the superposition of two log-normal distributions. However, most authors have provided evidence of web content popularity describing a power-law. Simkin and Roychowdhury [208] verify this claim through a thorough experimental evaluation process and Easley and Kleinberg [56] associate this characteristic to the "rich-get-richer" effect, where items that become popular have a higher probability of becoming even more popular because they are promoted. Previous work has denoted the high difficulty in predicting such cases of highly popular web content items [121, 123, 87, 206, 15].

This characteristic implies that the large majority of web content items gathers a low level of popularity, in contrast to a small set of rare web content items that obtain very high popularity levels. Standard learning tools and most evaluation metrics are focused on

predicting and evaluating the average behaviour of data. As such, a problem surges in the context of popularity prediction as items are not equally important, i.e. from a set of many web content items, one wants to recommend to users the most popular ones. To exemplify this issue, consider a classification task of predicting the popularity of web content, where 95% of the items have normal levels of popularity and only 5% of the items have high levels of popularity. In this case, the average item has a low level of popularity. As such, if a standard learning algorithm would predict all cases as having a low level of popularity, the widely used evaluation metric Accuracy would present an excellent result of 0.95. Since the items are not equally important, using this metric will inflate the evaluation, failing to convey that none of the highly popular items (those which we want to suggest) were correctly predicted. In classification tasks and others with nominal target variables, this issue can be more appropriately evaluated using metrics such as the F-Score or ROC curves.

Tasks with numerical target variables are equally prone to such learning and evaluation issues. To describe such issues, a prediction scenario is exemplified[15] in Table 2.3 and depicted in Figure 2.1 using synthetically generated data. In this scenario, two models (M_1 and M_2) provide their respective sets of predictions.

Table 2.3: Predictions made by two artificial models M_1 and M_2 with their respective error and the ground-truth values.

Predictions of Two Artificial Models										
True	2.71	3.35	3.36	3.63	4.08	4.16	4.31	5.55	5.78	6.40
M_1	2.67	3.29	3.43	3.73	3.97	4.28	4.54	5.91	7.03	4.72
Loss M_1	0.04	0.06	0.03	0.04	0.11	0.12	0.23	0.64	1.45	1.68
M_2	1.03	4.59	3.74	3.88	4.20	4.03	4.42	5.59	5.74	6.37
Loss M_2	1.68	1.24	0.72	0.45	0.88	0.77	0.89	0.04	0.04	0.03

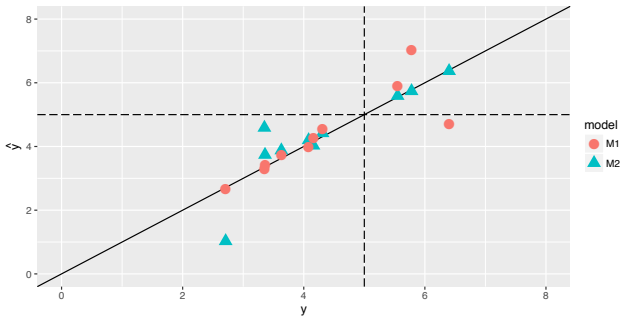


Figure 2.1: Misleading scenario for standard error metrics in a regression task, using artificial data.

Figure 2.1 shows that model M_1 obtains a superior predictive accuracy at low values of the data and that model M_2 is far more accurate at the highest values. However, if standard metrics such as Mean Squared Error and Mean Absolute Error[16] (*MSE* and *MAE*,

[15] This example is based on the scenario proposed in the work of Ribeiro [196].
[16] Also known as Mean Absolute Deviation (**MAD**).

respectively) are calculated, no difference between these two models is found: both models obtain a score of 0.461 for *MSE* and a score for *MAE* of 0.397. This occurs because these metrics are unable to distinguish where (in the domain of the target variable) the errors occur.

Considering the thorough review of previous work presented in this chapter, it is observed that evaluation metrics used in previous work concerning numerical prediction tasks are prone to this problem. Notwithstanding, proposals have been presented in order to deal with learning and evaluation scenarios such as those of web content popularity prediction. Based on the concept of utility, Ribeiro [196] proposes the formalization of such learning tasks as utility-based regression tasks, and describes appropriate evaluation metrics to account for non-uniform domain preferences of users.

Regarding the experimental methodology used in previous work, we observe that the proposed approaches were mainly validated using either the out-of-sample method [213] or k-fold cross validation [93]. The former consists of dividing the available data in two parts, train and test sets, which will be used to build the models and evaluate its performance, respectively. The latter partitions the data in k parts, and iteratively chooses one to predict, the remaining to learn the models, and validation results are averaged. As to the definition of k, previous work mainly defined it as $k = 5$ (e.g. [124, 23, 207]) and $k = 10$ (e.g. [140, 239, 118, 225, 217]).

Given the scope of our problem and the context of the data (i.e. temporal order), several issues concerning the usage of such experimental methodologies may be raised. Previous work has shown that results obtained using the out-of-sample method may lead to a unreliable measure of the models' quality [213]. The main caveat is its limitations in providing an optimal approximation of the error. In contrast, k-fold cross validation allows for the convergence of the validation scores, although at a higher computational cost. However, when the data is dependent, the validation of results using k-fold cross validation also raises caveats [16], since this methodology assumes that train and test sets are independent. In the case of dependent data, such as that of our scope where a temporal dependency exists, the use of such strategy may not provide an adequate simulation. Notwithstanding, other proposed methodologies are capable of tackling this caveat, such as forward validation [102] and multiple period out-of-sample tests [213]. The former guarantees that the test set will always be ahead of the train set, increasing incrementally the size of the train set. The latter can be loosely described as an application of the out-of-sample method on k samples of the entire data set, therefore maintaining the temporal order as well.

Finally, concerning ranking tasks, one of the characteristics of web content is that it is multi-source [152]. As such, it is possible for one given item to be published in several social media platforms, and therefore have multiple scores of popularity. In previous work, a set of approaches focused on this characteristic [179, 200, 41]. Their objective is to predict

the popularity of a given web content in a given social media source, using behavioural features of the same web content in other social media sources. The encouraging results obtained by the authors shows this connectivity between web content items in multiple sources. However, we should note that precisely due to the connection of web contents in several sources, using behavioural features from external sources in order to tackle the task of predicting the popularity of contents in a third-party source, considerably relaxes the prediction task. Despite the former, this multi-source characteristic of web content raises an important related research question, of how to rank items that have more than one objective target variable. Regardless of the extensive review presented in this chapter, this issue seems to have not been considered in previous work.

2.6 Conclusions

The discussion provided in the previous section describes the issues and questions raised by previous work. In addition, it also outlines the main focus points of the contributions in this thesis. These are described as follows.

- The main problem addressed is that of accurately predicting and ranking rare cases of highly popular web content items, by tackling the problems raised by standard prediction models and evaluation metrics. As such, this work is focused on numerical prediction tasks. Concerning the formalization of prediction tasks, the work of Ribeiro [196] concerning utility-based regression is leveraged.
- New approaches for both *a priori* and *a posteriori* prediction are proposed, and a new predictive strategy is proposed based on dynamic ensembles.
- Concerning evaluation, a robust framework is proposed. The objective of this framework is to allow a multi-level overview of the models' predictive accuracy and a comparison between error-based and utility-based evaluation metrics.
- Finally, the problem of multi-source ranking is addressed. This includes the proposal and evaluation of combinatorial methods concerning multiple feeds of popularity scores, taking into consideration the issue of imbalanced data.

In the following chapter, the characteristics of social media data are studied in depth. For this purpose, two novel data sets concerning online news feeds are presented and an exploratory analysis is detailed. This analysis motivates the aforementioned contributions, which are described in subsequent chapters.

Chapter 3

The Case of Online News Feeds

In this chapter the case of online news feeds is presented. A detailed description of its characteristics is provided as well as the motivation for using this type of web content in the experimental evaluations of the thesis' contributions. Two new data sets are introduced allowing for the analysis of scenarios involving both single and multiple official and social media sources. An exploratory analysis is performed in order to further discuss previous works' conclusions and new insights.

3.1 Introduction

With the evolution of technology and the increasing coverage of Internet access, news outlets, and journalism in general, have faced several problems. Apart from the obvious economic impact, with the noticeable downfall of physical newspapers' sales [165], this has had a great impact on traditional news distribution. Nowadays, most newspapers thrive on their online presence and greatly depend on web advertisement revenues [160].

This transition from offline to online newspapers also provoked an enormous growth in volume and diversity of online news articles, coupled with increased easiness in access [188]. Due to the low cost of publishing online, it has also caused a great overlap of news stories, where media outlets attempt to cover as many events as possible, in the shortest time period. The combination of such factors creates a setting of information overload, where the volume of accessible information is so overwhelming that it causes an inability to make decisions or to fully understand the issues [60].

These factors, in addition to the need of managing the galloping demand of news from users, presents several problems to decision-makers in news rooms. Namely, the ability to act quickly on breaking news. These lead to consecutive updates (new stories) requiring an almost constant readiness to react and to decide on which news to promote and advertise. On the other hand, it also raises concerns regarding the consumption of news by the users:

given the high volume of news stories divulged every day, spanning a variety of topics, the effort required by users to find the items most relevant to them is a major issue [60].

Addressing user demand, but also in order to ease the effort of going through multiple online newspapers, news recommender systems have become very popular. These serve as aggregation sites for news of multiple outlets, where news rankings are suggested to users, in order to facilitate the search of relevant items. To a considerable degree, the evolution and success of such platforms is due to research on collaborative filtering approaches [1].

Collaborative filtering leverages data concerning the interaction of users and items, allowing to derive a deep understanding of trends and similarities between both entities. However, these approaches are prone to several issues. Namely, when considering the high throughput of news and the dynamics of the topics, these systems are required to find relevant news early. This is difficult, as collaborative filtering requires data on related user-behaviour and recent items do not have such data. In addition, previous work suggests that traditional information retrieval and web page ranking approaches have demonstrate several shortcomings in identifying and recommending the most relevant information to users [59].

A comprehensive list of the challenges that recommendation systems face in this context of news feeds is described by Özgöbek et al. [262]. These include the cold-start problem of not having data related to a given item, and the recency problem where the available data is not sufficient for accurate recommendations. In this context, researchers have focused on tackling the task of predicting the popularity of news in social media. The cold-start problem is related to *a priori* prediction (without related social feedback), and the recency problem is mainly related to *a posteriori* prediction (when available social feedback is insufficient for accurate predictions), both extensively reviewed in the previous chapter. In both prediction tasks, the main assumption is that the data presents a realistic sample of the interest relation between users and news [165].

A recent international study [176] on news consumption has shown that half (51%) of the Internet users use social media as a source of news each week and that 24% of users share news in social media on a weekly basis. This relates to the large amount of news-related queries submitted by users, providing evidence that a great amount of activity in social media is a response to news events [170]. As an example, Kwak et al. [134] show that nearly 85% of Twitter posts are related to news.

In addition, news items have a very short lifetime in terms of capturing the attention of users, when compared to other types of web content, e.g. videos. Specifically, previous work has shown that news may receive the attention of users up until two [60] or four days [246] after they are published. This fast fading characteristic can be associated to habituation or the diversion of attention towards other news [242], as previously referenced, which in turn is manifested in the suggestions provided to users [208]. This amounts to the common

property of web content, where their popularity in social media is described by a heavy-tail distribution: most of the items are relatively unpopular, and a small set of rare cases are highly popular.

This exacerbates the popularity prediction task using news-related data, requiring highly reactive predictions, and the ability for them to be highly accurate, specially when concerning rare cases of highly popular news. The success of such prediction tasks can lead to considerable improvements in the quality of news recommendations. Namely, accurate and early news popularity predictions allow for more reactive suggestions of news, to improve the quality of popular items' suggestions [215] and faster discovery of relevant news [256]. Considering the plethora of previously proposed approaches concerning popularity prediction, it clearly shows the importance of solving this prediction task and how it continues to be an important goal, not only for users, but also for media professionals [12].

3.2 Data Sets

In this section two novel data sets are presented[1], which will be used in the subsequent chapters for experimental evaluations.

The goal of popularity prediction tasks using online news feeds' data is commonly to anticipate the popularity of a given news story in social media. News data can be obtained from two types of sources: *i)* official media, and *ii)* social media sources. The first type of source, official media, relates to the origin of news items and their original content. It may also provide an indication of the items' relevance according to the official media source, denoted by its ranking position. The second, social media, is the medium used to measure the attention received by the news items, i.e. popularity. In previous work, the provenance of the data used can be framed in one of three settings: *i)* solely using official media sources (*e.g.* [242, 142, 212]), *ii)* solely using social media sources (*e.g.* [207, 186]), or *iii)* using data from both official and social media sources (*e.g.* [141, 104, 60]).

Official media sources include legacy media[2] outlets (*e.g.* The Washington Post), news aggregation platforms (*e.g.* Digg, Slashdot) and news recommender systems (*e.g.* Google News, Yahoo! News). Most of previous work solely concerning official sources was focused on the second, with emphasis on the Digg platform. However, since one of the objectives of this thesis is to analyse the ranking ability of the prediction models proposed, the focus of this work is on the third type of official sources: news recommender systems. This choice has the benefits of *i)* providing a potentially large list of items, ordered by the relevance attributed by the system, and of *ii)* presenting a multitude of news outlets, including those

[1] Both data sets are available for download at `www.dcc.fc.up.pt/~nmoniz/Thesis/FullDataSets.zip`.
[2] Common expression to denote traditional media outlets.

that are not considered as being legacy media.

Regarding social media sources of news data, research shows that the micro-blogging platform Twitter covers almost all news-wire events but the opposite is not true [186]. Furthermore, Osborne and Dredze [181] have shown that Twitter is the preferred medium for breaking news, almost consistently leading Facebook or Google+. However, this claim has since grown out of date, as it has been shown [181] that Facebook is now the leading platform in accessing and sharing news, followed by Twitter. Newman et al. [176] provide a survey of over 50,000 people in 26 countries concerning news consumption. It shows that the most popular platform is Facebook, followed by Youtube, Twitter, Whatsapp, Google+, LinkedIn and Instagram. The data sets presented in this section include these sources with the exceptions of Youtube, Whatsapp and Instagram. The reason for this is that such platforms either do not allow the sharing of news items (although allowing it in comments) as in the case of Youtube and Instagram, or are essentially focused on direct messaging, significantly reducing the dynamics of popularity and social spread.

In order to allow the analysis and discussion of the interplay between these two types of data sources, the novel data sets use data from both types of sources: official and social media. The information portrayed in such sources is used and interpreted differently. From official data sources descriptors of news items are extracted, as well as their respective relevance according to each source, illustrated by their rank. As for social media sources, these are used to obtain the popularity of news, which may be denoted by different signals in each source.

The main difference between the two proposed data sets is the amount of official and social media sources. The first is a single-source data set, using one official media source and one social media source, Google News and Twitter. The second is a multi-source data set, using two official media sources (Google News and Yahoo! News) and three social media sources (Facebook, Google+ and LinkedIn). It should be clearly noted that not using Twitter data in this second data set was not a choice, but indeed due to the deprecation of its API functionality allowing for the extraction of such data[3].

Both data sets contain news-related data concerning four topics: *economy, microsoft, obama,* and *palestine*. These topics were chosen due to two factors: their worldwide popularity and the fact that they report to different types of entities (sector, company, person, and country, respectively). Notwithstanding, this choice raises some caveats. Being limited to a small number of topics might undermine conclusions concerning the ability to generalize the predictive ability reported in experiments. On the other hand, it does provide a deep insight into topics that have a daily activity of high magnitude, in opposition to having a news sample of a variety of topics. By approaching the development of data sets as proposed, possible

[3]Twitter API: `https://dev.twitter.com/docs/api`. The *count* method was deprecated on 20^{th} of November, 2015.

problems related to context-sensitivity (*i.e.* topic) characteristics of text-based tasks [31] are also tackled.

Concerning the prediction horizon, previous work has differed on the active time that one should consider to news items, since their publication. For example, alternative values include two days [60, 135], four days [246, 24] and 30 days [216]. However, based on the analysis provided by these works, most of the popularity dynamics of news develop in the first day, although showing that in some cases it may develop for more time. Nonetheless, in the latter cases, the increase is very residual in terms of proportion. Given this, an active time for the news items of two days is established, since their publication. This means that the popularity evolution of each item is stored for this period, since their appearance in official media sources.

Finally, it should be noted that a stochastic view of the popularity concept is assumed (i.e. aggregate behaviour [216]), considering all publications from every user equally, which are used as input in the prediction tasks. Different approaches have been proposed, as previously discussed, such as those focused on determining the number of "retweets" (Twitter functionality to re-publish a tweet) a given tweet will obtain [210, 254, 104] or those using data concerning the social network of individual users to predict the popularity of content they generate [95]. The data composed in these data sets is not focused on the popularity of content generated by a single or a given group of users, but on the general popularity of content in social media platforms, allowing for a source-wide perception of news stories' popularity.

3.2.1 Single-Source Data Set

The news recommender system Google News was queried during a period spanning roughly seven months (March 20^{th}, 2015 until October 23^{rd} 2015), in 20 minute intervals, for each of the four topics. For each query, the top-100 recommended news were collected. For each news recommended by Google News the following information was registered: title, headline, publication date, the news outlet and its position in the ranking. Figure 3.1 shows the total number of news per topic during this period (left) and a smoothed approximation of the amount of news per day for each topic (right).

Upon the retrieval of information from Google News, the popularity of all news items with an alive time below the defined period of two days, was obtained from Twitter. To obtain that information the Twitter API[4] was queried, using the *count* method. This method allows to check the number of times the news URL was published following its publication. Of the total number of news for all topics (81,469), in 4,622 cases (5.7%) it was not possible to obtain the number of tweets and in 10,951 cases (13.4%) the news items were not tweeted. Figure 3.2

[4]Twitter API: `https://dev.twitter.com/docs/api`

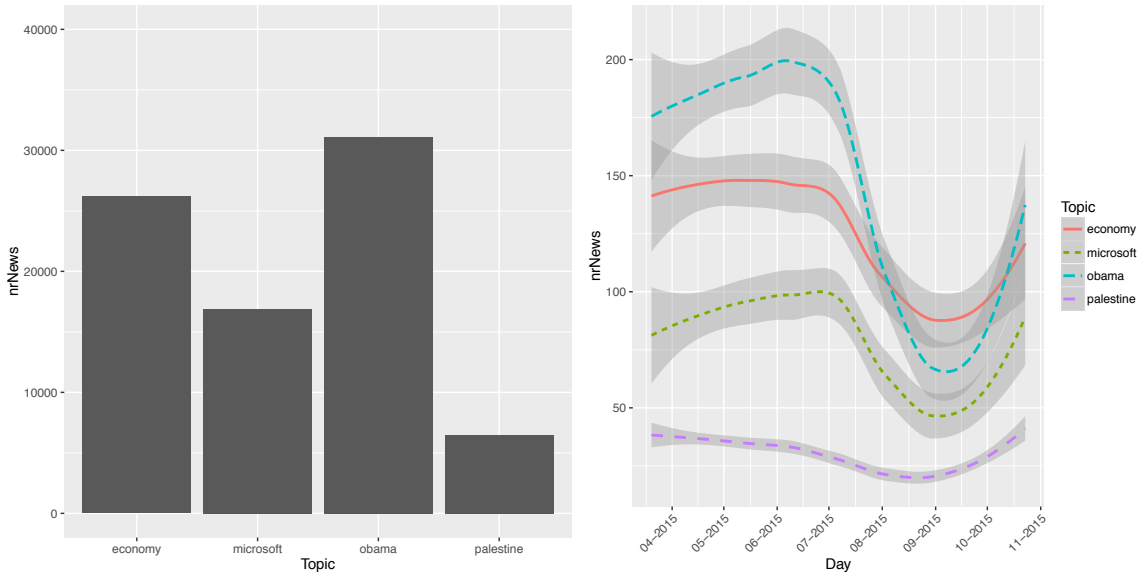


Figure 3.1: Single Source Data Set: Number of news per topic (left) and a smoothed approximation of the amount of news per day for each topic.

shows the distribution of popularity in all topics, estimated by the number of tweets (in logarithmic scale), for the news that were published in Twitter. For understandability purposes, the illustration is limited to 100 publications, although it should be noted that the remainder of the data abides by the same pattern of decay, seemingly following a heavy tail distribution, in accordance with previous work (e.g. [215, 118]).

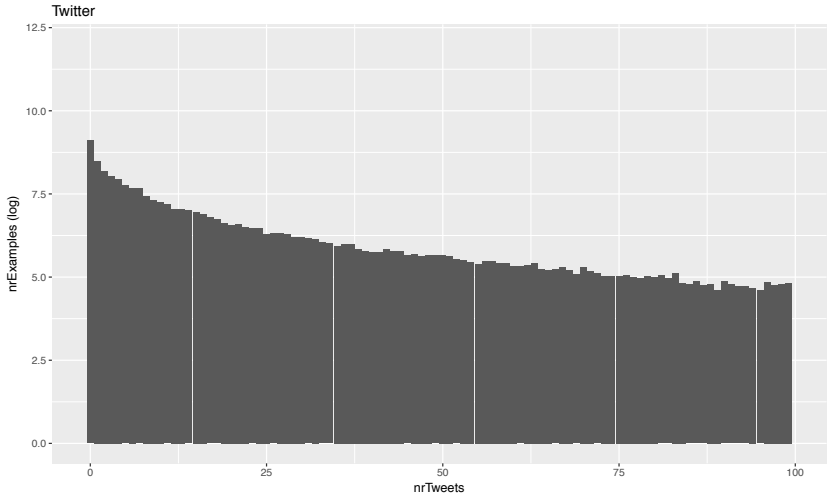


Figure 3.2: Single Source Data Set: Distribution of news popularity in Twitter in a logarithmic scale, limited to 100 publications.

3.2.2 Multi-Source Data Set

The process to build the second data set is similar to the former. However, in order to accommodate the information retrieved from multiple official and social media sources some processes were parallelized. The process is described as follows. The official media sources Google News and Yahoo! News were queried, during a period of approximately eight months (November 10^{th}, 2015 until July 7^{th} 2016), for each of the four topics (*economy*, *microsoft*, *obama* and *palestine*). The queries were done simultaneously, in 20 minute intervals. For each query, the top-100 recommended news of the respective official media sources were collected. As in the previous data set, for each news recommended, the following information was collected: title, headline, publication date, the news outlet and its position in the ranking.

In order to deal with duplicated cases, the following procedure was applied. First, cases with the same title, headline and from the same news outlet, are grouped. Second, the oldest case is kept and the remaining are removed from the data set. Finally, the identifiers of the removed cases are replaced by the case that was kept. Figure 3.3 shows the global number of news per topic during the retrieval period (left) and a smoothed approximation of the amount of news per day for each topic (right). The total amount of news retrieved in both official media sources is 93,239.

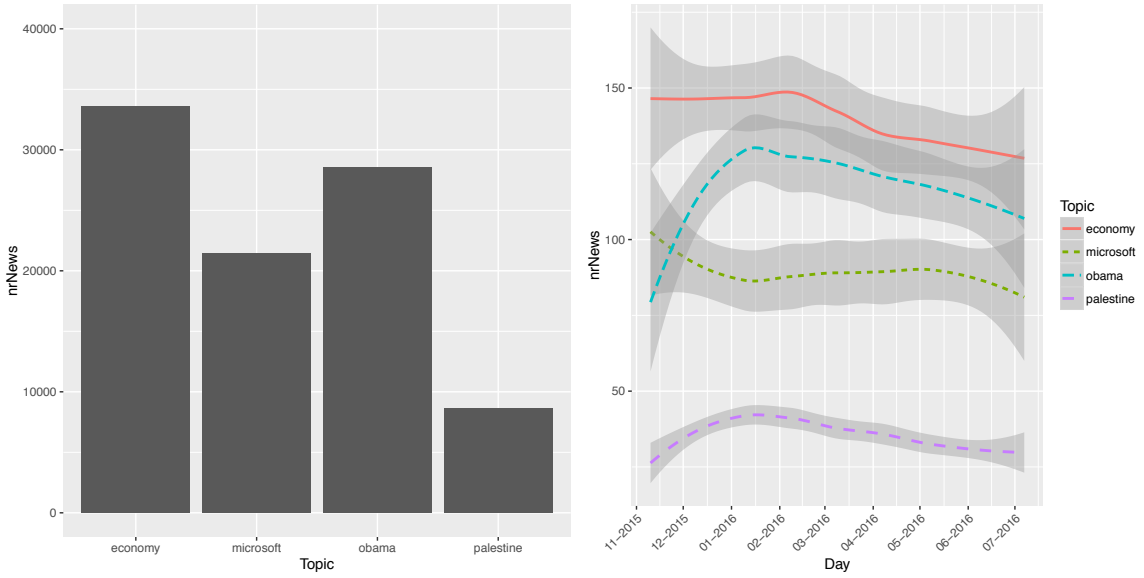


Figure 3.3: Multi Source Data Set: Number of news from both Google News and Yahoo! News (left) and a smoothed approximation of the amount of news per day, for each topic.

The contribution of each official media source to the total amount of news is not equal. In Figure 3.4 a Venn diagram is depicted, illustrating the intersection of both news sets. As shown in the figure, the official media source Google News provides the majority of items (71,481) in comparison to Yahoo! News (24,474). Additionally, data shows that there is only

a small set of 2,702 news items (2.9%) that are used by both sources.

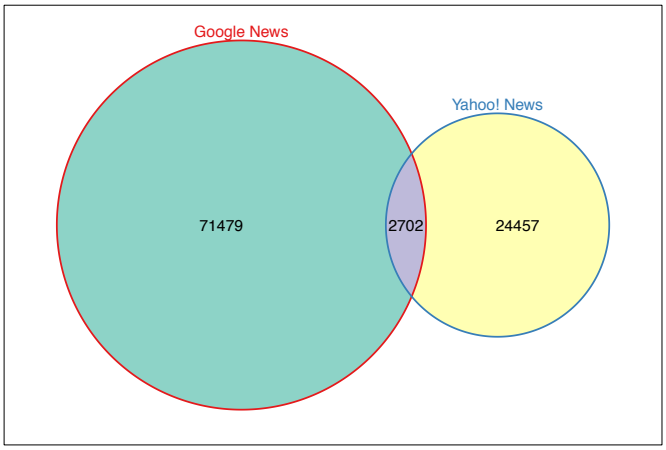


Figure 3.4: Venn diagram of published news items in official media sources.

After retrieving the query-data from the official sources, the popularity of all known news items, with an alive time below the defined period of two days, is obtained from the social media sources Facebook, Google+ and LinkedIn, simultaneously. Considering the differences between each social media source, the procedures for obtaining popularity are different. They are described as follows.

- For obtaining information from Facebook, the Facebook Graph API[5] is used, by querying for information concerning the URL of each news item. The data retrieved reports the number of "likes", shares, clicks, comments and the total number of interactions concerning each unique URL. For consistency, the number of shares is used as the main popularity measure. In 11,602 cases (12.4%) it was not possible to obtain the number of shares, and in 26,919 cases (28.9%) the news items' were not shared on Facebook;

- The social media platform Google+ does not allow to obtain the number of shares of a given URL. Nonetheless, it allows one to check the number of times users have "liked" the URL's. Despite the differences with other social media sources, it is nonetheless a valid metric of received attention by news stories. This process is carried out by querying a public end-point[6] in order to obtain the amount of "+1" (similar to "like" in Facebook) a given URL received. In 5,744 cases (6.2%) it was not possible to obtain

[5]The Graph API: `https://developers.facebook.com/docs/graph-api`

[6]The number of "+1" received by a given URL in Google+ is obtained by appending the respective URL to `https://plusone.google.com/_/+1/fastbutton?url=`.

the number of "+1", and in 55,114 cases (59.1%) the news items' did not obtain any "+1";

- Finally, concerning the LinkedIn platform, the number of times each news story URL was shared is obtained by querying its public end-point[7], designed for such purposes. Concerning the overall statistics of the news items' presence in the platform, in 5,745 cases (6.2%) it was not possible to obtain information concerning the news item URL. In addition, in 54,413 cases (58.4%) the news were not shared on the LinkedIn platform.

Figure 3.5 shows the distribution of popularity for the news that were published in all the social media sources used in this multi-source data set, in a logarithmic scale (for understandibility purposes). The illustration is limited to 100 publications, noting that the distribution of the data follows the same trend, consistently, in each of the sources, until the respective maximum of popularity. According to the data collected, the maximum popularity obtained by a news item in Facebook (shares) is 49,211, in Google+ ("+1") 1,267, and in LinkedIn (shares) 20,341.

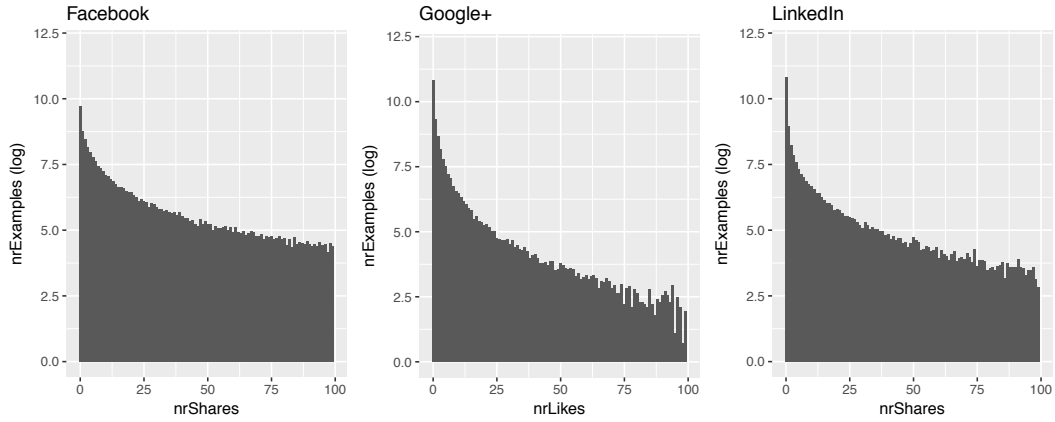


Figure 3.5: Multi Source Data Set: Distribution of news popularity in Facebook, Google+ and LinkedIn, limited to 100 publications.

Similarly to the case of official media sources, in Figure 3.6 the intersection of the sets of news that were published in each of the social media sources is illustrated. Results show that most of the news published in Google+ or LinkedIn are also published in Facebook (91.7% and 88.7% respectively). Conversely, data shows that roughly a third (36%) of news published in Facebook are also published in both Google+ and LinkedIn. Additionally, only 2,006 (7.9%) news stories were published in both Google+ and LinkedIn, and not on Facebook.

[7]The number of times a given URL was shared in LinkedIn is obtained by appending the respective URL to `https://www.linkedin.com/countserv/count/share?format=json&url=`.

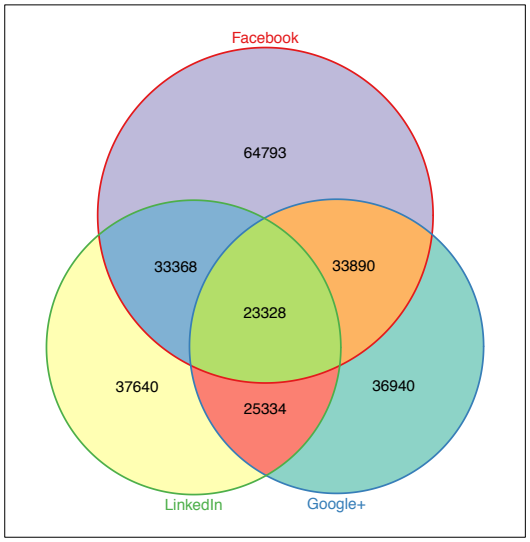


Figure 3.6: Venn diagram of published news items in social media sources.

3.3 Analysis

In this section an extensive study of the characteristics of the data sets described in this chapter is presented. The study concerns various aspects of the data, such as: *i)* temporal dynamics of popularity, *ii)* the agreement between sources, *iii)* the distribution of popularity, *iv)* the presence of sentiment words, *v)* impact of named entities and *vi)* news outlets, and *vii)* the contribution of temporal data. Considering these aspects of the data, ten analysis questions are formulated and addressed:

Q1. How does popularity evolve with time? Considering the importance of accurate predictions soon after a news item is published, it is crucial to understand the temporal dynamics of the popularity of such items. Using data from both data sets, the evolution of popularity according to the available social media sources is analysed.

Q2. Do official media sources agree? Using data from the multi-source data set, the recommendations of both Google News and Yahoo! News are analysed. The objective is to understand the degree of overlap amongst these official media sources. Understanding their intersection allows the derivation of insights regarding the diversity or superposition of news recommendations.

Q3. Do social media sources agree? Also using data from the multi-source data set, given the availability of data from three different social media sources, a study is carried out in order to understand which news are more popular according to each social media source. This allows to understand if popularity is a transverse concept along various sources. If not, do they show clear patterns of popularity evolution amongst them?

Q4. Do official media and social media sources agree? Building on the two former questions, the possibility of agreement between news recommendations from official media sources with the suggestions provided by the aggregate behaviour of users in social media sources is explored.

Q5. Is news popularity in social media sources best described by a power-law distribution? Despite the widespread claim that the popularity of web content is described by a power-law distribution [215, 208, 56], this has been previously contested. Some authors propose that the popularity of web content is better fitted by a log-normal distribution. The objective is to study the popularity distribution of all social media sources available in the presented data sets, in order to confirm whether the distribution of news popularity is best described by a power-law distribution.

Q6. How does sentiment analysis of news relate to their popularity? Previous work has shown the positive impact of using sentiment analysis as features in prediction models [18]. The objective is to understand the existence and magnitude of the relation between sentiment in news and the magnitude of their popularity.

Q7. Are highly popular news related to mentions of popular named entities? As in the previous question, the objective is to discover the relation between entities mentioned in news and the degree of popularity the news items obtain.

Q8. Are popular news published by popular news outlets? The outlet of news items has been used in previous work [24], claiming that such a predictor greatly contributes to the prediction of popularity. The objective is to ascertain if news outlets are good indicators of news popularity.

Q9. Does the publication hour influence news popularity? Concerning the temporal aspect, previous work has studied the circadian pattern of user activity and its impact in the evolution of web content's popularity [124]. Here the objective is to study the commonalities and disparities within the circadian pattern of news in multiple social media sources, w.r.t. to their publication hour.

Q10. Does the publication weekday influence news popularity? The previous study is repeated, focusing on the publishing weekday of news items.

Dynamics of Popularity

The first question addressed in this analysis of the presented data sets concerns the evolution of popularity (**Q1**). One of the main caveats concerning the prediction of web content popularity relates to recency. This occurs when content is very recent, and the related social feedback provided by users is nonexistent or insufficient to enable accurate predictions. As

such, it is important to understand how popularity evolves. Additionally, given that the data sets presented include several social media sources, it is also important to understand if such dynamics are similar amongst different sources.

With this objective, the evolution of popularity in each of the topics available in the data sets is studied, as well as in each of the social media sources: Twitter, using the single-source data set, and Facebook, Google+ and LinkedIn using the multi-source data set. Figure 3.7 illustrates the evolution of popularity for each of the query periods in the data sets, referred to as time slices. It should be reminded that the query periods/time slices in both the single-source and the multi-source data are set to 20 minutes. For example, the first time slice t_1 refers to alive-time period of a news item between 0 and 20 minutes, and the third time slice to the period between 40 and 60 minutes. Given the prediction horizon of two days, the final time slice (t_f) is 144[8].

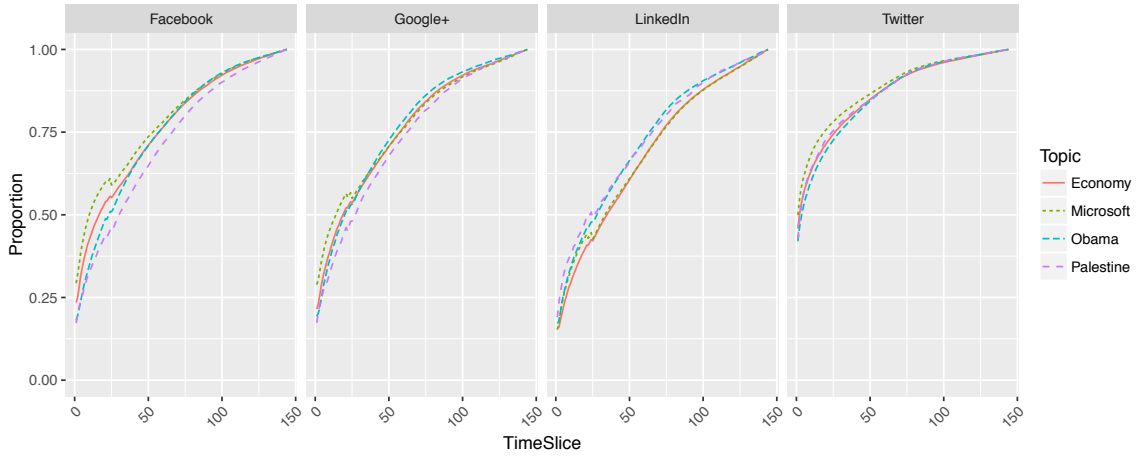


Figure 3.7: Evolution of popularity (as proportion of final popularity) in each topic, for social media sources of both data sets. Each time slice represents a 20 minute period.

Results in Figure 3.7 are depicted as the average proportion between the popularity scores at each given time slice, and the respective final popularity scores. Based on this, it is observed that in most social media sources, news items obtain close to half of their final popularity in a short amount of time. However, it should be noted that the case of Twitter presents a different dynamic from all others. Although Facebook, Google+ and LinkedIn show that in the first moments after publication, news items quickly obtain on average close to 20% (22%, 21.8% and 16.8%, respectively) of their final popularity, in the case of Twitter this proportion is close to 50% (45.1%). This shows that Twitter presents a more reactive response to news items than other social media sources.

[8]144 time slices corresponds to 144 query periods of 20 minutes.

Agreement Between Sources

In this section the analysis of agreement between the various data sources (and types of sources) included in the data sets is addressed (**Q2**, **Q3** and **Q4**). The objective is three-fold: *i)* to understand the intersection of news in official media, *ii)* in social media, and *iii)* amongst both types of sources. To do so, the rankings proposed by the sources are analysed and compared.

To assess the level of agreements between official sources (**Q2**), the multi-source data set is used, as it contains data from both Google News and Yahoo! News. Formerly, when describing this data set, it was shown that the intersection of news between both official media sources is small (see Figure 3.4). Only in the case of 2,702 news did such items appear in rankings of both official sources. This represents approximately 3.8% of news items in Google News rankings, and 11.1% of those in Yahoo! News. Furthermore, the discrepancy between the amount of news suggested by both sources was pointed out (illustrated in Figure 3.4). Yahoo! News proposed 24,473 news items in the span of roughly 8 months, whilst Google News proposed 71,480 items.

Before proceeding with the study of agreement between official media sources, it is important to understand the reason for the discrepancy in amount of news between these two sources. As such, each of these official media sources is analysed, concerning the alive-time of news items in their respective news recommendations, illustrated in Figure 3.8.

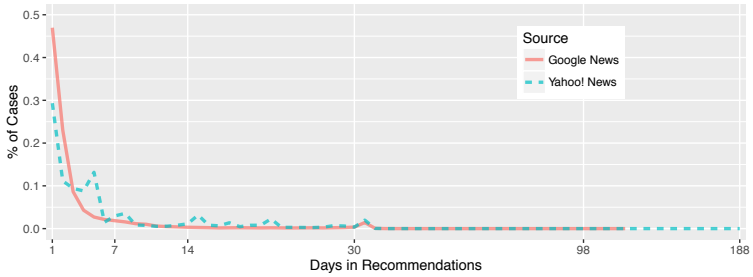


Figure 3.8: Distribution of the amount of days news items appear in official media sources' recommendations.

Results show a clear distinction between Google News and Yahoo! News concerning the alive-time of news items in their respective recommendations. With a higher throughput, results show that nearly 50% of news items in Google News are included in recommendations for a single day. Conversely, in Yahoo! News, this proportion is roughly a third (31.8%). Moreover, it is interesting to note that only 10% of news items in Google News are included in recommendations for a period of over a week. In the case of Yahoo! News, nearly a quarter (24.4%) of news items are included in recommendations for more than a week. Therefore, it may be concluded that the discrepancy in the amount of news included in recommendations

of both official sources is mainly related to the alive-time of each sources' recommendations: Google News has a higher throughput, and therefore, news will usually not remain long in their recommendations, whilst in Yahoo! News the data shows a tendency for news to remain in recommendations for a longer period of time.

Given the amount of news in both official media sources, it is clear that the overlap between sources is not significant. However, it is nonetheless important to understand on which type of news they tend to agree. Concretely, it is analysed if such shared news, despite their amount, are considered by both sources to be relevant news. The degree of relevance according to the sources is indicated by their position in the respective rankings. As such, an evaluation of the news rankings suggested by the sources is carried out, focusing on two perspectives: *i)* is the occurrence of news items in both official media sources constant, or is it dependent on to a given temporal window; and *ii)* are these news jointly considered to be relevant. In Figure 3.9, the distribution of news ranking positions for both official sources is depicted (left), as well as the their temporal dynamics (right).

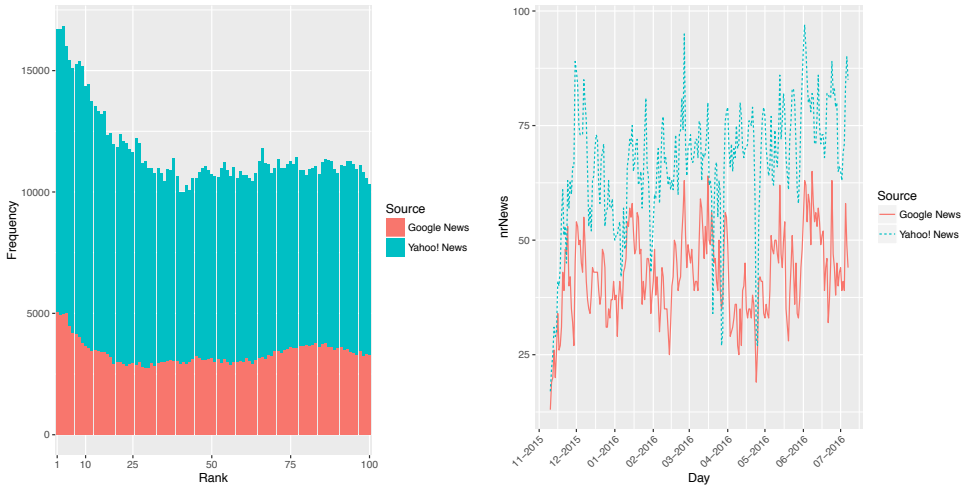


Figure 3.9: Distribution of news ranking positions in both official sources (left) and their temporal dynamics (right).

The results illustrated in Figure 3.9 show that news suggested in both official media sources are commonly considered by both sources as being relevant items. By observing the left side of the aforementioned figure, it is clearly observable that news suggested in both official media sources, are often jointly considered top suggestions. Also, considering the temporal dynamics of this set of news (right side of the figure), results show that the occurrence of news items suggested in each official media sources roughly follows the same tendencies.

Thus, given the results obtained, it is concluded that the official media sources Google News and Yahoo! News are mostly discordant, as can be argued primarily by the observation that the intersection of suggestions from both sources is minimal (2.7%). However, it should be noted that news suggested by both official media sources are tendentiously considered to be

top suggestions, and that the existence such news are not related to specific time periods.

Building on the former conclusions, question **Q3** of whether social media sources agree on the relevance of news items, is approached. To carry out this analysis, the following procedure was implemented. Using the multi-source data set, the data was partitioned according to both the day of news publication and its respective topic. Using this information, daily sub-sets of news were derived in order to obtain news rankings for each of them. For each combination of topic and day of publication, three rankings were derived, considering each of the social media sources available (Facebook, Google+ and LinkedIn). Iteratively, the rankings of each source was considered to be the ground-truth and an evaluation of the remainder sources was carried out. The evaluation resorts to the metric Normalized Discounted Cumulative Gain ($NDCG@k$), defined in Section 2.4, considered to be a robust ranking evaluation metric [203]. This metric requires the definition of cut-off ranking positions. Three were used: 10, 25 and 50. The metric also requires the setting of degrees of relevance for ranking positions intervals. As such, a scale of relevance based on the ground-truth ranking positions was defined, described as follows: items in the top-10 have a relevance of 3, the remaining positions in the top-25 a relevance of 2, a relevance of 1 is attributed to other items in the top-50 and the remainder have a relevance of 0. Results are described in Table 3.1.

Table 3.1: Evaluation of daily news rankings from social media sources using $NDCG@k$ (with 10, 25 and 50 as k values), for each topic. Results in bold denote the best scores for a given baseline source.

		Facebook			**Google+**			**LinkedIn**		
Topic	**Baseline**	@10	@25	@50	@10	@25	@50	@10	@25	@50
Economy	Facebook	•			**0.590**	**0.650**	**0.738**	0.492	0.569	0.668
	Google+	**0.629**	**0.652**	**0.722**	•			0.594	0.643	0.716
	LinkedIn	0.529	0.571	0.656	**0.594**	**0.645**	**0.730**	•		
Microsoft	Facebook	•			**0.633**	**0.714**	**0.765**	0.569	0.657	0.710
	Google+	**0.678**	**0.721**	**0.756**	•			0.632	0.693	0.734
	LinkedIn	0.593	0.660	0.709	**0.606**	**0.688**	**0.742**	•		
Obama	Facebook	•			**0.685**	**0.747**	**0.829**	0.523	0.575	0.692
	Google+	**0.699**	**0.747**	**0.819**	•			0.494	0.570	0.673
	LinkedIn	**0.506**	**0.563**	**0.654**	0.489	0.555	0.652	•		
Palestine	Facebook	•			**0.775**	**0.845**	**0.869**	0.650	0.747	0.771
	Google+	**0.806**	**0.844**	**0.874**	•			0.654	0.739	0.771
	LinkedIn	**0.690**	**0.762**	0.783	0.679	0.758	**0.793**	•		

In each of the topics, results obtained with the evaluation metric $NDCG@k$ for different cut-off ranking positions, show that all three sources show a considerable degree of agreement. Nonetheless, it is observed that some of the sources are more similar to others: Facebook and Google+ share a high degree of similarity amongst themselves, in comparison to LinkedIn; and Facebook and Google+ show different degrees of similarity with LinkedIn depending on

the topic.

Finally, the similarity of recommendations from official and social media sources is analysed (**Q4**). This issue has been previously addressed [59], presenting evidence that would suggest, in this case, that official and social media sources do not agree well. Nonetheless, in order to test this conclusion and to provide further insights, an analysis of the data in both the multi-source and single-source data sets described in this chapter is performed. The procedure is as follows. For each ranking proposed by Google News and Yahoo! News, an evaluation using the $NDCG@k$ metric is performed, using each of the social media sources as ground truth. In this case, the focus is on the top 10 positions of the ranking ($NDCG@10$), since those are the most likely to be suggested to users.

However, unlike the analysis performed in the previous question (**Q3**), an issue may arise concerning the magnitude of popularity. As shown, news items remain for a given period of time in the proposed rankings. As such, using their final popularity may not provide a fair comparison between the different types of sources. Therefore, a linear decay factor is applied to the ground-truth popularity of social media sources in order to simulate the deprecation of news items' popularity. This decay factor relates to the alive-time of the news items, translated as the number of existing query periods (time slices). As previously noted, in both the single- and multi-source data sets this period is defined as 20 minutes, and the final time slice (t_f) if 144. The decay factor $d(t_i)$ at a given time slice t_i is then defined as:

$$d(t_i) = 1 - \frac{t_i - 1}{t_f}, \tag{3.1}$$

and the popularity of a given news item is the product of its final popularity and a decay factor $d(t)$. Results of the aforementioned analysis are described in Table 3.2.

Table 3.2: Evaluation of news rankings from official media sources with social media sources' data as baseline, for each topic, measured by $NDCG@10$.

Topic	**Official/Social**	Facebook	Google+	LinkedIn	Twitter
Economy	Google News	0.299	0.401	0.400	0.433
	Yahoo! News	0.342	0.267	0.329	
Microsoft	Google News	0.612	0.687	0.625	0.713
	Yahoo! News	0.326	0.295	0.345	
Obama	Google News	0.493	0.502	0.490	0.526
	Yahoo! News	0.474	0.335	0.342	
Palestine	Google News	0.601	0.528	0.387	0.662
	Yahoo! News	0.346	0.228	0.180	

Results show a different conclusion when compared to the former analysis of agreement amongst social sources. Concerning the multi-source data set, the outcome of this analysis shows that official and social media sources are mostly discordant. However, it should be noted that the case of the official source Google News and the topic *mirosoft* is an exception since it obtained similar results to those of agreement amongst social sources. As for the

single-source data set, results show that the social media source Twitter and the official media source Google News have a considerable level of agreement.

Given the overall results, the degree of agreement between official and social media sources shown in this analysis points to the confirmation of the conclusions by DeChoudhury et al. [59]. The authors state that information retrieval and web page ranking approaches present several issues when identifying the most relevant content for end users, which is mostly confirmed by the results obtained. Nonetheless, it should be highlighted that Google News and Twitter show a distinct level of agreement when compared to the other official/-social media sources pairs.

Furthermore, the impact of using the decay factor should be noted. Although it arguably provides a fairer comparison, it is not based on any official procedure made public by official media sources.

Popularity Distribution

The distribution of web content popularity has been subjected to a wide variety of analysis, in previous works. Specifically concerning the case of online news feeds and the popularity of such items, conclusions on how to better describe their distribution are not consensual. Although authors agree that the distribution is better described as heavy-tail, there is disagreement in regards to the type of heavy-tail distribution that best fits web content popularity. On one hand, some argue that this distribution is best fitted with a log-normal distribution [209, 146], and on the other hand, other authors have described this distribution as a power-law [215, 118, 208, 56].

Addressing question **Q5**, the distribution of online news popularity is studied, considering several social media sources: Facebook, Google+ and LinkedIn from the multi-source data set, and Twitter from the single-source dataset. The objective is to provide evidence of whether or not the notion that online news' popularity is better described by a power-law distribution is correct, and if not, which distribution provides the best fit.

The procedure applied to study the distribution of online news popularity is as follows. A Kolmogorov-Smirnov goodness-of-fit test is applied, via a bootstrapping procedure with 100 simulations, to the popularity of news items according to each of the social media sources, following the guidelines of Clauset et al.[9] [50]. This procedure estimates the parameters of the power-law distribution - the minimum value for which the power law holds x_{min} and the exponent α of the power law, and the p-value for likelihood of the distribution fitting the power-law distribution.

However, it is still necessary to discard the possibility that no other distribution provides

[9]This procedure was carried out by using the framework available in the **R** package **poweRlaw**[94].

a better fit. As such, the goodness-of-fit of the popularity distribution is compared with two other types of distribution: log-normal and exponential. This comparison is carried out by using the value of x_{min} estimated for the power-law distribution, and by estimating the parameters for the comparisons with log-normal and exponential distributions. The parameters estimated for the former are the location μ and scale σ, and for the latter, the rate (or inverse scale) λ. Using this outcome, the test statistic R proposed by Vuong [231] is applied. The test statistic R is the sample average of the log-likelihood ratio, standardized by a consistent estimate of its standard deviation. The outcome is bounded by $[-1, 1]$, and its sign indicates the best fit: if positive, the baseline model is a better fit, and if negative, the opposite is concluded.

This procedure is applied to all the distributions of news popularity according to each of the social media sources in both the single- and multi-source data sets, and for each topic. Results are presented in Table 3.3 where the type of distribution providing the best fit is denoted in bold.

Table 3.3: Goodness-of-fit tests via bootstraping procedure as described by Clauset et al. [50]. The $p-$value evaluates the power-law goodness of fit and the test statistic R compares it to the log-linear and exponential distributions.

		Power-Law			**Log Linear**			**Exponential**	
Topic	**Source**	x_{min}	α	p	σ	μ	R	λ	R
Economy	Twitter	190	2.57	0.01	**2.35**	**1.61**	**-1.95**	0.004	3.54
	Facebook	573	2.47	0.35	**-2.05**	**2.58**	**-0.48**	0.001	2.17
	Google+	**45**	**2.92**	**0.9**	-1100.94	24.36	0.20	0.019	2.70
	LinkedIn	37	1.99	0	**1.75**	**1.94**	**-3.51**	0.007	11.44
Microsoft	Twitter	536	3.24	0.41	**4.65**	**1.05**	**-1.31**	0.003	1.17
	Facebook	84	2.17	0.09	**-7.76**	**3.46**	**-0.59**	0.004	5.44
	Google+	51	2.47	0.07	**1.50**	**1.60**	**-0.73**	0.013	3.26
	LinkedIn	**413**	**2.56**	**0.85**	-926.98	24.55	0.10	0.002	3.25
Obama	Twitter	452	2.69	0	**4.11**	**1.35**	**-2.68**	0.002	4.13
	Facebook	**8824**	**3.99**	**0.83**	-245.73	9.32	0.08	0.001	2.46
	Google+	137	3.87	0.47	**3.12**	**0.93**	**-0.58**	0.014	0.81
	LinkedIn	44	2.19	0	**1.18**	**1.88**	**-1.76**	0.009	4.46
Palestine	Twitter	143	2.41	0	**3.34**	**1.44**	**-1.57**	0.004	2.71
	Facebook	75	1.94	0	**2.66**	**1.94**	**-2.88**	0.003	5.31
	Google+	**41**	**3.16**	**0.63**	-37.45	4.50	0.24	0.028	2.38
	LinkedIn	**30**	**2.66**	**0.62**	-707.67	21.31	0.20	0.02	2.01

Results show that, in contradiction to the majority of previous works' conclusions, the log-linear distribution is predominant in providing a best fit for online news popularity distribution, over power-law. However, it should be stressed that the power-law distribution provides the best fit in some cases, and as such should not be disregarded. The exponential distribution shows the worst results, by not providing any advantage over the power-law distribution. To illustrate the differences in the goodness-of-fit of the various distributions, a set of cases described in Table 3.3 are depicted in Figure 3.10.

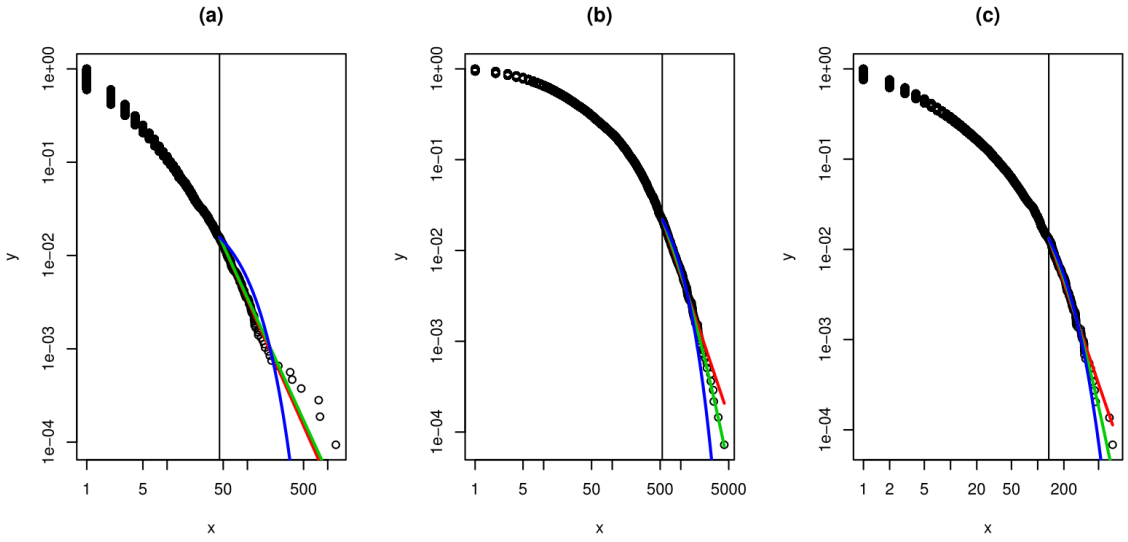


Figure 3.10: Goodness of fit for the power-law (red), log-linear (green) and the exponential (blue) distributions in three scenarios: using (a) data from LinkedIn in the topic "palestine", (b) data from Twitter in the topic "microsoft", and (c) data from Google+ in the topic "obama".

These three cases were chosen due to their different conclusions or uniqueness. The first case (a) reports the distribution of news' popularity of the topic "economy" according to Google+; the second case (b) uses the social media source Twitter, to illustrate news' popularity distribution of the topic "microsoft"; finally, the third case concerns news' popularity of the topic "obama" according to Google+. In the first case, the power-law distribution provides evidence as to being the best fit, and in the second and third case, the log-linear distribution is considered to be a best fit. The third case is featured in the figure due to it being the case where the exponential distribution provides its best results.

Sentiment Analysis

In *a priori* prediction tasks, models often extract information from the content of news (e.g. title, headline) in order to predict their popularity - one of the most popular features is derived using sentiment analysis.

Sentiment analysis, also known as opinion mining, is a field of text mining which studies people's opinions, judgments and ideas towards entities [151]. The process of sentiment analysis mainly relies on the detection of sentiment words (also known as opinion words or opinion-bearing words).

Sentiment is usually described by its polarity (positive, negative or neutral). Examples

of positive sentiment words are "*good*", "*excellent*" and "*amazing*". Examples of negative sentiment words are "*bad*", "*awful*" and "*horrible*". It is possible to represent it numerically, portraying different degrees of polarity strength. This numeric representation depicts a regression-like approach to the strength of the sentiment polarity. Therefore, sentiment words may be associated to a given polarity strength value in order to express their positive or negative polarity (1 and -1, respectively) or to portray levels of polarity strength (for example, between 5 and -5). This information is encapsulated in sentiment lexicons, providing the necessary structure for applying sentiment analysis.

Several approaches can be used to calculate the sentiment score of a given text. These range from the more naïve approaches (e.g. difference between the number of positive and negative words), to those more elaborate.

In this thesis, a more elaborate approach is applied, accounting for context words, the polarity strength of positive and negative words and also for additional types of words: negation, amplification and de-amplification words. Negation words invert the polarity of words, e.g. "*not* good"; amplification words multiply the polarity strength of words by a factor greater than 1, e.g. "*very* good"; and de-amplification words multiply the polarity strength of words by a factor below 1, e.g. "*rarely* good". This structure is implemented by the **R** package **qdap** [199], which is used to calculate the sentiment scores of news items[10].

It has been established that sentiment analysis is influenced by the domain of the items to which it is applied [31]. This context-sensitivity characteristic is important as it carries effects on the vocabulary-level where words may express different polarities, depending on the domain. As such, a collocation assumption [86] is usually enforced, where it is assumed that given a sentence which expresses sentiment polarity, its expression is directed to the domain in question.

Considering this, question **Q6** is addressed, in order to study the relation between sentiment analysis and the popularity of news. It has been previously indicated that sentiment analysis provides a positive impact in prediction models, when used as features [18]. Therefore, using several sentiment lexicons, a study of the relation between sentiment score and news popularity is provided, according to each of the social media sources, for all available topics.

For this analysis four different sentiment lexicons are considered: AFINN [177], the sentiment lexicon used in the early work of Hu and Liu [107] henceforth referred as SentLex, the SentiStrength lexicon [218] and SentiWordNet 3.0 [21]. Each of these lexicons have been developed for and used in different contexts but all have in common their proven added-value to the detection and analysis of sentiment in short-text environments. Table 3.4 summarizes the aforementioned sentiment lexicons[11].

[10]The formulas used to calculate the sentiment scores are also described in [199].

[11]The sentiment lexicons were standardized in order to provide fair comparisons. This process included the translation of positive/negative labels to numeric (1/-1), the removal of neutral and repeated sentiment

Table 3.4: Description of known sentiment lexicons according to number of positive and negative words, scale and reporting the use of n-grams.

Lexicon	#Pos	#Neg	Scale	n-Grams
AFINN	878	1598	$[-5, 5]$	✓
SentLex	2003	4779	$\{-1, 1\}$	
SentiStrength	593	1971	$[-5, 5]$	
SentiWordNet	9021	11127	$[-1, 1]$	✓

Apart from their differences in terms of the amount of positive and negative words inscribed in the lexicon, their differences also include the scale in which they present the polarity/polarity strength of a given sentiment word and if they include n-grams (*e.g.* a good deal) or solely unigrams (*e.g.* good). A significant difference concerns the process of polarity and strength labelling, which in most cases was done by using manual labelling. AFINN and SentLex were manually labelled by the respective authors; the SentiStrength lexicon was manually labelled by a seed of annotators and optimized by trying different values of strength with each term in classification tasks; and the SentiWordNet lexicon uses synonym distance for attributing polarity strength to words.

Considering the diversity in sentiment lexicons, as well as news popularity according to each of the social media sources in both the data sets presented in this chapter, the relation between these two components is analysed. Due to understandability and space constraints, the focus of this analysis is on the case of the topic "economy"[12] as it provides a comparable illustration to the outcome of other topics. Given that the headlines of news items provides much more information than the title, the sentiment analysis procedures were applied to such component of the news. Results are shown in Figure 3.11. Note that in the figure, a logarithmic scale is applied to the popularity of news in order to ease the effort of understanding the results.

Results concerning the setting depicted in this figure show that *1)* there is no significant change in the relation between the various sentiment lexicons used and the popularity of news according to the available social media sources. Additionally, results also show that *2)* the magnitude of sentiment in news' headlines is not necessarily related to higher popularity, and that *3)* most of the news headlines have a sentiment score near 0, and that the most rare cases (those with high popularity) are included in such cases. Also, it is observed that these conclusions extend to the analysis of other topics using either the headline and the title of news.

words and aggregating multiple word-sentiment records based on the work of Gatti and Guerini [92]

[12]Results concerning the remaining topics and the application of sentiment analysis to the title of news are accessible in `http://tinyurl.com/kx7xekz`.

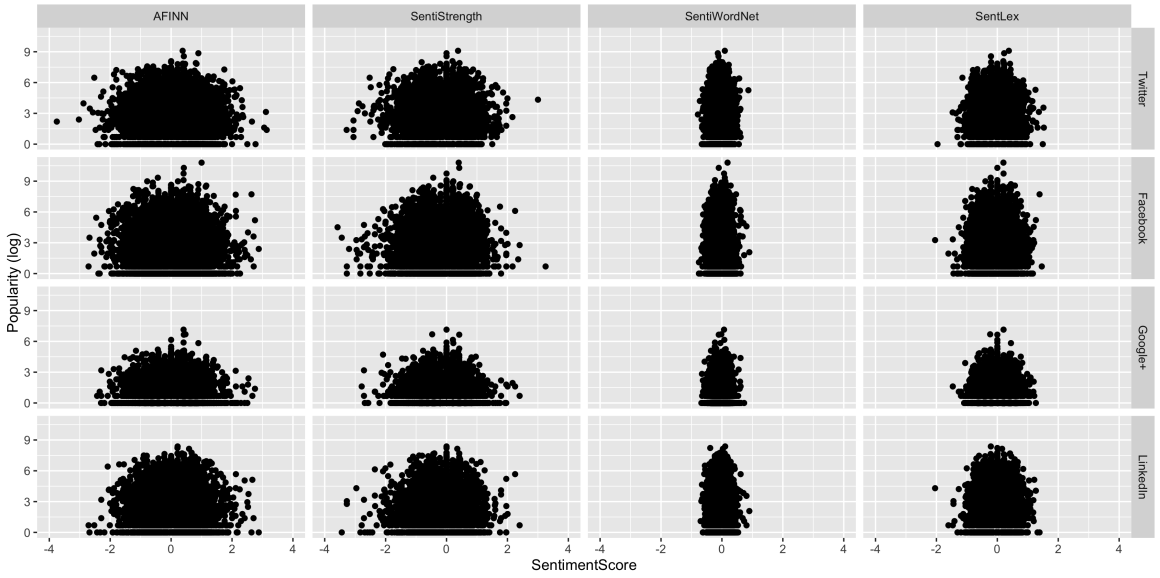


Figure 3.11: Scatter plot between sentiment score (headline) and the popularity of news items (logarithm) using four different sentiment lexicons and data from all social media sources concerning the topic "economy".

Named Entities

In addition to text mining tools such as sentiment analysis, news data also allows for the recognition of named entities. In the title and headline of news, a diverse type of entities may be mentioned, entailing further information.

One may postulate that by mentioning a certain entity in the title or the headline (e.g. Hillary Clinton, Republicans), this could have an impact on the attention that news gathers. This is the objective of this section, by addressing question **Q7** on the relation between the popularity of news items and mentioned named entities.

This analysis is carried out as follows. Using the infrastructure provided by the **openNLP** [105] package in **R**, a process of named entity recognition is applied to both the title and headline of all news items in each topic. This process is carried out separately for each of the data sets presented in this chapter. The recognition of named entities is focused on three types of entities: locations, people and organizations. For each type of entity, a different language model was used, leveraging available tools: the location, person and organization finder models made available by **openNLP**[13]. Building on this information, the average popularity of each entity is calculated by averaging the popularity of news in which they were mentioned. It should be stressed that this process is carried out independently for each topic and each data set. This is justified as to certifying that the analysis of entities are topic-dependent,

[13]These resources are available in `http://opennlp.sourceforge.net/models-1.5/`.

i.e. an entity may be highly popular in only one topic.

Using the outcome of this procedure, a score is attributed to each news item. In Figure 3.12 the scatter plot between news popularity and the average popularity of entities mentioned in each of these news is depicted. The depiction relates to the topic "economy" using data from both data sets. The values of each axis are transformed with a logarithmic scale to improve readability.

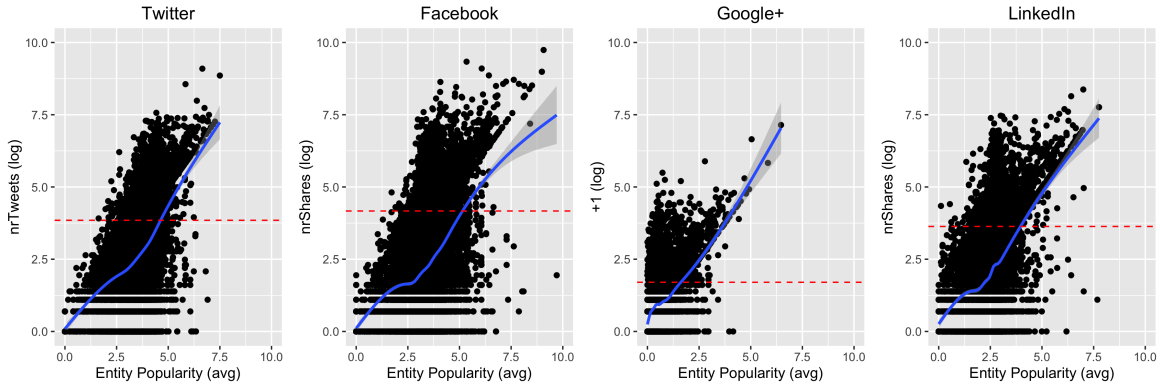


Figure 3.12: Scatter plot between the average popularity of mentioned named entities and news items' popularity in topic "economy", using data from available social media sources. The dashed line (red) represents the logarithm of the mean popularity in each scenario, and the smoothed conditional mean is also illustrated (blue).

By analysing the data depicted in Figure 3.12, the existence of a considerable degree of correlation between the popularity of mentioned entities and the popularity of the news in which they are mentioned is confirmed. Specifically, evaluating this correlation with the broadly used Pearson correlation coefficient (detailed in Equation 2.20), it is observed that in the topic "economy", the correlation between these two variables is considerable: 0.57, 0.65, 0.81 and 0.56 according to the social media sources Twitter, Facebook, Google+ and LinkedIn, respectively.

Nonetheless, it should be pointed out that results concerning all topics show that above-average news in terms of popularity, or even highly popular news, are not exclusively related to high average of mentioned entities' popularity. As such, it is concluded, concerning question **Q7**, that a significant degree of correlation between mentioned entities popularity and news popularity is observed, but when focusing on highly popular news, these are not exclusively related to highly popular entities.

News Outlet

Recommendations made by the official media sources sources include news from a large number of news outlets. These are not limited to traditional legacy media such as the

Washington Post or BBC News. News outlets in official media sources such as Google News and Yahoo! News range from blogs (e.g. The Nation) to radio stations (e.g. Radio New Zealand), traditional newspapers (e.g. Financial Times), and television networks (e.g. CNBC).

As previously mentioned, earlier work by Bandari et al. [24] provided important insights on the influence of news outlets and the popularity of items. Mainly, authors report results that point to the conclusion that this influence relates to news outlets that are widely recognized, and/or have a greater presence online, thus having a greater probability of obtaining high popularity levels for their news. Question **Q8** addresses this issue by studying the influence of news outlets in the final popularity of news. This is carried out by aggregating the popularity of news items by outlet and topic, guaranteeing that the scope (i.e. topic) of the news is considered, and that the impact of the news outlets' is accounted.

In Figure 3.13, results concerning the topic "economy" are shown, using data from social media sources of both data sets (single- and multi-source) and illustrating the relation between the average popularity of news outlets and the popularity of each item it published. For this figure, the popularity scores are transformed with a logarithmic scale, in order to provide a more understandable illustration of the results. Also, a dashed line (in red) is added in order to demarcate the logarithm of the mean popularity in each topic-social media source pair, and also a smooth approximation of news' popularity in relation to the average popularity of outlets' news (in blue).

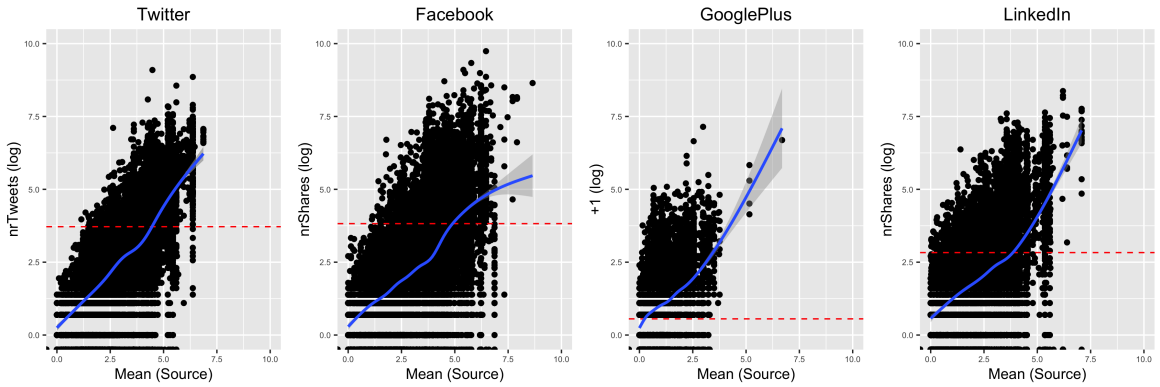


Figure 3.13: Scatter plot between the average popularity of news outlets and their respective news items' popularity, in topic "economy", using data from available social media sources. The dashed line (red) represents the logarithm of the mean popularity in each scenario, and the smoothed conditional mean is also illustrated (blue).

As expected, results show that there is a clear relation between the popularity of the news outlet and the popularity of their items. Although the figure only depicts the results concerning the topic "economy", it is observed that this tendency is also clear for the remaining topics, in both data sets.

However, news items included in the data sets presented only report news suggested by Google News and Yahoo! News, and therefore do not include all news from each outlet. Given this, it should be noted that results show that the most popular news outlets commonly show a small amount of news. Furthermore, results also show that, despite the greater probability of highly popular news being related to highly popular news outlets, these are not exclusive to such outlets. In fact, it is observed that in many cases the most popular news items published in a given topic (according to the popularity values by a given social media source), did not originate from the most popular news outlets.

Given these results, it is concluded that there is a clear relation between the popularity of news items and their respective news outlet. Also, it is observed that the average popularity of news outlets' items does provide an interesting indicator of news popularity. However, it should also be noted that the variation of news' popularity within news outlets is of great magnitude, and that the most popular news are not necessarily published by top news outlets, in terms of the average popularity of its items.

Temporal Patterns

The circadian[14] nature of user behaviour on the Internet is an important factor in the analysis provided in this chapter. In previous works, authors have considered this issue. For example, Kobayashi and Lambiotte [124] proposed an approach that modelled the data based on this circadian behaviour, and Tatar et al. [215] analysed the hourly performance of their proposed approaches in ranking news items. Also, Kong et al. [126] have indicated that temporal features are important for popularity prediction tasks.

In order to better understand the temporal patterns of user activity a study on the commonalities and disparities within this circadian pattern is provided, considering news of multiple social media sources. However, first and foremost, it is important to understand the temporal patterns involved in news publishing. As such, an illustration of the hourly average number of published news for each topic is provided, according to the official media sources available in the multi-source data set (Figure 3.14).

Results show that the dynamics of online news publishing is similar in both the single- and multi-source data set. It reveals that most topics see an increase in the amount of news published throughout the day. This conclusion does not hold for the topic "palestine", which is likely related to its lower throughput of news. However, in all topics a peak of news publishing at the turn of the day is observed. This can be explained by the scheduling of news for automatic publishing.

Based on these conclusions, the behaviour of users w.r.t. news items is analysed, within the

[14] Natural recurrence on a twenty-four-hour cycle.

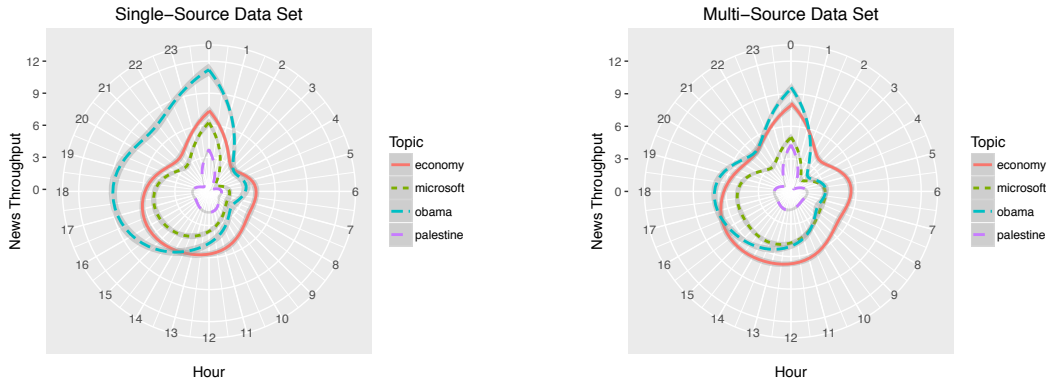


Figure 3.14: Daily average of news published for each topic, in both data sets.

scope of a circadian pattern. In Figure 3.15 the distribution of popularity according to the social media sources available in both the single- and multi-source data sets is illustrated, for each daily hour. As in previous cases, the popularity scores of each social media source are transformed with the application of a logarithm in order to improve understandability. Also, this illustration solely refers to the topic "economy".

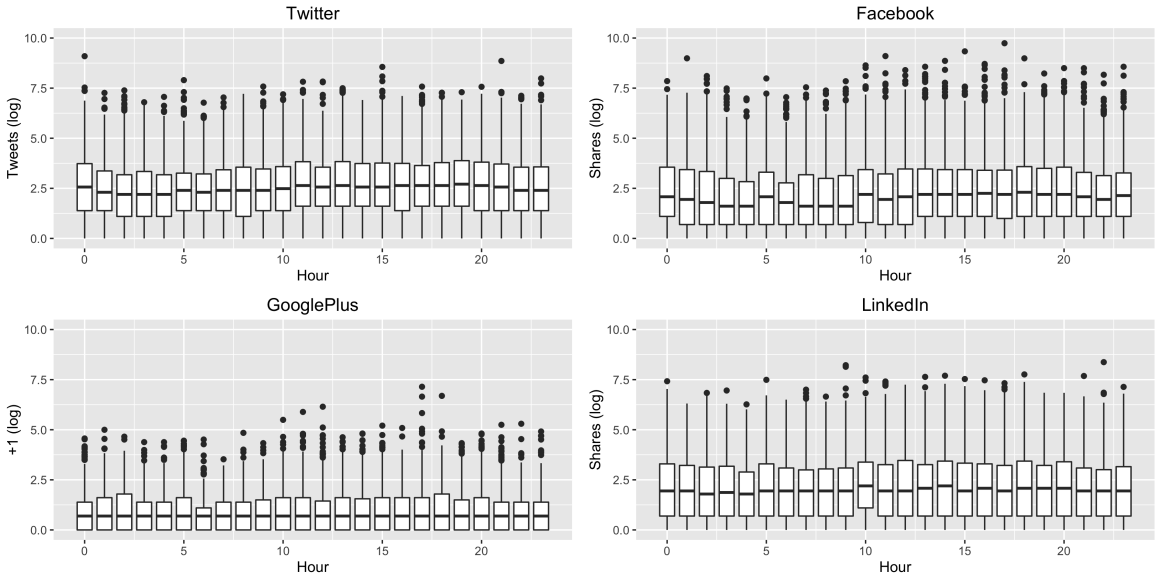


Figure 3.15: Distribution of news popularity (logarithm) per publishing hour, for all social media sources available, on the topic "economy".

Results show that the popularity of news is not related to the news publication temporal dynamics, previously depicted in Figure 3.14. Regarding question **Q9**, concerning the influence of publication hour in news popularity, results also demonstrate that in most cases the average popularity of news is rather stable throughout the day. Concerning highly popular news, results do not show any evidence of such cases being typically framed within a certain period of the day. Finally, concerning the other available topics, it is observed that

results provide evidence to support the aforementioned conclusions.

Notwithstanding, it is also important to frame the issue of popularity patterns within the space of the week (question **Q10**). As such, the process implemented to answer question **Q9** is repeated, but concerning the dynamics of news popularity in each weekday. First, in Figure 3.16 the throughput of news from official media sources in each of the weekdays is depicted, for each of the topics, in both the single- and multi-source data sets.

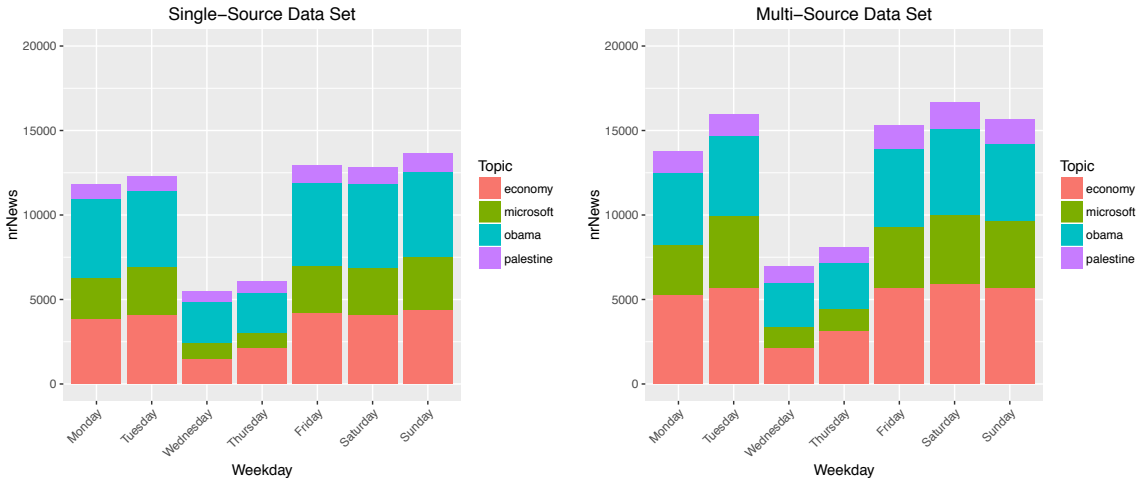


Figure 3.16: Throughput of news per weekday concerning each topic, in both data sets.

Results from both data sets agree, showing that for most days, a similar amount of news is published. However, it is observed a considerable decrease in daily throughput on Tuesdays and Wednesdays. This could be related to the scheduling process for news publishing: in order to maximize the number of online readers and newspapers' sales, the publishing of news deemed as impactful is planned for days of the week when the audience is larger, such as the beginning and the end of the week[15].

Based on these observations, the correlation of news popularity and their publication weekday is studied. Figure 3.17 shows the distribution of news items' popularity according to the social media sources available in both the single- and multi-source data sets, for the topic "economy". As in previous cases, popularity scores were transformed to a logarithmic scale in order to improve understandability.

Results show similar conclusions to question **Q9**: there is no apparent relation between the popularity of news items and the weekday they were published. As well as in the previous case, the average popularity of news items w.r.t. each of the social media sources is fairly stable. Results do not show any evidence of a given day or group of days explaining a significant amount of highly popular cases.

[15]Given the lack of empirical evidence to support this hypothesis, it should be solely considered as intuition.

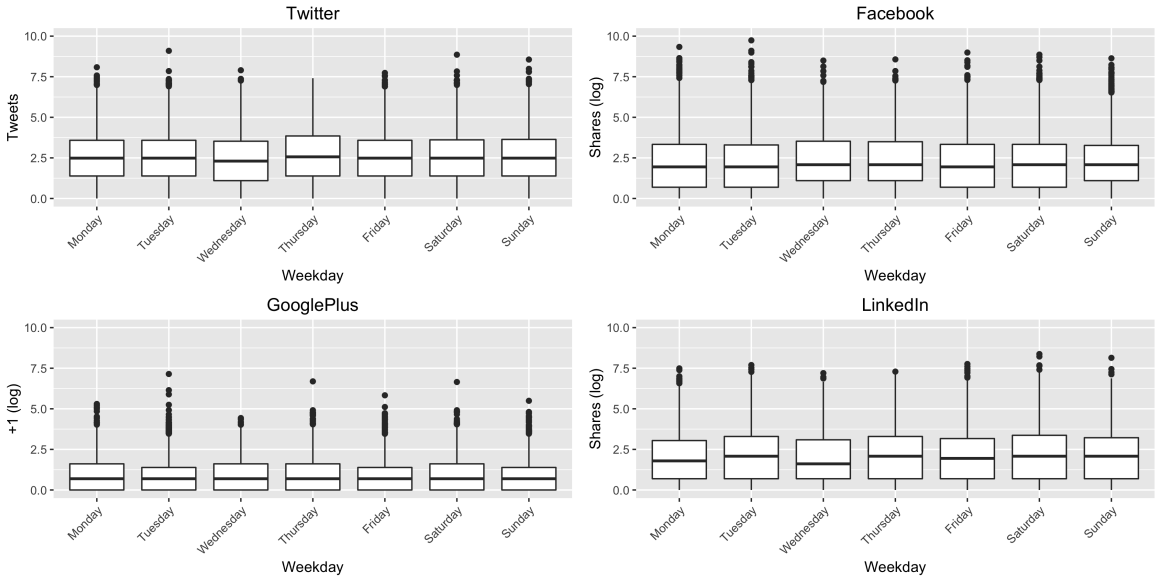


Figure 3.17: Distribution of news popularity (logarithm) per publishing weekday, for all social media sources available, on the topic "economy".

Summary

This analysis was focused on studying several aspects of the proposed data sets: the dynamics of popularity, the agreement between types of sources (official and social), the distribution of popularity, the analysis of sentiment in news, the impact of named entities and news outlets, as well as the contribution of temporal data. In summary, the following conclusions were drawn from the aforementioned analysis:

- The evolution of news items' popularity is quite rapid, achieving roughly 50% of each items' final popularity in only a few hours, in the case of Facebook, Google+ and LinkedIn. In the case of Twitter, results show that this social media source is more reactive, providing evidence that items obtain close to 50% of their final popularity under the first hour.

- Results obtained show that Google News and Yahoo! News (official media sources) share a small amount of news items in their recommendations. However, it also shows that this intersection of items is constant, and that they are usually considered by both sources as relevant items. As for agreement between social media sources, results show a considerable degree of agreement, revealing that this agreement is larger amongst Facebook and Google+. Finally, concerning the agreement between official and social media sources, results show that the user activity in Facebook, Google+ and LinkedIn is mostly discordant with the suggestions of Google News and Yahoo! News. Nonetheless, results show that Twitter and Google News present a considerable level of agreement.

- In previous work, a recurrent discussion is related to the characterisation of the distribution of popularity. On one side, several authors have provided empirical evidence of it being better described by a power-law distribution, whilst others have empirically defended that it is best described by a log-linear distribution. Resorting to diverse combinations of settings (different topics and social media sources), results using the data sets described in this chapter provide strong evidence indicating that in most cases the log-linear provides a better fit. Results also include the exponential distribution as a possibility, which provides the worst results in terms of goodness-of-fit.
- A study of the relation between the sentiment scores of both the title and headline of news items is applied. Results show that the magnitude of sentiment is not necessarily related to higher popularity and that most news have a sentiment score nearing 0. These conclusions were consistent in diverse settings where different sentiment lexicons were used, and throughout all possible combinations of topics and popularity data from all social media sources.
- The analysis of entities mentioned in news and their popularity, in relation to the popularity of their respective news, shows that a strong correlation exists between the two. This conclusion is transverse to all social media sources. However, it should be mentioned that results also show that the rare cases of highly popular news are not exclusively related to highly popular named entities. This same analysis procedure was applied to the popularity of news outlets and the popularity of their respective news. Results indicate similar conclusions, although denoting that highly popular news outlets present a small amount of news in the official media sources recommendations, and that the deviation of popularity in popular outlets is significant.
- Finally, results show that both the hour and weekday of publishing do not provide any evidence for further understanding the dynamics of highly popular news items. Also, it should be mentioned that results show a rather stable average of news popularity throughout publishing hours and weekdays. Finally, it is worth noting that, concerning the analysis of news items' popularity and the weekday of publishing, results from both data sets show that Wednesday and Thursday present a significant difference in number of news published.

3.4 Conclusions

In this chapter a particular case of web content is presented: online news feeds. This type of web content is described by a set of distinguishing characteristics in comparison to other types of web content. Most importantly, news items have a short lifetime, which exacerbates the complexity of popularity prediction tasks, requiring accurate and very early predictions of the level of attention that these items will receive.

A general description of the pressing issues that news outlets face is presented. These include the problems raised by the transition from offline to online environments, creating the need to publish large amounts of news from diverse topics, under the pressure of increasing user-demand. For news rooms, the online environment presents several problems which have been increasingly addressed. Given the evolving dependency on web advertisement revenues, not only does it require a permanent promptness in reacting to "newsworthy" events, but also in deciding on issues such as which news to promote or advertise. These issues are detailed in this chapter, using them as motivation for the use of online news feeds data in web content popularity prediction tasks.

Two data sets are detailed, using data from two different types of sources: official media sources, i.e. news recommender systems such as Google News and Yahoo! News, and social media sources (Twitter, Facebook, Google+ and LinkedIn). Their main difference resides in the amount of sources used in each data set. The first data set is single-source, using one official media source and one social media source, Google News and Twitter, respectively. The second data set uses multiple sources, resorting to Google News and Yahoo! News as official media sources, and to Facebook, Google+ and LinkedIn as social media sources. These data sets present the opportunity to study these two distinct scenarios in popularity prediction tasks. Upon the description of both data sets, an extensive exploratory data analysis was performed focusing on some of the key characteristics of online news feeds.

Using the data sets presented in this chapter, and building on the results from the extensive analysis provided, the task of web content popularity prediction is discussed and formalized in the following chapter. Unlike previous work, the work presented in this thesis is focused on the prediction and ranking of the rare cases of highly popular web content. The issues raised by standard prediction tasks when focused on such type of items are discussed, motivating the use of utility-based regression [196]. Also, building on this concept, an evaluation framework for web content popularity prediction tasks is presented. Using this evaluation framework and the data sets hereby presented, the contribution of a diverse set of predictors in popularity prediction tasks is analysed, when focusing on highly popular items.

Chapter 4

Learning with Imbalanced Domains

In this chapter the problem of learning with imbalanced domains is introduced, showing how skewed distributions impact standard evaluation approaches in the context of predictive modelling tasks. The concept of utility-based regression, a non-standard approach for such tasks, is described. Considering the primary goals of this thesis - the prediction and ranking of highly popular web content - a new approach to utility surfaces is proposed, based on rule-knowledge derived from user preferences, and a robust evaluation framework is presented. Finally, an experimental evaluation is carried out with two objectives: i) to assess the contribution of several types of predictors in terms of accuracy towards highly popular items, in a priori popularity prediction tasks; and, ii) to provide insights concerning the caveats raised by standard evaluation methods used in previous work.

4.1 Introduction

Building on the analysis of the web content domain provided in previous chapters of this thesis, it is noticeable how large amounts of data may impede decisions that until recently were the responsibility of human actors. For example, news outlets today face the challenge of providing an overwhelming degree of throughput in news stories, on an hourly basis. Decisions regarding the promotion or the advertisement of such items were feasible to decision-makers in newsrooms on an offline environment. However, with the transition to online news, this task is now virtually unfeasible for human actors.

To tackle this issue, one of the main contributions of machine learning and data mining relate to learning and data analysis tasks such as predictive modelling, allowing the prediction of outcomes. Predictive modelling is a statistics-based task where the goal is to map a given set of features to a given target (or targets). Formally, given a variable Y from a domain $\mathcal{Y}$, predictive modelling attempts to approximate a function $Y = f(X_1, X_2, \cdots, X_p)$, where Y is the target variable, $X_1, X_2, \cdots, X_p$ are describing features, i.e. predictors, and $f()$ is

the unknown function one attempts to approximate. In order to obtain an approximation $h(X_1, X_2, \cdots, X_p)$ of this unknown function, a set with examples of the function mapping (known as a training set) is used, *i.e.* $D = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n}$. Depending on the type of Y variable, this task can be denoted as a classification task, when Y is nominal, or a regression task when Y is continuous.

Successful predictive modelling approaches allow the automation of many previously human-dependent tasks, and the ability to handle a greater number of such tasks. These can be solved using many existing algorithms, where a trial-and-error process is applied in order to find the best solutions. This performance-based process is commonly driven by the optimization of a given standard error criterion.

However, previous work [155, 128, 35] suggests that the use of standard learning algorithms and evaluation metrics is prone to misleading conclusions and under-performing models in domains with skewed distributions of the target variable, such as the case of web content popularity. This characteristic of data sets is commonly denoted in related literature as imbalanced data.

4.2 Imbalanced Domains

Learning tasks are commonly faced with imbalanced data. This problem has been extensively studied within the scope of classification tasks [155]. In such tasks, this issue is presented when a significant gap amongst the prior probabilities of different classes is observed [101], i.e. the probabilities of a given example belonging to distinct classes are significantly different. For example, in a binary classification task, this occurs when one class is severely under-represented in relation to the other. This problem is also common in regression tasks. Given the continuous nature of the target variables in such tasks, the problem is observed when specific intervals of the domain are severely under-represented. However, despite the interest in solving such tasks [128], the problem of imbalanced data has not yet been thoroughly explored in regression tasks.

In order to provide a coarse definition of imbalanced data, capable of covering both nominal and continuous prediction tasks, the following is presented:

Definition 4.2.1. Imbalanced Data. A characteristic of data from domains where the distribution of target values is severely skewed. Depending on the type of variable, this characteristic translates to the existence of a given class(es) or numeric interval(s) which is (are) extremely under-represented.

Despite the widespread observation of imbalanced data, there are scenarios in which it may not be a relevant issue. Consider the problem of handling outliers in traditional statistics

literature [190]. An outlier is considered to be a data point showing a significant deviation from the remaining data [3], therefore similar to the under-represented items of skewed distributions. Several reasons for the presence of outliers may be presented, ranging from data input and sampling errors, to intentional misreporting. Previous work [190] argues that the treatment of outliers depends on such reasons and the objective of the task. In addition, some authors argue that when outliers are considered as illegitimate or error-based, these may be removed in order to obtain the best parametrization possible [117] (e.g. [119]).

However, in many domains, these rare and under-represented items are considered by users to be the most important cases in terms of predictive accuracy implying a great cost when incorrectly predicted [72]. The combination of such factors (skewness of data distribution and domain preferences towards less common cases) is the basis for the task of imbalanced domain learning.

Regarding classification tasks, Krawczyk et al. [128] observe that this is a common situation in many domains. Examples include activity recognition [88], behaviour analysis [20], cancer malignancy grading [129], sentiment analysis [244], text [172] and video mining [89], amongst others. As for regression tasks, examples of domains with imbalanced data sets include those related to meteorology [82], electricity [127] and water consumption [261] or financial markets [8].

The application of standard algorithms in learning tasks with imbalanced domains has been thoroughly studied in previous works [155, 128, 35], showing how these may favour the over-represented (majority) items, which in many cases, users consider to be non-relevant [180, 90]. Standard learning algorithms are commonly focused on optimizing a given standard evaluation metric. Since these metrics are focused on the average error, the results will be greatly influenced by the behaviour of the models towards the majority of items, to the detriment of the minority items.

Consider the example of online news data, detailed in the previous chapter. The goal of the related predictive task is to forecast the amount of popularity that news will obtain in social media, i.e. a numerical prediction task. Evidence shows that the distribution of web content popularity is best described by a heavy-tail distribution. As such, at a given moment, it is possible to have a set of news where the majority of items has a small amount of popularity, whilst a small group of rare cases has a very high level of popularity. Coincidentally, these highly popular news are those that should be accurately predicted, in order to quickly place them in the top of recommendations to users.

To illustrate the issues associated to the evaluation of learning tasks in the imbalanced domain of web content using standard evaluation metrics, a ranking scenario is depicted in Figure 4.1. This scenario uses data from the single-source data set (described in Sec-

tion 3.2.1), in which the popularity of 100 news[1], according to the social media source Twitter is presented. A naïve set of artificial predictions is used, by generating random values from a normal distribution, using the median of ground-truth popularity values and a standard deviation of 3. In this example, the evaluation metric Normalized Mean Squared Error ($NMSE$) described in Equation 2.16 (Section 2.4), is used. This metric is bounded by $[0, 1]$, where 0 denotes the optimal result.

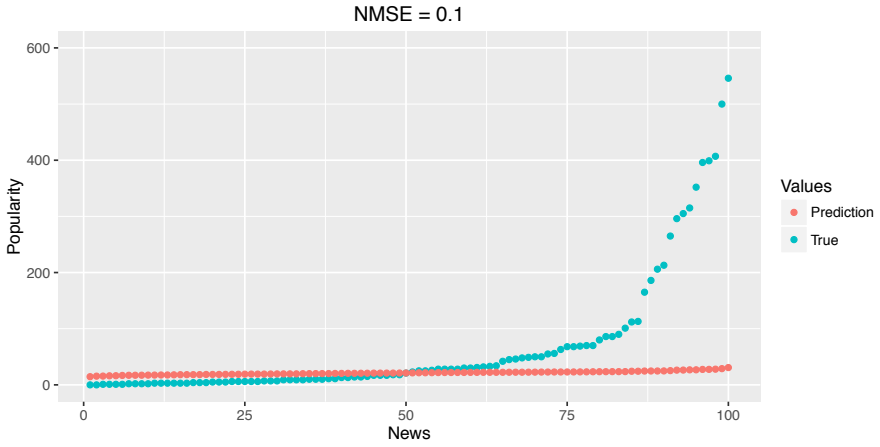


Figure 4.1: Google News ranking with 100 news, ordered by their respective popularity according to the social media source Twitter, and a set of 100 random predictions generated by a normal distribution.

Due to the data imbalance of web content popularity, and in this case the online news feeds data, it is observed that for most of the items the error between the true and predicted values, is relatively small. This is confirmed by the resulting $NMSE$ score: 0.1. As such, by according to this standard evaluation metric, one could conclude that this median-based naïve approximation is fairly acceptable. However, by observing the illustration, predictions show significant errors concerning the small set of highly popular items.

This problem relates to one of the main subjects of this thesis. Focusing on regression tasks, i.e. numerical prediction, our objective is to enable a fast and precise anticipation of web content items' popularity, targeting the small set of highly popular items, i.e. those that should be placed in the top of recommendations to users.

Based on the description of imbalanced domains and the formal definition of predictive modelling tasks, the problem of learning with imbalanced domains abides to the verification of two conditions [35]:

1. a subset of the target variable domain is attributed more importance by the user, w.r.t. its predictive performance in the obtained model;

[1]This set of news is given by a Google News ranking on the topic "economy", with the time stamp 2015-05-29 05:40.

2. the most relevant cases for the user in the training set are severely under-represented, causing a poor predictive performance in such cases.

Unlike standard learning tasks, these conditions describe a scenario where users assign different levels of importance to distinct types of cases. This uneven judgment of importance may be denoted in different manners, relating the attribution of different benefits to the accurate prediction of target values, or the association of different costs to different types of prediction errors. Additionally, this non-uniform importance may also be characterized by a combination of different costs and benefits.

4.2.1 Utility-based Learning

Several learning approaches have been proposed to address problems similar to that of learning with imbalanced domains. These are mainly related to utility-based data mining [237, 238, 253] or utility-based learning [85]. The concept of *utility* is originally found in the field of economics, used as a metric of usefulness, i.e. benefits, concerning the consumption of a given good. More recently, authors have observed how this concept can also be applied to machine learning and data mining tasks. In this context, Elkan [72] defines utility as a function combining positive benefits and negative benefits (costs), being applied as a domain-specific metric of approaches' usefulness, with the main goal of utility maximization.

As an illustration of the practicality of the utility concept, consider the example provided by Ribeiro [196] concerning fraud detection. In such scenario, decisions are related to the triggering (or not) of actions. If one suspects that a given transaction is fraudulent, an action can be triggered. However, if an action is triggered and the transaction is not fraudulent, a cost is associated. This is also the case if a fraudulent transaction is not detected and no action is triggered.

Amongst the extensive work concerning utility-based data mining, the most popular approach is cost-sensitive learning [72]. Usually associated to classification problems, previous work [252, 150] shows that there are three main approaches to this type of learning task: *i)* minimization of expected cost, *ii)* example weighting, and *iii)* cost-sensitive classifiers. The first associates non-uniform costs to classification errors, expecting the learner to minimize such costs. The second tackles the problem by sampling the data set before applying a classifier algorithm. Finally, the third is based on the conversion of classification algorithms to being cost-sensitive.

One of the main differences between these cost-sensitive learning approaches is the manner in which misclassification costs are represented. The first approach assumes the existence of prior knowledge or the ability to estimate misclassification costs associated to the target

domain. The second approach does not make this assumption. Instead, it weighs examples of the data set in a cost-proportionate manner, similarly to the boosting technique [83]. The third approach may assume costs by any of both approaches.

Given the characteristics of the target domain of this thesis, web content popularity, the available information allows for the first assumption. As such, we will focus on the first approach, cost-sensitive learning by minimization of expected cost, henceforth referred to as cost-sensitive learning for simplification purposes.

Early work regarding cost-sensitive learning includes several proposals (e.g. [62, 72]) as to formulations of cost matrices which incorporate domain knowledge. These proposals associate costs to all possible classification combinations between true and predicted values. Domingos [62] proposes the cost-sensitive meta learning algorithm MetaCost, based on Bayes risk theory [67], and Elkan [72] proposes the formulation of benefit matrices, in which costs are measured in relation to a given benefit baseline. This proposal is an extension of cost matrices, allowing for positive and negative values. Also, it stipulates that accurate predictions have non-negative values in the matrix, and that the remaining values in the matrix should have negative values.

Based on the definitions proposed by Ribeiro [196], cost and benefit matrices may be defined as follow:

Definition 4.2.2. Cost Matrix. Given a $n \times n$ matrix $C := [c_{ij}]$, where n is the number of classes existing in domain $\mathcal{Y}$, the value c_{ij} denotes the cost of classifying a case of true class j as class i. The structure of the cost matrix is then defined such that,

$$c_{ij} = \begin{cases} 0, & \text{if } i = j; \\ > 0, & \text{otherwise} \end{cases} \tag{4.1}$$

Definition 4.2.3. Benefit Matrix. Given a $n \times n$ matrix $B := [b_{ij}]$, where n is the number of classes existing in domain $\mathcal{Y}$, the value b_{ij} denotes the benefit of classifying a case of true class j as class i. The structure of the cost matrix is then defined such that,

$$b_{ij} = \begin{cases} \geq 0, & \text{if } i = j; \\ < 0, & \text{otherwise} \end{cases} \tag{4.2}$$

These matrices provide the foundation that allows the application of cost-sensitive learning in solving tasks such as learning with imbalanced domains. However, it should be noted that the relation between the two problems (cost-sensitive and imbalanced domain learning) is dependent on the verification of the two previously mentioned conditions: based on users preferences: *i)* the judgment of cases importance is uneven, and *ii)* the most relevant cases are severely under-represented.

Users' domain preferences are a crucial factor in cost-sensitive learning. Due to their importance, several caveats may be raised as to the quality and detail of such specifications. For example, previous work notes that the quality of information provided by users has an impact on the evaluation and comparison of models, as well as the optimization of such models [35]. This is known as the "problem-definition issue" [236].

Ideally, for each cost-sensitive learning task, one should possess a detailed specification of the utility function provided by the user, U, denoting the utility value (i.e. usefulness) of making a given prediction $\hat{y}$, given a true value y. However, in most cases, this is not possible. For example, real-world domains are commonly affected by systematic changes in the distribution of observed values, i.e. non-stationarity. As such, this would require the user to provide regular updates concerning the utility values associated to all possible pairs of predicted and true values, $\langle \hat{y}, y \rangle$. Also, given the amount of combinations, providing such complete specifications of utility functions may be extremely difficult.

Furthermore and most importantly, as previously stated, the majority of work concerning cost-sensitive learning is focused on classification tasks. Considering the issues raised by the "problem-definition issue", the complexity of the problem is much larger when regarding regression tasks, given their continuous domains. In such tasks, the user is required to provide, in ideal scenarios, the utility function $u(\hat{y}, y)$ for an infinite domain, which is unobtainable. Even in the case of providing a constrained utility function, the amount of effort required to perform such task would be significant.

4.3 Utility-Based Regression

According to Crone et al. [55] the majority of research concerning regression tasks does not consider uneven judgments of values' importance, i.e. assumes uniform costs. For example, according to the extensive review of previous work in web content popularity prediction (Chapter 2), this problem has not yet been addressed when considering numeric prediction tasks.

In standard regression tasks, the objective is to optimize models according to a given loss function, $L(\hat{y}, y)$. This loss function may be denoted by different error criteria, such as the absolute or the squared error of predictions w.r.t. the true values. The underlying assumption is that the usefulness of a prediction is inversely proportional to the respective loss values. As such, given uniform domain preferences, standard regression tasks assume that utility U is a function of the prediction error $L(\hat{y}, y)$, and that the utility function is inversely proportional to the loss function, $U \propto L^{-1}$. Therefore, given a domain $\mathcal{Y}$, the following properties [196] are verified for any pair of predictions $\langle \hat{y_1}, y_1 \rangle, \langle \hat{y_2}, y_2 \rangle \in \mathcal{Y}$:

1. Equally accurate predictions have the same utility.
   $L(\hat{y_1}, y_1) = L(\hat{y_2}, y_2) \Rightarrow U(\hat{y_1}, y_1) = U(\hat{y_2}, y_2)$;

2. More accurate predictions have greater utility
   $L(\hat{y_1}, y_1) < L(\hat{y_2}, y_2) \Rightarrow U(\hat{y_1}, y_1) > U(\hat{y_2}, y_2)$;

3. Less accurate predictions have lesser utility
   $L(\hat{y_1}, y_1) > L(\hat{y_2}, y_2) \Rightarrow U(\hat{y_1}, y_1) < U(\hat{y_2}, y_2)$.

Although these properties may be valid for tasks where users' domain preferences are uniform, when regression tasks involve imbalanced domains, they could be misleading. Based on the work of Branco et al. [35] a set of properties is provided below, which are verified[2] in regression tasks with imbalanced domains and non-uniform domain preferences.

1. Equally accurate predictions may have different utility.
   $L(y_1, y_1) = L(y_2, y_2) \not\Rightarrow U(y_1, y_1) = U(y_2, y_2)$;

2. More accurate predictions may have lesser utility.
   $L(\hat{y_1}, y_1) < L(\hat{y_2}, y_2) \not\Rightarrow U(\hat{y_1}, y_1) > U(\hat{y_2}, y_2)$;

3. Less accurate predictions may have greater utility.
   $L(\hat{y_1}, y_1) > L(\hat{y_2}, y_2) \not\Rightarrow U(\hat{y_1}, y_1) < U(\hat{y_2}, y_2)$.

These properties also represent a formalization of the first condition (Section 4.2), which must be verified in order to consider a problem as learning with imbalanced domains: a subset of the target variable domain is more relevant to the user, in terms of its predictive performance.

Given a regression task on an imbalanced domain, the main problem of using cost-sensitive learning is related to the "problem-definition issue". In regression tasks, the information concerning domain preferences is potentially infinite, therefore requiring alternative approaches as to the definition of utility functions.

This was addressed in the work by Ribeiro [196] on utility-based regression. Concretely, the author focuses on how to enable tasks of evaluation, comparison and model selection while accounting for uneven judgments of items importance. To our knowledge, no other approach on how to handle imbalanced domains learning with regression has been proposed. Nonetheless, it should be noted that several authors have addressed the problems related with the average-based evaluation in regression tasks [49, 55] in the domain of financial applications. However, these approaches are specifically focused on the distinction between

[2]These properties may be verified singularly, or jointly.

under- and over-predictions, and therefore do not address the problem of users' non-uniform domain preferences.

Coarsely, the approach of utility-based regression is based on two concepts: *i)* relevance functions, and *ii)* utility surfaces.

4.3.1 Relevance Functions

In order for the user to define the importance of values in a given domain, Ribeiro [196] proposes the use of relevance functions. By definition, the relevance function allows the user to assign a relevance score to each of the values in a certain target variable, concerning a given domain. Therefore, in formal terms, a relevance function $\phi()$ maps the values of a target variable Y, from a given domain $\mathcal{Y}$, to a range of importance/relevance. This function is bounded by $[0, 1]$, where 0 corresponds to minimum relevance, and 1 to the maximum relevance, and is described as

$$\phi(Y) : \mathcal{Y} \rightarrow [0, 1] \tag{4.3}$$

The advantages of this approach are described by Branco et al. [35]. The authors note that it relaxes the problem associated to the definition of utility functions in regression as follows: *i)* instead of requiring the definition of a function which depends on two variables ($\hat{y}$ and y), the relevance function solely requires one (y); and *ii)* relevance functions require a significantly smaller amount of information in comparison to utility functions.

In order to facilitate the users' task of providing relevance functions, Ribeiro [196] also proposes an approach for the interpolation of relevance, given a set of user-defined pairs of relevance scores and target values, i.e. control points. These control points report to values in the target variable of the imbalanced domain for which the user is aware of the relevance score. Given this information, the author proposes the use of Piecewise Cubic Hermite Interpolating Polynomials [64] (*pchip*) in order to define an appropriate relevance function.

In addition, Ribeiro also proposes a distribution-based approach to the automatic definition of relevance functions. This is appropriate when users have no domain knowledge, or when domains are highly dynamic, thus requiring regular updates to previous relevance judgments. This proposal is based on box plot statistics [227], which provide a visualisation of key elements of a continuous variable distribution. This method is non-parametric (i.e. does not assume any underlying distribution), and provides an emphasis on the tails of the distribution.

The elements illustrated by the box plot visualisation w.r.t. the distribution of a given

target variable Y are: *i)* the median ($\bar{Y}$), *ii)* the lower and upper hinge, describing the first ($Q1$) and third ($Q3$) quantile, and *iii)* the lower and upper whisker. The middle 50% of the distribution is depicted within a box, delimited by $Q1$ and $Q3$. The range between $Q1$ and $Q3$ is known as the inter-quartile range (IQR). The lower and upper whiskers denote the distribution space outside of the box, where values are not considered as outliers, i.e. the box plot rule [51]. According to Tukey [227], these are defined as $Q_1 - \alpha \times IQR$ and $Q_3 + \alpha \times IQR$, respectively. The α factor represents a coefficient that is variable depending on the preferences of the user, although traditionally defined as 1.5.

It should be noted that, concerning our target domain of web content popularity, cases which are considered as outliers, are also extreme values. Extreme value analysis [189] is based on 1-dimensional data, assuming that outliers are values which present a value too large or too small. Conversely, the traditional definition of outliers (e.g. Hawkins [100]) is related to the *generative probability* of the values and not their extremity. Thus, it should be clarified that in the scope of this thesis, outliers and extreme values report the same type of cases: rare cases of highly popular items. Therefore, such type of values will be referred to using both terms interchangeably.

Based on the information given by the box plot statistics, one is able to define a set of control points which are then used to automatically define a relevance function, resorting to the *pchip* interpolation method. These control points are as follows, given a target variable Y:

- a relevance score of 0 is attributed to the median of Y;
- a relevance score of 1 is attributed to the values of the lower and upper whiskers, i.e. boundaries for the definition of values as outliers;
- a relevance score of 1 is attributed to all outliers.

To exemplify the application of this method, the data formerly employed in Figure 4.1 (Section 4.2) is used. Figure 4.2 shows the interpolation of relevance ($\phi(Y)$) based on box plot statistics (top) of the popularity values Y.

Results show that this approach is capable of depicting the distribution of a target variable Y, in accordance with box plot statistics. To elaborate, it is observed that the relevance of values between 0 and the median of the distribution (22) have a relevance of 0; the relevance value of the third quantile ($Q3 = 68$) is 0.51; and all values equal or greater than the upper whisker (113) have a relevance of 1. In addition, it should be stressed that in the case of web content popularity, its distribution has only one tail. Therefore, one can only expect high extreme values (i.e. upper whisker). This is not the case for many other domains (e.g. electricity [127]) presenting extreme values on both sides of the box.

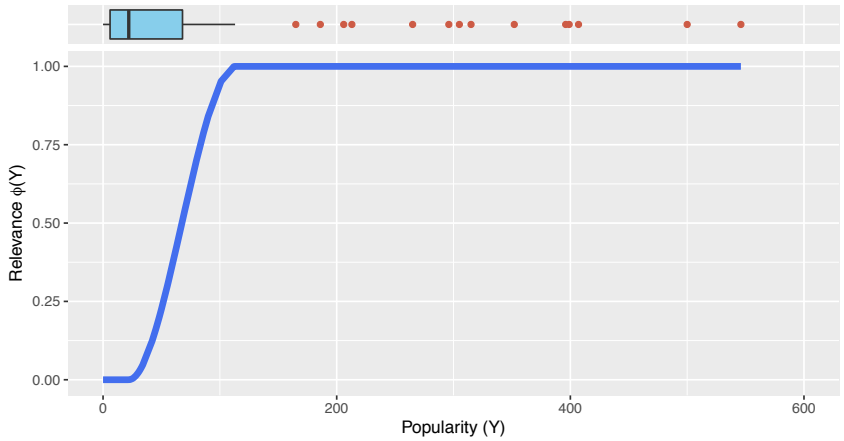


Figure 4.2: Relevance function of data in Figure 4.1 with boxplot statistics (top).

Relevance Threshold

In many imbalanced domains, for both classification and regression tasks, the users may introduce thresholds, thus defining importance boundaries. For example, in a financial market application, the user may define a minimum threshold for the transaction to be considered interesting, i.e. having a relevant level of return. As such, it may be assumed that in addition to the relevance function provided by the user, or automatically generated as previously described, the user may also define a relevance threshold, t_R.

The relevance threshold represents a boundary for the user-definition of relevant values. It should be stressed that this boundary does not serve the purposes of discretization (definition of classes) or the definition of irrelevant values. Its objective is to define the values that, according to the user, are the most relevant in a given domain.

Using this threshold, it is possible to consider the domain $\mathcal{Y}$ of a given target variable Y as having two types of values, according to their relevance: a subset of the domain where values are considered highly relevant, and a second domain subset with values considered to be normal. The first subset $\mathcal{Y}_R \subset \mathcal{Y}$ contains all domain values that have a relevance greater than the defined threshold, $\mathcal{Y}_R = \{y \in \mathcal{Y} : \phi(y) > t_R\}$; and the second subset $\mathcal{Y}_N \subset \mathcal{Y}$ contains all remaining values $\mathcal{Y}_N = \mathcal{Y} \setminus \mathcal{Y}_R$.

Using this notation, a formalization to the second condition (Section 4.2) required in order to consider a problem as learning with imbalanced domains is provided. Given a subset D_R of training data D where $y \in \mathcal{Y}_R$, and D_N is a subset containing the values considered by the user to be normal $D_N = D \setminus D_R$, imbalanced domain learning tasks must verify the condition $|D_N| \gg |D_R|$: relevant cases are severely under-represented w.r.t. normal cases.

4.3.2 Utility Surfaces

Based on the concept of relevance functions, Ribeiro [196] defines the principle of utility for imbalanced domain regression tasks:

Definition 4.3.1. Principle of Utility in Non-Uniform Regression. Utility is a function of the error of predictions, $L(\hat{y}, y)$, and the relevance of both predicted ($\hat{y}$) and true (y) values.

This definition depicts the differences between standard and utility-based regression. The former assumes that users' domain preferences are uniform and therefore utility is an inverse function of the loss function $L(\hat{y}, y)$. However, in utility-based regression, non-uniform domain preferences are assumed, and as such, the utility of a given prediction $u(\hat{y}, y)$ not only depends on the loss function, but also on the relevance of true and predicted values.

Nonetheless, despite the relaxation of the "problem-definition issue" [236] using the concept of relevance, it is still necessary to translate such relevance functions to utility values. Accordingly, Ribeiro [196] proposes the following formalization of utility functions for utility-based regression tasks:

$$u_{\phi}^{p}(\hat{y}, y) = B_{\phi}(\hat{y}, y) - C_{\phi}^{p}(\hat{y}, y) \tag{4.4}$$

$$= \phi(\hat{y}) \cdot (1 - \Gamma_B(\hat{y}, y)) - \phi^{p}(\hat{y}, y) \cdot (1 - \Gamma_C(\hat{y}, y)), \tag{4.5}$$

where $B_{\phi}(\hat{y}, y)$ and $\Gamma_B(\hat{y}, y)$ are the benefit and bounded benefit loss functions, and $C_{\phi}^{p}(\hat{y}, y)$ and $\Gamma_C(\hat{y}, y)$ are the cost and bounded cost loss functions, respectively[3].

According to this proposal, both the benefit and cost loss functions, abide to the preposition of a maximum admissible loss. It stipulates that the maximum prediction error is equal to double the smallest amplitude between relative or absolute minimums and maximums. Predicted values presenting an error of equal or larger magnitude, obtain minimum utility, -1. As an example, given the ranking scenario illustrated in Figure 4.2, the maximum admissible loss is defined by the amplitude between the highest value y with relevance $\phi(y) = 0$ ($y = 21$), and the smallest value y with relevance $\phi(y) = 1$ ($y = 113$).

Using this process, one is able to obtain utility functions for continuous domains, also denoted as *utility surfaces*. These can be interpreted as a continuous version of benefit matrices used in cost-sensitive learning, with classification tasks [72]. Figures 4.3 and 4.4 depict a utility surface, when applied to the data used to illustrate Figure 4.2.

Results show that by applying this approach, one is able to account for non-uniform users'

[3]Each of these functions is thoroughly presented in [196], and as such they will not be described in detail.

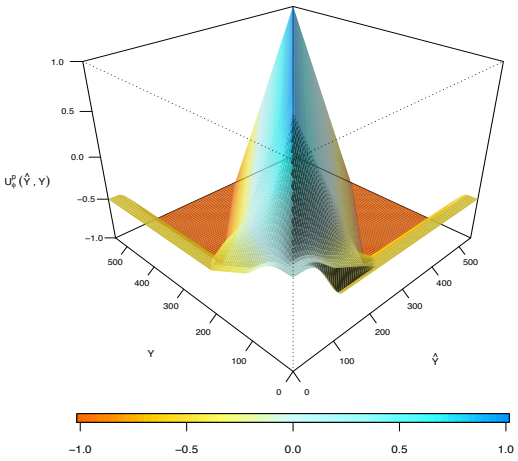


Figure 4.3: Example of utility surface (3D).

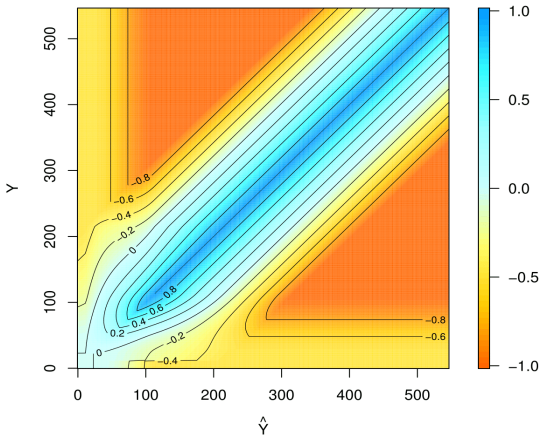


Figure 4.4: Example of utility surface (isometrics)

.

domain preferences, allowing a utility-based evaluation of models' predictions. This evaluation accounts for the relevance of both the predicted and true values, according to a given relevance function, and the prediction error (loss).

4.4 Rule-Based Utility Surfaces

The approach proposed by Ribeiro [196] to obtain utility surfaces is originally motivated by the problem of actionable forecasting tasks [28], which spawns from the concept of activity monitoring [78]. Actionable forecasting entails the process of predicting the correct action, inferred by a function of a predicted numerical variable.

As an example of such tasks, a previously described scenario (Section 4.3.1) is restated. Given a financial market application, a given user may define a minimum threshold (i.e. relevance threshold) for the transaction to be considered positive, i.e. renders a given level of profit. As such, if the outcome is much more lucrative than a borderline positive prediction, this does not translate to a case of negative utility. Conversely, it could be considered as a near-zero utility transaction, since the outcome for the user could have been much more lucrative. Objectively, in such scenarios, users will not consider correct predictions of positive or negative actions as non-useful, i.e. with negative utility. This is also the case in the domain of web content popularity: amongst a considerable size of items, correctly predicting rare items as highly relevant should not render a negative utility, and vice-versa.

The proposal of utility by Ribeiro [196] states two criteria for considering a prediction as

beneficial (*i.e.* positive utility): *i)* the predicted value leads to the correct action, and *ii)* the deviation of the predicted value is within a maximum admissible loss (prediction error). Given a domain where the ground assumptions are similar to that of the previous example scenario, it should be noted that this approach raises an important caveat. As observed in the previous illustrations of utility surfaces and isometrics (Figures 4.3 and 4.4), the original concept of utility by Ribeiro allows for highly relevant values which are predicted as such, to have a negative utility. This is due to the definition of the maximum admissible loss (double of the smallest amplitude between relative or absolute minimums and maximums), and the impact of this constraint on domains that have highly extreme values.

Based on these observations, a new approach for utility surfaces is proposed. As in the work of Ribeiro, this proposal also considers utility as a function of both the relevance of true and predicted values, and a given loss function. For simplification purposes, the utility function is denoted as follows [35]:

$$u(\hat{y}, y) = g(\phi(\hat{y}), \phi(y), L(\hat{y}, y)), \tag{4.6}$$

where y and $\hat{y}$ are true and predicted values and $\phi()$ is a relevance function.

As an alternative to the constraint of maximum admissible loss, a rule-based approach to derive utility surfaces is proposed, resorting to interpolation methods. The baseline rules of this proposed approach are as follows:

1. If a case is correctly predicted as highly relevant or non relevant, its utility is bounded by $[0, 1]$;
2. If a case is incorrectly predicted as highly relevant or non relevant, its utility is bounded by $[-1, 0]$;

By using rule-based knowledge and the relevance function $\phi()$ for a target variable Y of a given domain $\mathcal{Y}$, a set of predicted ($\hat{y}$) and true (y) paired values is obtained, where the corresponding score in the utility function $u(\hat{y}, y)$ is known. These are described as such:

- When the predicted value is equal to the true value ($\hat{y_1} = y_1$), the utility of the prediction is equal to the relevance of the true value: $u(\hat{y_1}, y_1) = \phi(y_1)$;
- Pairs of predicted and trues values where one corresponds to its maximum value, e.g. $max(\hat{y})$, and the other is a value with the same relevance as the relevance threshold, $\phi(y_2) = t_R$, the utility value is 0: $u(max(\hat{y}), y_2) = 0$;
- Cases where the prediction error is maximized (*e.g.* $(max(\hat{y}), min(y))$), the utility value is equal to -1: $u(max(\hat{y}), min(y)) = -1$.

In summary, *i)* accurate predictions obtain a utility corresponding to the true value's relevance, $\phi(y)$; *ii)* the minimum utility for cases correctly predicted as highly relevant or normal is 0; and *iii)* predictions corresponding to the largest prediction error possible have a utility of -1.

To interpolate the remaining values of utility, a scattered-data surface fitting approach proposed by Akima [9] is used. This approach provides a framework for the interpolation of irregularly and regularly spaced data[4]. This interpolation process can be achieved with several other surface interpolation methods, such as those based on multilevel B-splines [138] or *kriging* [132].

In Figures 4.5 and 4.6 a comparison of the utility surfaces proposal by Ribeiro [196] and the proposal for rule-based utility surfaces is depicted. These are based on the same settings used in Figures 4.3 and 4.4. The comparison is carried out using the isometrics illustration of the surface, and for example purposes, the relevance threshold t_R is defined as 0.9 (dashed line). As previously noted, this threshold defines a boundary for values of a given domain, which the user considers to be highly relevant.

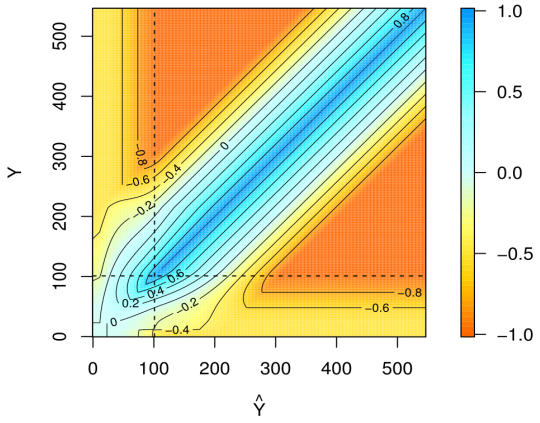


Figure 4.5: Example of utility surface as proposed by Ribeiro [196], with relevance threshold (dashed line).

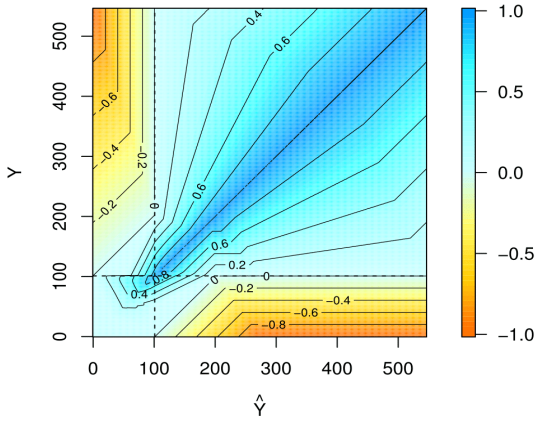


Figure 4.6: Example of rule-based utility surface with relevance threshold (dashed line).

Results show that the proposal by Ribeiro (left) may not provide the best fit for the exemplified scenarios. Due to the interpolation of utility based on the concept of maximum admissible loss, there can be negative values of utility in cases where highly relevant values are predicted as such. In contrast, the proposed approach (right) better accounts for the user preferences, and is capable of defining a utility function that better approximates the user-defined boundaries, i.e. relevance threshold.

[4]This approach is implemented in the R package **akima** [10].

Therefore, in regression tasks where, in addition to providing domain preferences, users also set a relevance threshold and establish that the usefulness of predictions is bounded by such a threshold, the proposed approach of rule-based utility functions is capable of providing a better approximation to the user preferences, when compared to the work of Ribeiro [196].

4.5 Evaluation Metrics for Imbalanced Learning

As previously noted, one of the crucial aspects of predictive modelling is the application of appropriate evaluation procedures, since the development and improvement of predictive models are driven by the reduction of forecasting errors. These are measured by employing evaluation metrics.

Concerning learning tasks with imbalanced domains, it is very noticeable that most of the efforts have been focused on classification problems. In such tasks, the common approach is to use the precision/recall evaluation framework [57] and their composite metric, F-Score (F_β). Given a data set from an imbalanced domain, the metric Precision provides an indication on how accurate the model is in predicting under-represented cases. As for the Recall metric, it signifies how frequently these rare cases were identified as such by the model. The F-Score metric [198], combines both previous metrics based on a β factor, which allows to define the importance attributed to Precision and Recall. The F-Score is often used in order to account for the known trade-off between the Precision and Recall measures, i.e. predicting all items as "rare" will result in a perfect Recall score, but a very poor score concerning Precision.

Nonetheless, regarding regression tasks with imbalanced domains, the efforts made concerning their appropriate evaluation have been negligible. In such tasks, researchers commonly resort to standard metrics such as the Mean Squared Error (MSE). These standard metrics assume uniform domain preferences and are solely focused on the magnitude of the prediction error. As such, they raise severe issues when evaluating imbalanced domain learning tasks.

In the financial domain, several authors have proposed evaluation methods capable of considering uneven costs of predictions, e.g. [49, 55]. Such proposals are based on asymmetric loss functions with the prime objective of differentiating between under- and over-predictions. The idea of these proposals is that such types of errors should be considered differently, i.e. different prediction costs for the same absolute error. However, these proposals prove to be inadequate for imbalanced domains learning given that they do not account for differentiated domain preferences of users.

Based on the concept of utility-based regression, Ribeiro [196] proposes several utility-based evaluation metrics, providing a framework for the assessment of predictive errors in imbalanced learning tasks. These metrics include the Mean Utility (MU) and the Normalized Mean Utility (NMU). The former (MU) allows the estimation of models' performance given

uneven domain preferences by users, which are depicted by relevance functions $\phi()$. The latter metric is a normalization of the Mean Utility. Given a set of paired predicted ($\hat{y}$) and true (y) values of size n and a utility function $u()$, these metrics are defined as follows:

$$MU = \frac{1}{n}\sum_{i=1}^{n} u_\phi(\hat{y}_i, y_i) \qquad (4.7) \qquad NMU = \frac{\sum_{i=1}^{n} u_\phi(\hat{y}_i, y_i) + n}{2n} \qquad (4.8)$$

In addition, based on the concept of the precision/recall framework, Ribeiro also proposes an adaptation of such metrics to the scope of numerical prediction tasks with imbalanced domains. This framework for utility-based regression considers both the user preference biases and the issue of numeric accuracy. This is carried out by resorting to the previously described concept of utility. Formally, Ribeiro [196] defines the Precision ($prec^u_\phi$) and Recall (rec^u_ϕ) metrics as follows:

$$prec^u_\phi = \frac{\sum\limits_{i:\hat{z}_i=1,z_i=1} 1 + u(\hat{y}, y)}{\sum\limits_{i:\hat{z}_i=1,z_i=1} 1 + \phi y_i + \sum\limits_{i:\hat{z}_i=1,z_i=0} 2 - p(1-\phi(y_i))} \qquad (4.9)$$

$$rec^u_\phi = \frac{\sum\limits_{i:\hat{z}_i=1,z_i=1} 1 + u(\hat{y}_i, y_i)}{\sum\limits_{i:z_i=1} 1 + \phi(y_i)} \qquad (4.10)$$

where p is a weight differentiating the types of errors, and $\hat{z}$ and z are flags associated with the presence of a highly relevant case.

In order to relax the definition of the metrics proposed by Ribeiro, an alternate definition of these metrics is proposed, by identifying normal and highly relevant cases solely resorting to relevance functions $\phi()$ and the respective relevance threshold t_R. As such, instead of using a binary flag in order to identify signalled (predicted as highly relevant) and true (highly relevant) cases, the signalled and true highly relevant cases are those where the relevance of their values is above the relevance threshold: $\phi(\hat{y}) > t_R$ and $\phi(y) > t_R$, respectively. Given this, the proposed alternative definitions of Precision and Recall for utility-based regression are as follows:

$$prec^u_\phi = \frac{\sum\limits_{\phi(\hat{y}_i)>t_R,\phi(y_i)>t_R} (1 + u(\hat{y}_i, y_i))}{\sum\limits_{\phi(\hat{y}_i)>t_R} (1 + \phi(\hat{y}_i))} \qquad (4.11) \qquad rec^u_\phi = \frac{\sum\limits_{\phi(\hat{y}_i)>t_R,\phi(y)>t_R} (1 + u(\hat{y}_i, y_i))}{\sum\limits_{\phi(y_i)>t_R} (1 + \phi(y_i))} \qquad (4.12)$$

where $\phi(y_i)$ is the relevance associated with the true value y_i, $\phi(\hat{y}_i)$ is the relevance of the predicted value $\hat{y}_i$, t_R is a user-defined threshold, above which cases are signalled as highly relevant for the user, and $u(\hat{y}_i, y_i)$ is the utility of making the prediction $\hat{y}_i$ for the true value y_i, normalized to $[-1, 1]$. It should be noted that the proposed alternate definition of Precision and Recall is motivated by the work of Branco [33] which also presents a different

definition for the utility-based metrics. However, unlike the original proposal by Ribeiro [196] and the traditional definition of Precision and Recall, in Branco's proposal, the numerators and denominators of the equations report to the same amount of true and predicted pairs of values. The alternate definition proposed in this section does not make such assumption, encapsulating the same assumptions on which both the traditional definition and the proposal by Ribeiro are based: the numerators of Precision and Recall concern the cases where highly relevant values were predict as such.

As for the F-Score metric for utility-based regression (F^u_β), it is based on the traditional definition of the metric, previously described in Section 2.4 (Equation 2.5). It combines both Precision ($prec^u_\phi$) and Recall (rec^u_ϕ) according to our alternate definition, with an harmonic mean, including a β factor which denotes the importance attributed to the components:

$$F^u_\beta = \frac{(1+\beta) \cdot prec^u_\phi \cdot rec^u_\phi}{\beta^2 \cdot prec^u_\phi + rec^u_\phi}. \tag{4.13}$$

4.6 Utility-Based Evaluation Framework

In this section, an evaluation framework for the imbalanced learning task of web content popularity prediction is proposed. Building on previous work concerning this issue, this proposal uses both standard and utility-based evaluation metrics. The motivation for such decision is related to allowing the comparison of the performance of models from multiple perspectives.

The proposed evaluation framework is composed of three evaluation metrics. Two of these metrics have already been described and formalized: the *i)* Root Mean Squared Error ($RMSE$) and the *ii)* utility-based F-Score (F^u_β). The former ($RMSE$) assumes uniform domain preferences, and is focused on determining the average squared error of models' predictions. The resulting score is given by the root of this average, in an attempt to mitigate the effects of extreme values of errors. This metric is commonly used in previous work concerning web content popularity prediction. The latter, utility-based F-Score (F^u_β), is a composite measure of the utility-based regression framework proposal for the Precision ($prec^u_\phi$) and Recall (rec^u_ϕ) metrics. The $RMSE$ and F^u_β metrics are defined in Section 2.4 (Equation 2.15) and in Section 4.5 (Equation 4.13), respectively.

In addition to these evaluation metrics, a novel evaluation metrics is proposed: the Relevance-Weighted Root Mean Squared Error ($RMSE_\phi$).

Relevance-Weighted Root Mean Squared Error

In order to provide a notion of the impact caused by non-uniform domain preferences in standard evaluation metrics, a relevance-weighted adaptation of the Root Mean Squared Error ($RMSE$) is proposed, focusing on prediction errors of highly relevant cases.

It should be noted that applying this weighted approach to all cases is prone to severe evaluation caveats [196], as it could ignore the impact of cases where true values are considered as normal w.r.t. their relevance, but are predicted with highly relevant values.

To illustrate this issue, consider the following scenario. Given a domain $\mathcal{Y}$, consider two true values, $y_1 = 0$ and $y_2 = 100$, where the former is signalled as "normal", with a relevance score of 0, and the latter as highly relevant with a relevance score of 1. Consider that their respective predictions are $\hat{y}_1 = 100$ and $\hat{y}_2 = 0$, representing the same absolute error. By weighing these cases according to the relevance of their true values, the first case would be discarded, despite representing a considerable error by the prediction model. Based on this insight, the proposed metric focuses on a subset of cases – those considered to be highly relevant – avoiding the issues raised by weighted evaluation metrics.

The Relevance-Weighted Root Mean Squared Error $RMSE_\phi$ is formalized in the following manner: given a paired set of true and predicted values from cases considered highly relevant, i.e. those with a relevance score $\phi()$ of the true value above the relevance threshold t_R, the proposed adaptation consists of weighting the average squared errors of these pairs, using the relevance score of the respective true values,

$$RMSE_\phi(\hat{y}, y) = \sqrt{\frac{\sum\limits_{\phi(y)>t_R} \phi(y) \times (\hat{y} - y)^2}{|y : \phi(y) > t_R|}} \tag{4.14}$$

where predicted and true values are respectively denoted as $\hat{y}$ and y; $|y : \phi(y) > t_R|$ denotes the number of highly relevant true values; $\phi(y)$ is the relevance of a given items' true value and t_R is the user-defined relevance threshold.

This proposal allows the evaluation of prediction errors' impact, when focusing on cases regarded by users' domain preferences as highly relevant. In comparison to the originally proposed $RMSE$, our proposal accounts for the users' domain preferences, denoted by the true values' relevance. However, it does not account for the relevance of the predicted values and the utility of such predictions, as considered by the utility-based F^u_β evaluation metric.

In summary, the proposed evaluation framework is characterized by the use of the following metrics, with different objectives:

- **Root Mean Squared Error** $RMSE$. A standard evaluation metric accounting for

squared prediction errors. The outcome is given by the root of the squared errors average. It assumes uniform domain preferences.

- **Relevance-Weighted Root Mean Squared Error** $RMSE_{\phi}$. Using user-defined relevance functions, and focusing on highly relevant cases, it weighs the squared error of predicted and true paired values with the relevance of the latter. Results are described by the root of the weighted squared errors average.

- **Utility-Based F-Score** F^{u}_{β}. A composite evaluation metric ($prec^{u}_{\phi}$, rec^{u}_{ϕ}) based on the definition of a relevance function, which evaluates the accuracy of models' predictions towards user-defined highly relevant cases, based on the concept of utility.

Finally, it should be noted that the proposed evaluation framework can be extended in order to include the factors (Precision and Recall) of the F-Score metric. Regarding parametrization, this evaluation framework depends on the definition of a relevance function $\phi()$ and the relevance threshold t_R to account for uneven domain preferences, and a β factor used by the F-Score metric, in order to weigh its respective Precision and Recall metrics.

4.7 Experimental Analysis

In this section an experimental analysis is provided, focusing on the target domain of this thesis: web content popularity. As stated in this chapter, we consider the task of web content popularity prediction to be an imbalanced domain learning task. As such, it assumes that both the following conditions are verified: *i)* the distribution of the target variable is highly skewed, and *ii)* the cases that are severely under-represented are those which require the most predictive accuracy.

This task can be framed in two scenarios: *a priori* and *a posteriori*. The former concerns the scenario where the evolution of the items' popularity one aims to predict is not available. This scenario is common when predicting the popularity of items before they are published, or shortly after. The latter, *a posteriori* prediction, refers to the scenario where the predictive task is focused on modelling the evolution of items' popularity, in order to forecast future popularity. In the experimental analysis provided in this section the focus is on *a priori* prediction tasks.

One of the most crucial procedures in the efforts to boost predictive accuracy in *a priori* tasks concerns feature selection. As analysed in the literature review provided in Chapter 2, several types of features[5] have been used in related works: *i)* behavioural, *ii)* social network, *iii)* content, *iv)* temporal, *v)* meta-data, *vi)* external sources. One of the main distinctions

[5] A thorough description of these types of features with illustrative examples is provided in Section 2.3 of Chapter 2.

between *a priori* and *a posteriori* tasks is that the former does not resort to behavioural features.

The objectives of the experimental analysis presented in this section are two-fold: *i)* to study the contribution of several types of features concerning predictive accuracy in *a priori* popularity prediction tasks, and *ii)* to analyse and discuss the conclusions provided by standard and utility-based evaluation metrics.

4.7.1 Materials and Methods

In this section, the materials and methods used to perform the experimental analysis are detailed. The description of the predictors used from each type of features, as well as the data set, is presented. The evaluation methods are motivated and presented, including their required parametrization. The learning algorithm used is described and the evaluation methodology motivated. Finally, results are presented and a discussion is provided.

Data

The experiments performed in this analysis use the data sets concerning online news feeds presented in the previous chapter. Online news feeds are a type of web content, with particular characteristics. The most important characteristic is that it has a very short alive-time, therefore requiring that prediction models are both quick and accurate in predicting the popularity of news items.

The two data sets were presented in Sections 3.2.1 and 3.2.2. The first is a single-source data set, and the second is a multi-source data set. The former includes news items obtained from the news recommender system Google News, and the respective popularity of such news in the social media source Twitter. The latter data set, concerns news obtained from both the Google News and Yahoo! News news recommender systems, and the popularity of such news items in the social media sources Facebook, Google+ and LinkedIn. In both data sets, from the news recommender systems, i.e. official media sources, the retrieved information involves the title, headline, publication date, the news outlet and the position of news in the respective source rankings. From social media sources, the amount of attention received by the news items, i.e. popularity, was queried in intervals of 20 minutes.

For the purposes of this experimental evaluation, the goal is to predict the final popularity of news items, i.e. the popularity accumulated by news until 2 days after their publication (prediction horizon), before such items are published. Therefore, behavioural features such as the popularity of the items at the moment of prediction will not be considered. Also, in this evaluation social network and external sources features are not considered. As previously discussed in Section 2.5, social network features raise several issues concerning privacy and

scalability of data access. As for external sources features, previous work by Martin et al. [161] shows that such features do not provide significant additional performance and, in the scope of this experimental evaluation, the focus is on data provided by official and social media sources.

The news items obtained in both data sets concern four topics: *economy*, *microsoft*, *obama* and *palestine*. These are very active topics, and they report to different types of entities: sector, company, person, and country, respectively.

Using the data available in this type of data sets, previous work provides evidence pointing to variables that have strong predictive power and those that do not. Bandari et al. [24] show that the news outlet (media source) of the news items is a strong predictor of popularity. Additionally, the authors use named entity identification, adding predictors concerning average behaviour (as to popularity) when certain entities are referenced in news. Textual features are used by Tsagkias et al. [224], by extracting the top-100 most discriminative terms and observing that they show a strong predictive performance. Subjectivity of language and sentiment analysis have also been used in numeric prediction tasks of news popularity. Bandari et al. [24] used subjectivity of language, reporting little predictive power by this predictor. In contrast, Berger et al. [29] show that sentiment analysis features are a good indicator of articles' virality. Also, concerning temporal features, Ahmed et al. [6] show that date and time information concerning news can be successfully employed as features in web content popularity prediction tasks.

Considering the indications provided by previous work, this study will focus on three types of predictive features: content, temporal and meta-data features. Concerning content features, the contribution of bag-of-words features and sentiment scores obtained by the text mining technique, sentiment analysis, is assessed. As for temporal features, the day of the week and the hour of news' publication are considered as predictors. Finally, regarding meta-data features, the popularity of entities in the title and headline of the news items', as well as that of news outlets is considered. The features are grouped by their type, and as such, this experimental evaluation will test seven different possible feature sets (combinations of each type of features) in each predictive scenario combining the available social media sources and news topics. Table 4.1 summarizes the predictive features described, and Table 4.2 presents the combinations of feature types tested in the experimental evaluation.

Type	**Variable**	**Description**
Content	$T_1, T_2, \cdots$	Frequency of terms from bag of words approach applied to headlines.
	SentTitle, *SentHeadline*	Sentiment scores of the news title and headline.
Temporal	$D_1, D_2, \cdots, D_7$	Day of the week the news was published (flag).
	$Hour$	Hour of news publication.
Meta-Data	*OutletAvg*	Average popularity of news' outlets.
	EntitiesAvg	Average popularity of entities mentioned in title and headline of news.

Table 4.1: Set of predictors tested in the experimental evaluation.

Feature Sets	Bag-of-Words	Sentiment Scores	Day of Week	Hour of Day	Outlet Popularity	Entities Popularity
C	✓	✓				
T			✓	✓		
M					✓	✓
CT	✓	✓	✓	✓		
CM	✓	✓			✓	✓
TM			✓	✓	✓	✓
CTM	✓	✓	✓	✓	✓	✓

Table 4.2: Feature sets with combinations of predictors tested in the experimental evaluation.

Feature Generation. The generation of features described in Table 4.1 is carried out using the following procedures.

- A standard bag-of-words approach is applied in order to obtain a set of terms describing it, using the infrastructure provided by the R package **tm** [79]. This is applied to the headline of the news[6], given that it provides a larger amount of information. Punctuation, numbers and stop words are removed. Sparse terms that appear in less than 1% of cases are also removed. Depending on the topic, this translates to between 100 and 200 terms describing news items' text content.

- The two sentiment scores, concerning the title and headline of news, are obtained by applying sentiment analysis techniques, using the sentiment lexicon described by Hu and Liu [107]. This is carried out using the framework provided by the R package **qdap** [199].

- Data concerning the weekday and hour of a given news publication are extracted directly from information retrieved from official media sources (Google News and Yahoo! News).

- The popularity of news outlets is obtained by averaging the amount of attention their respective news obtained. This is carried out by applying simple statistics to historical data, i.e. training set.

- The popularity of entities mentioned in news items is obtained by averaging the amount of attention obtained by news in which they were mentioned. To achieve this, the R package **openNLP** [105] is used. It allows one to recognize named entities based on language models. This process focuses on three types of entities: locations, people and organizations.

[6]The full news text was not considered as this would require following the available link to the original news site and have a specific crawler to obtain this text.

It should be stressed that these procedures are applied separately to each combination of social media source and topic of news. This allows for a more precise generation of features, accounting for the specific topics in which the prediction tasks are framed. As a reminder, the process for obtaining the popularity of the items in each of the social media sources is described in the previous chapter (Section 3.2).

Learning Algorithms

In order to analyse the contribution of the types of features mentioned, it is necessary to select the regression tools used to learn the predictive models. A similar study by Arapakis et al. [15] claims that the best results in numerical prediction tasks of web content popularity are given by linear models (**lm**). As such, in this experimental evaluation this modelling approach is used. To allow easy replication, the implementation of **lm** in the R package **stats** is applied.

Regardless, the performance of learning algorithms are highly dependent on the predictive features used. As such, it should be made clear that the outcome of this experimental evaluation is specific to the use of linear models, given the claim by Arapakis et al. [15] that such approach provides the best outcome. However, if another learning algorithm would be employed, it is possible that the results concerning the best feature set for the web content popularity prediction task would provide different conclusions.

Evaluation Metrics

Regarding the evaluation of prediction models' performance, the focus of our problem is accurately predicting the popularity of a small amount of cases, i.e. those which are rare due to their high level of popularity. As thoroughly described in this chapter, standard evaluation approaches are prone to issues in such predictive settings. Concerning the scope of our predictive task, the main issue with standard evaluation metric is that such metrics assume uniform domain preferences.

As such, the utility-based evaluation framework proposed in Section 4.6 is used. This framework includes three evaluation metrics: *i)* the root mean squared error $RMSE$, the relevance-weighted root mean squared error $RMSE_\phi$, and a utility-based F-Score F^u_β. Given our focus on highly popular items, experimental evaluations will focus on the latter two metrics, since those are focused on assessing the predictive accuracy of models towards such type of cases. The relation of such results will be compared and discussed with those obtained by the $RMSE$ metric.

The utility-based evaluation framework requires the definition of several parameters: *i)* a β factor determining the importance attributed to the factors (Precision and Recall) in the

metric F^u_β, and *ii)* the relevance functions noting the importance attributed by users to the values of the domain and the threshold for considering a case as highly relevant. Also, given the use of the proposal by Ribeiro [196] for the automatic definition of relevance functions, the boxplot coefficient α must also be defined.

Boxplot Coefficient. According to the seminal work on boxplot statistics by Tukey [227], outliers can be assigned to two different categories: *i)* mild outliers, and *ii)* extreme outliers. This categorization is related to the coefficient α used in order to define the boundary of the lower and upper whiskers[7]: $Q3 + \alpha \times IQR$, where $Q3$ is the third quantile of the distribution, and IQR is the difference between $Q3$ and $Q1$.

Given a domain $\mathcal{Y}$, mild and extreme outliers are defined as follows:

- Mild Outliers ($mOut$):

$$mOut := \{y \in \mathcal{Y} | y \leqslant Q1 - 1.5 \times IQR \vee y \geqslant Q3 + 1.5 \times IQR\} \tag{4.15}$$

- Extreme Outliers ($eOut$)

$$eOut := \{y \in \mathcal{Y} | y \leqslant Q1 - 3 \times IQR \vee y \geqslant Q3 + 3 \times IQR\} \tag{4.16}$$

Considering that our objective is to accurately predict the rare cases of highly popular news, i.e. extreme cases of high popularity, the focus of the experimental evaluation concerns extreme outliers. As such, the boxplot statistics coefficient α is defined as 3.

F-Score β Factor. As previously described, the F-Score is a composite evaluation metric, combining the scores of the metrics Precision and Recall with an harmonic mean. Precision indicates the model's accuracy on under-represented cases, and Recall specifies how frequently models correctly identify such cases.

The β factor determines the importance of the Recall measure in relation to Precision. Commonly, the β factor assumes one of three values: 0.5, 1 and 2. The first weighs Precision higher than Recall, and reduces the impact of false negatives[8]; the second weights Precision and Recall equally; and the third weighs Recall higher than Precision, thus accentuating the effect of false negatives.

Given the scope of our objectives, the accurate prediction of news items' popularity and the focus towards the rare cases of extreme popularity, the information provided by both Precision and Recall are equally important. As such, the value of β is defined as 1.

[7] As previously described, in the domain of web content popularity there is only one extreme of large numbers.

[8] In the scope of our predictive task, false negatives report to situations where a highly popular news item is predicted with a popularity level of a normal case.

Relevance Functions. The definition of relevance functions is crucial for the efforts of formalizing the described popularity prediction tasks as utility-based regression. In ideal scenarios, it is expected that the user would provide such relevance functions. These contain the users' domain preferences and the description of the importance attributed to target variable values.

As previously discussed, one of the main caveats raised by this ideal scenario concerns highly dynamic domains. For example, if a domain is prone to non-stationarities, affecting the distribution of its values, this would require the user to update the function regularly. Another example concerns scenarios where there is no previous domain knowledge, and therefore it is not possible to provide such specifications. The domain of web content popularity can be framed in both of these situations. To tackle this, Ribeiro [196] allows for the automatic definition of relevance functions based on the distribution of the target variable, using boxplot statistics. Such approach is used, in order to obtain relevance functions.

However, it is still necessary to provide an additional parameter related to the relevance of values: the relevance threshold. This threshold regards the users' boundary definition in order to consider items as highly relevant. In order to obtain an appropriate amount of highly relevant items, given that the accurate prediction of such cases is the focus of our predictive tasks, several values were tested: 0.5, 0.75 and 0.9. Considering the number of cases provided by the automatic relevance functions and each of these relevance threshold alternatives, the value of 0.9 was selected since it presents an appropriate (small) number of cases: results show that with a threshold of 0.9 the amount of cases considered as highly relevant is roughly 10%, for each combination of social media source and news topic.

Evaluation Methodology

The evaluation methodology used in experiments regards the approach applied to estimate the error that a predictive model incurs when applied to future data, i.e. unseen data. In related work concerning web content popularity prediction, the most popular methodologies are the out-of-sample method [213] and k-fold cross validation [93].

However, given the implicit temporal order of information commonly observed in web content data sets, approaches such as the out-of-sample method and k-fold cross validation raise severe caveats. Previous work [213] shows that the application of out-of-sample estimation may lead to unreliable estimates of the models' prediction errors. As for the latter, the problem of using such method is related to the temporal order of the data, since this method assumes that the data is independent w.r.t. the temporal dimension.

In this experimental context, one needs to be careful in terms of the process used to obtain

reliable estimates of the selected evaluation metrics. This means that the experimental methodology should make sure that the original order of the data is kept, so that models are trained on past data and tested on future data, avoiding over-optimistic estimates of their scores.

In order to obtain reliable estimates of the selected evaluation metrics for each of the alternative models (based on different combinations of predictive features' types), the Monte Carlo simulation method is used as the experimental methodology[9]. This methodology randomly selects a set of points in time within the available data, and then for each of these points selects a certain past window as training data and a subsequent window as test data, with the overall training+test process repeated for each random point. All alternative models are compared using the same training and test sets to ensure fair pairwise comparisons of the obtained estimates.

4.7.2 Results

This experimental evaluation is focused on assessing the contribution of different types of features in terms of their predictive accuracy. To evaluate the various types of features and their combinations, the previously described utility-based evaluation framework is applied, and the evaluation methodology employed is the Monte Carlo simulation. Results are obtained through 10 repetitions of a Monte Carlo estimation process with 50% of the cases used as training set and the subsequent 25% as test set. This process is done using the infrastructure provided by R package **performanceEstimation** [220].

Figure 4.7 presents an illustration of the results according to the F_1^u evaluation metric, concerning the combination of news data on each of the topics *economy*, *microsoft*, *obama* and *palestine* and the popularity of such news according to each of the social media sources available: Twitter in the single-source data set, and Facebook, Google+ and LinkedIn in the multi-source data set. Results concerning all metrics are presented in Annex A, where for each combination of social media source and news topic, the best result according to each evaluation metric is denoted in bold.

Results show that meta-data features provide the best predictive performance and that temporal features provide the worst. Focusing on the combinations of feature types, results show that the best predictive performance is obtained by approaches using a combination of meta-data and temporal features. However, approaches solely based on meta-data features provide a better evaluation than such combination. Experimental results also show that these conclusions are valid for most combinations of social media source data and news topic.

[9] An implementation of the Monte Carlo simulation method for R is provided by Torgo [219].

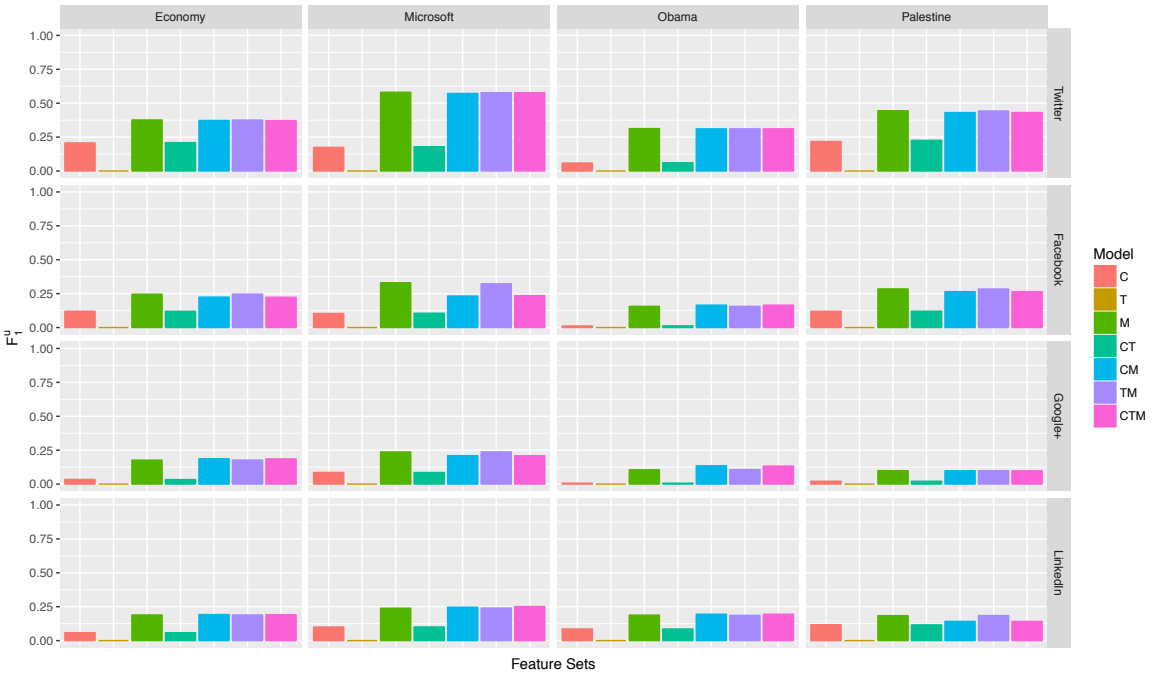


Figure 4.7: Evaluation of prediction models using distinct feature sets in all combinations of social media sources and news topics, according to the utility-based evaluation metric F_1^u.

Concerning the comparison of standard evaluation metric $RMSE$ and the utility-based metric F_1^u, results show that in the majority of cases these do not agree. This shows the impact that standard evaluation metrics may have when evaluating learning tasks where the objective is to accurately predict an underrepresented set of cases in the data. This observation is also valid for the relevance weighted version of $RMSE$ ($RMSE_\phi$). Indeed, in most cases, $RMSE_\phi$ does not agree with either the standard version of the evaluation metric ($RMSE$) or the F_1^u metric.

These discrepancies between the outcome of different evaluation metrics show the ability of the proposed evaluation framework in providing a broad and multi-faceted analysis of the predictive accuracy of regression models in this domain. It shows that, for most cases, the models with the best overall prediction error (i.e. uniform domain preferences), are not the models with the best numeric accuracy on highly relevant cases ($RMSE_\phi$). Also, it shows that such models do not present the best results in this experimental evaluation when accounting for both the relevance of the true and predicted values (F_1^u).

Despite this overall analysis of results, it is still not clear if the outcome of the models represent statistically significant performance differences. Therefore, critical difference diagrams [61] according to the Friedman test are applied, in order to further understand the difference between the prediction models tested, concerning the F_1^u metric. A lower rank represents better performance, and the horizontal lines connecting the methods show the

significance of the difference among ranks. Pairs of models not connected with a horizontal line indicate significant ($p-$value < 0.05) difference in their ranks for a given experiment. Results are depicted in Figure 4.8

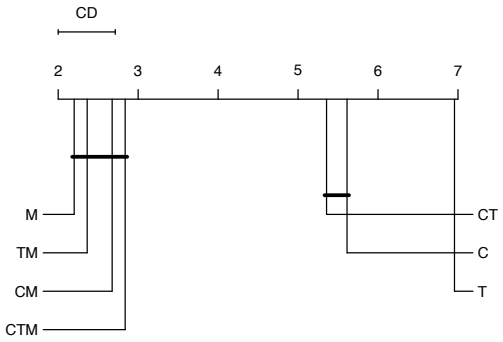


Figure 4.8: Critical difference diagram concerning the results of the evaluation metric F_1^u for models with different combinations of features.

According to the significance tests carried out, results show that models using meta-data features provide the best results, with statistical significance over models solely based on content and/or temporal features. Also, results show that models using content features provide a statistically significant advantage over models solely based on temporal features. Overall, best results are obtained by models solely based on meta-data features, although not presenting a significant improvement over feature sets combining content and/or temporal features with meta-data features. However, it should be noted that the models using such feature space show F_1^u scores that are non-optimal, with a global average of roughly 0.25. This shows the difficulty of standard learning algorithms in learning with imbalanced data, and in accurately identifying the items that are more relevant for the users, in the context of *a priori* prediction.

4.7.3 Discussion

Based on the definition of the popularity prediction task as an imbalanced learning problem, this experimental evaluation is focused on two objectives: *i)* to provide a comparison between standard and utility-based evaluation metrics, and *ii)* to derive conclusions as to the best feature space in order to predict highly popular news items.

As previously stated, related work shows how standard evaluation metrics are prone to misleading results. This issue is related to their assumption of uniform domain preferences, i.e. every case is equally important. To overcome this problem, a robust evaluation framework combining standard and utility-based evaluation metrics is proposed.

Results of the utility-based F-Score (F_1^u) show that concerning the prediction of highly

popular items, the models which provide the best predictive accuracy towards important and under-represented cases are based on a feature space solely composed of meta-data features (average popularity of named entities mentioned and of the news outlet). This metric is the most robust of the framework, accounting for the numerical error of predictions, and the relevance of both the true and predicted values.

However, it should be noted that a comparison of results concerning the utility-based F_1^u metric and both the standard evaluation metric $RMSE$ and the relevance-weighted $RMSE_\phi$, which does not account for the relevance of predicted values and the utility of such predictions, show some discrepancy regarding the top performing models. In Figures 4.9 and 4.10, critical difference diagrams [61] are depicted, illustrating the statistical significance ($p-$value < 0.05) between the prediction models using the different combinations of features. The diagrams concern results obtained by evaluation metrics $RMSE$ and $RMSE_\phi$, respectively, in the previously detailed experimental evaluation conditions.

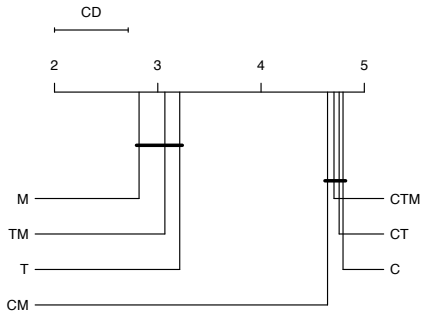


Figure 4.9: Critical difference diagram concerning the results of the evaluation metric $RMSE$ for models with different combinations of features.

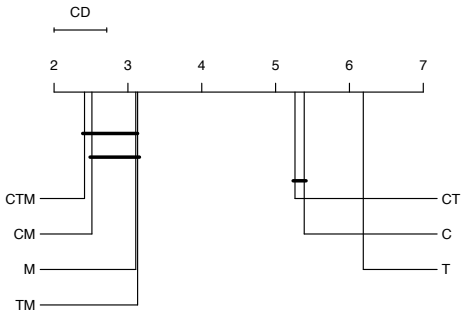


Figure 4.10: Critical difference diagram concerning the results of the evaluation metric $RMSE_\phi$ for models with different combinations of features.

According to the results of the statistical tests, it is observed that the conclusions according to $RMSE$ and $RMSE_\phi$ metrics are different from those obtained with the F_1^u metric. Concerning the $RMSE$ metric, results show that models based on meta-data and/or temporal features provide a significant advantage over all other combinations, and that models using content features obtain the worst results. As for the $RMSE_\phi$ metric, results are similar to those obtained by the F_1^u, with small differences as to which provides the best outcome (although without statistical significance). These differences highlight the impact of solely accounting for the relevance of the true values and the numerical error, and the additional impact of also accounting for the relevance of predicted values and the utility of such predictions, as done by the F_1^u metric.

Regarding the sets of features used in the predictive modelling tasks, the conclusions derived from this experimental evaluation require a comparison with previous work. As previously

described, several authors have provided insights on the predictive ability of features in popularity prediction tasks. Bandari et al. [24] show that the popularity of media outlets are strong predictors of popularity, whilst Tsagkias et al. [224] show evidence of discriminative terms providing good predictive performance. Sentiment analysis has shown [29] to be a good indicator of virality, and Arapakis et al. [15] prove that temporal features are well correlated with articles' popularity in Twitter.

In comparison, our results show that content features (bag-of-words and sentiment analysis), although providing average results (in comparison to the best performing feature sets), do not provide a good predictive performance, which is also the conclusion in the work of Martin et al. [161]. Also, regarding temporal features, it is noted that these provide the worst predictive performance, in contradiction with the conclusions by Tsagkias et al. [224]. Concerning the meta-data features, experimental results provide similar evidence to that of Bandari et al. [24], where media outlets' and named entities' popularity provide a strong predictability of popularity.

However, one of the features used in the experiments provides a singular contradiction with all previous work: sentiment scores. The use of sentiment analysis and features derived from such approach, have shown throughout previous experiments to be good predictors [29, 161, 18].

Given this contradiction, the experimental evaluation is repeated in order to provide further evidence to support the evaluation outcome, with the aim of comparing the predictive performance of models using solely meta-data features, and their combination with sentiment scores of both the title and headline. In order to reduce the bias of using a given sentiment lexicon, results include the use of four different lexicons, previously described in Section 3.3: AFINN [177], SentLex [107], SentiStrength [218] and SentiWordNet 3.0 [21]. Due to the extent of the results, they are described in Annex B (best results denoted in bold). In Figures 4.11 and 4.12 critical difference diagrams are presented in order to assess the statistical significance ($p-\text{value} < 0.05$) of the models.

Results show that according to both $RMSE$ and F_1^u evaluation metrics, the models based on meta-data and sentiment scores' features provide the best outcome. Results also show that all models combining both types of features provide a statistically significant advantage over models solely based on meta-data features. As for the best combination, results concerning the F_1^u metric show that this is obtained by using the SentiWordNet lexicon to obtain sentiment scores, despite not presenting a significant advantage over the second best combination (when using the AFINN lexicon).

In summary, the experimental evaluation carried out in this section shows that meta-data features provide the strongest predictability of news items' popularity, and temporal features the worst. Also, regarding the combination of features, results show that the best

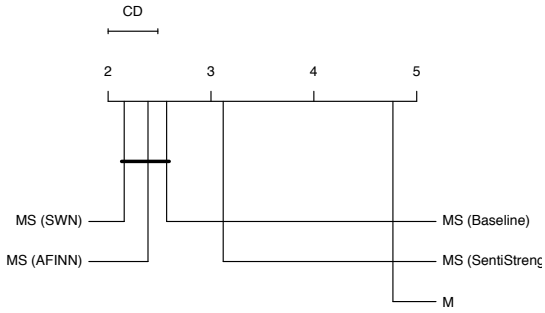


Figure 4.11: Critical difference diagram concerning the results of the evaluation metric F_1^u for models using meta-data and sentiment scores' features.

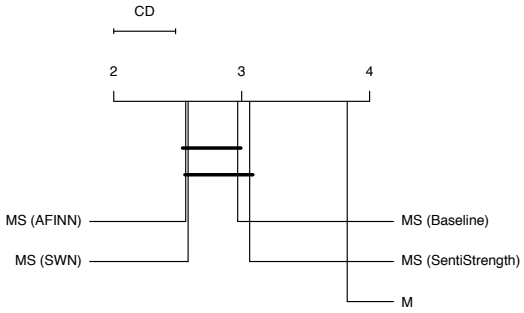


Figure 4.12: Critical difference diagram concerning the results of the evaluation metric $RMSE$ for models using meta-data and sentiment scores' features.

performing models are observed when combining meta-data features (average popularity of named entities and news outlets) and sentiment scores derived from the application of sentiment analysis to the title and headline of news items, using the SWN sentiment lexicon. Concerning the proposed evaluation framework, results show how standard evaluation metrics are prone to ambiguous optimization of models, and that the improvement of results w.r.t. standard evaluation metrics may not signify an improvement of predictive accuracy in under-represented cases. In fact, results show that the optimization of models according to the $RMSE$ metric may lead to a deterioration of prediction models' ability to accurately forecast rare cases of highly relevant items.

It should be stressed that such conclusions are specific to the application of linear models, following the results presented by Arapakis et al. [15] where it is concluded that such learning approach provides the best outcome in web content popularity prediction tasks. Given that learning algorithms are greatly influenced by the feature sets used, it is possible that if a different algorithm would be used, the conclusions would also be different.

4.8 Conclusions

In this chapter the problem of predicting the popularity of web content is formalized as a numeric predictive modelling task. As previously observed, this domain is severely skewed, given that the majority of items at a given time have a low amount of popularity, and a small set of rare (and relevant to the user) cases has an abnormal (high) level of popularity, i.e. imbalanced domain.

To tackle the task of web content popularity prediction, related work has proposed a considerable amount of approaches. However, their commonality is the assumption of uniform user domain preferences. Unlike previous work, such an assumption is not made here. The

problem of web content popularity prediction is defined as a non-standard learning task, framed within the concept of imbalanced domain learning.

Learning with imbalanced domains is based on two conditions: *i)* the user is focused on the predictive accuracy of a subset of target values, and *ii)* that subset is severely under-represented in the data. It is argued that this is the case in the web content popularity domain. As such, in this thesis, it is established that the focus of popularity prediction tasks concern under-represented cases of highly popular items, since those are the ones which should likely be recommended to end-users.

To solve this problem, the concept of utility is described, as well as its application to data mining and machine learning problems. Previous work has mainly focused on this problem within the scope of classification tasks, i.e. nominal target variables, and the related work concerning regression is negligible. The proposal of Ribeiro [196] concerning utility-based regression is presented. This proposal allows for the definition of a framework capable of formalizing the problem of web content popularity prediction with non-uniform domain preferences. It is based on two concepts: *i)* relevance functions and *ii)* utility surfaces. The first, relevance functions, allows users to attribute relevance scores to values of the target domain. It presents a relaxation of the "problem-definition issue" defined by Weiss [236] where the quality of information provided by users has an impact on the evaluation, comparison and optimization of prediction models. The latter, utility surfaces, allows the evaluation of prediction errors of models, not only based on numerical error, but also considering the relevance of both true and predicted values.

Based on the shortcomings of the proposal by Ribeiro concerning utility surfaces in the context of web content popularity prediction, a new approach to deriving them is proposed, based on rules knowledge instead of a maximum loss criterion. Also, a robust evaluation framework is proposed in order to provide a broad and multi-faceted evaluation of popularity prediction tasks. This framework includes a novel evaluation metric, focused on assessing the numerical error of items considered by users as highly relevant. This evaluation framework allows to understand the impact of using a standard evaluation metric such as $RMSE$ and its shortcomings when focusing on the prediction errors of highly relevant cases ($RMSE_\phi$). Furthermore, the use of the utility-based evaluation metric F_1^u is able to illustrate the impact of accounting for both the numeric error and the relevance of the true and predicted values by the models.

Using the proposed evaluation framework, an experimental analysis is carried out focused on the problem of *a priori* popularity prediction tasks. The experimental evaluation uses the data concerning online news feeds, a type of web content, thoroughly described in the previous chapter. The objectives of the experimental analysis are two-fold: *i)* to analyse the predictive ability concerning highly popular items of different types of predictive features, and *ii)* to compare and discuss the impact of standard and utility-based evaluation metrics.

Concerning the first objective, experimental results show that the best performance is obtained by models based on a feature set composed of meta-data features (average popularity of named entities and of news outlets) and sentiment scores of the news title and headline, using the SWN sentiment lexicon. As for the second objective, results show that the use of standard evaluation metrics is prone to ambiguous conclusions when the objective of the predictive tasks is accuracy on under-represented items. Concretely, it is observed that in some cases, standard evaluation metrics consider ineffective prediction models w.r.t. highly popular cases as being within the group of top performing models. Notwithstanding, it should be noted that these conclusions are based on the use of linear models, following the work of Arapakis et al. [15]. As such, given the sensitivity of learning algorithms to different feature sets, such conclusions could be different when applying other algorithms.

In this chapter, the problem of standard evaluation metrics and their impact on predictive tasks focusing on under-represented cases is thoroughly discussed. However, the issue of standard learning tasks focusing on accuracy towards the average behaviour of the data, i.e. the majority of cases (items with low levels of popularity), as well as its implications in predictive tasks such as those described in this chapter, is not addressed. In the following chapter, such issues are discussed. In addition, several proposals for imbalanced domain learning approaches are described, for both *a priori* and *a posteriori* prediction of web content popularity. Using such proposals, an extensive experimental evaluation is carried out, providing a comparison with state-of-the-art approaches in order to assess their ability to accurately predict the popularity of rare cases of highly relevant web content items.

Chapter 5

Popularity Prediction Models

In this chapter, the use of standard learning algorithms in imbalanced domain learning tasks is addressed. Previous work suggests that the application of such algorithms in learning tasks with imbalanced domains are prone to several issues. In order to tackle such issues, previous work concerning imbalanced domain classification tasks have addressed this issue extensively. However, concerning regression tasks, related work is negligible. Framed within the imbalanced domain of web content popularity, this chapter presents several proposals to tackle the issues associated to the use of standard learning algorithms when the distribution of the data is skewed, and the user domain preferences target under-represented cases. An extensive experimental evaluation is carried out concerning a priori and a posteriori prediction tasks, comparing the ability of the proposed approaches in accurately predicting the popularity of highly popular items to several state-of-the-art baselines.

5.1 Introduction

Solving the predictive task of anticipating the popularity of web content has received increasing attention over past years. The interest in solving this task may be observed by the quantity and diversity of approaches proposed to solve this problem. However, as thoroughly analysed in the previous chapter, the majority of related work has neglected the impact of an important characteristic of the web content popularity domain: the skewness of its distribution. This translates to an imbalanced distribution setting, where most of the web content items receive low levels of popularity and only a small set of such items gather a high degree of popularity.

The uneven distribution of popularity could not be a problem in the context of web content data, if the under-represented items were not relevant. However, intuitively, the objective of prediction tasks involving such domain is mostly concerned with accurately predicting such under-represented cases, i.e. those with high popularity levels. These are the items that

will be suggested to users, or automatically promoted to improve user-experience in online platforms.

In the former chapter, the dimension of evaluating imbalanced domain learning tasks was addressed and solutions to tackle the shortcomings presented by standard evaluation metrics were described. In this chapter, the impact of commonly used standard learning algorithms to solve web content popularity prediction tasks is discussed. Novel proposals to improve the prediction of highly popular web content items are presented, and an extensive experimental evaluation is carried out including state-of-the-art approaches.

5.2 Strategies for Imbalanced Domains

Standard learning algorithms commonly optimize models by attempting to reduce a given standard evaluation metric, which is focused on the average behaviour of the data. Such approach may lead to models which are specialized towards the well-represented cases of the data, causing the models to obtain a sub-optimal performance towards under-represented cases. The reasons for the impact of imbalanced domains in standard learning algorithms is mostly related to the following reasons:

1. Using standard evaluation metrics may provide misleading results, causing the models to be optimized towards the average behaviour of the data;
2. Standard learning algorithms may disregard under-represented cases due to their small coverage;
3. Algorithms may denote these rare cases as noise, discarding them from the learning process.

The problems posed by standard learning algorithms in imbalanced domain learning tasks have been studied for over two decades, mostly concerning classification tasks. Several surveys on this topic have been published (e.g. [155, 128, 35]), thoroughly describing approaches to overcome such difficulties, and providing important insights that can be applied to the problem of web content popularity prediction.

According to Branco [35], there are four main approaches to learning tasks when using data from imbalanced domains: *i)* data pre-processing methods, *ii)* special-purpose learning methods, *iii)* prediction post-processing methods, and *iv)* hybrid methods. The first type of methods operate by modifying data sets in order to provide more balanced distributions. The second type attempts to relax the bias towards majority cases by modifying existing learning algorithms and adapting them to imbalanced distributions. The third type, manipulates the

predictions of models according to both user preferences and the imbalance of the data. The fourth type of methods work by combining both the first and second types of methods.

In this thesis the prediction post-processing methods are not studied. As such, the objective is to propose several approaches for data pre-processing methods, special-purpose learning methods and hybrid methods, and investigate if they are capable of improving the ability of models in predicting highly relevant cases in web content prediction tasks. Such cases are the most important concerning an accurate and early prediction, as they are the ones that should be placed in the top positions of suggestions to users. For simplicity, the three types of approaches will be referred to as data-level methods, algorithm-level methods and hybrid methods, respectively. In the following sections, each of these methods are described.

Data-Level Methods

Methods operating at a data-level are applied as pre-processing procedures[1]. The objective of the methods is to change the distribution of the training sets in order to balance rare (minority) and normal (majority) cases, following user preferences. As such, instead of applying a given learning algorithm directly to the original training data, this data is first pre-processed, providing a new data set.

Data-level methods are independent w.r.t. the learning algorithms used, and therefore provide a broad applicability to various predictive scenarios, presenting an effective solution to the imbalance problem [76, 80]. However, the main issue concerning such methods is that it requires considerable efforts regarding parametrization. This is necessary as to discover the optimal new distributions, capable of accurately translating the preferences of users, and maximizing predictive accuracy towards the target cases.

Different approaches have been proposed in previous work concerning data-level methods, which can be clustered in three major groups: *i)* undersampling, *ii)* oversampling, and *iii)* a combination of the previous approaches.

Undersampling approaches operate by reducing the number of examples which are considered as normal. These cases compose the majority of the data set. By reducing the amount of such cases, the objective is to aid standard learning algorithms in better capturing the dynamics of under-represented (rare) cases. Oversampling approaches are based on the generation of new examples concerning under-represented cases, which can be obtained by simply replicating existing cases, or by interpolating new cases based on other under-represented cases. Finally, the third type of data-level methods (hybrid methods) combine undersampling and oversampling approaches, attempting to balance the distribution by simultaneously reducing the number of normal cases, and adding new examples of under-

[1] Data-level methods are also commonly denoted as resampling strategies.

represented cases.

Concerning previous work, the main distinction between the data-level proposals pertain to case selection procedures. Commonly, the selection of cases to remove (undersampling) or to add (oversampling) in the new data set, is carried out randomly [66, 76]. However, several caveats may be raised by the random selection of cases. Namely, concerning its use in undersampling, this could lead to the removal of relevant cases for the learning process. As for the case of oversampling, since it commonly consists of replicating existing examples, this could increase the likelihood of overfitting.

To avoid such issues, other approaches to undersampling and oversampling have been proposed. These include distinct case selection procedures, such as those based on *i)* the distance between cases [30], *ii)* the clusters obtained from the data [113] or *iii)* those focusing on the recognition of cases from the majority class [110]. In addition, one of the most sophisticated approaches in data-level methods is the "Synthetic Minority Ovesampling Technique" [45] (SMOTE). This approach is focused on generating new under-represented (rare) cases by interpolating the existing ones. As such, instead of replicating such cases (as in traditional oversampling proposals), it generates synthetic cases, thus reducing the hazard of overfitting.

Algorithm-Level Methods

In order to relax the bias of standard learning algorithms towards the most common type of examples in data sets, algorithm-level methods provide modified versions of such algorithms. These approaches are considered as special-purpose learners, where knowledge concerning the domain is introduced into the learning process, biasing the algorithms towards users' preferences. This requires an extensive knowledge of the learning algorithm, and the understanding of the reasons related to its failure in accurately predicting under-represented cases.

The most common approach to algorithm-level methods consists of modifying preferences criteria of learning algorithms, incorporating costs and/or benefits in order to detail the utility of the predictions, i.e. cost-sensitive learning [72]. Examples of such approach include applications to algorithms such as decision trees [159], k-Nearest Neighbours [27], kernels [243], Support Vector Machines [7] (SVM) or neural networks [11]. Also, this approach has been applied to ensembles of SVMs [232] and Random Forests [46], as well as boosting proposals [116], among others.

Hybrid Methods

Hybrid methods combine methods at a data- and algorithm-level. Their objective is to potentiate the strengths and to reduce the shortcomings of the previously detailed meth-

ods. One of the first proposals concerning hybrid methods is presented by Estabrooks and Japkowicz [75]. This proposal is based on a framework that combines 10 classifiers using oversampling and 10 classifiers using undersampling methods, with different sampling percentages. A simple heuristic scheme is applied to combine the methods and to discard classifiers that are considered unreliable. Phua et al. [187] proposes the application of undersampling on several partitions of the data set, which are then individually used to train models using several classifiers. Classifiers trained by the same algorithm are combined using the bagging technique, and predictions are obtained by applying the stacking technique to combine the various classifiers. Similarly, Liu et al. [154] also propose the combination of data-level approaches with both bagging and boosting techniques, creating an "ensemble of ensembles".

Several contributions to the semi-supervised strategy of active learning can also be framed within the concept of hybrid methods. These contributions are commonly described by the combination of undersampling methods and SVM classifiers [74, 259, 73]. Nonetheless, research has shown that models based on the active learning strategy show a performance degradation as the domain imbalance increases [19].

5.2.1 Strategies for Regression Tasks

In spite of the interest that the problem of imbalanced domain learning has received for over two decades, proposals concerning this problem are mostly related to classification tasks, with emphasis in binary imbalanced classification tasks [211]. Concerning the proposal of strategies within the scope of numerical prediction tasks, such as regression, the attention it received is negligible. Nonetheless, a small number of proposals has provided important insight concerning the applicability of such strategies in continuous domains.

Concerning data-level methods, Torgo et al. [222] propose an adaption of the random undersampling method and of the SMOTE method for regression tasks (SMOTEr). The authors formalize the problem as a utility-based regression task, described in Section 4.3. Using such formalization and the concept of relevance functions, where cases are described by a given relevance score (Section 4.3.1), and considered "normal" or "rare" based on a given relevance threshold value. Concerning the proposed SMOTEr method, this combines the random undersampling method with the synthetic generation of new rare cases. Also, Branco et al. [34] propose an adaption of the random oversampling method, for which the objective is to randomly select rare cases and to replicate them, in order to provide a more balanced data set. As for algorithm-level methods, some approaches have been presented concerning regression trees, by proposing to change their splitting criterion [221, 197].

Motivation

Before proceeding, it is important to clarify how these methods for tackling imbalanced learning tasks may be applied within the context of web content popularity prediction tasks. As previously stated, such tasks are presented in two scenarios: *a priori* and *a posteriori* prediction.

In *a priori* tasks of web content popularity prediction, the predictive modelling approaches are solely based on leveraging features related to descriptors of the content. One of the main problems in providing prediction models which are accurate in predicting the rare cases of highly popular content concerns the diversity of the data. For example, concerning online news feeds data, it is possible to have a considerable number of similar or even identical news, for which the only difference is the time they were published and the respective news outlet, where all but one obtains a minimal level of popularity. Given the available methods for imbalanced domain learning, data-level methods provide an interesting approach, by reducing the number of cases that hinder the ability to accurately predict the target cases, or by providing a greater coverage for such cases.

For *a posteriori* tasks of web content popularity prediction, the issues which hinder the predictive ability of models is slightly different from *a priori* tasks. Such predictive modelling tasks are commonly based solely on observations of items' popularity in consecutive periods. A major difficulty in such predictive scenario is that the models must be able to provide a prediction of the final value of popularity with different levels of available data, and in the first moments, it may be difficult to distinguish which observations relate to cases that will obtain a high level of popularity. Given the focus of standard learning algorithms in optimizing the predictions of models towards the average behaviour of the data, these will only be able to detect the highly relevant cases that obtain an extremely high level of popularity in a very short amount of time. However, this problem can be tackled if the the focus of such learning algorithms is altered, in order to be more sensitive to highly relevant cases. As such, in order to tackle the *a posteriori* prediction tasks, it is proposed the use of algorithm-based methods, which are focused on altering learning algorithms in order to relax the influence of well-represented cases in the predictions.

Notwithstanding, it should be noted that the distinction of *a priori* and *a posteriori* tasks may be artificial in many situations. For example, once a news item is published, there is no social feedback from users available. Therefore, only an *a priori* prediction approach is possible. Conversely, if some time has passed since the items' publication, the level of available observations will probably be enough to provide an accurate prediction of the final value of popularity. However, in the first moments after a news item is published, given a minimal amount of popularity observations, it is indeed possible to approach the problem in both *a priori* and *a posteriori* scenarios. As such, in order to explore this approach, hybrid

methods are proposed in order to study the predictive ability of approaches combining data-level and algorithm-level methods.

Concerning the data-level methods proposed in this thesis, such methods are focused on taking advantage of the context of web content data, namely its temporal order and the relevance of the items. Instead of using a random approach to the selection of cases for under- and/or oversampling, the proposed methods incorporate the notion of temporal and relevance bias.

Regarding algorithm-level methods, new proposals are presented concerning altered versions of kernel and k-nearest neighbour methods, attempting to improve the predictive accuracy towards under-represented cases in comparison to state-of-the-art approaches.

Finally, concerning hybrid methods, a time-based ensemble method is proposed in order to combine approaches using the data- and algorithm-level methods proposed in this thesis. This proposal is focused on combining such approaches in a time-dependent manner, in order to account for the strengths and issues of both types of approaches.

5.3 Context-Bias Resampling Strategies

Traditional approaches of data-level methods, also known as resampling strategies, operate by randomly selecting cases that are removed from and/or replicated in the original data set. This leads to a more balanced data set, aiding standard learning algorithms to improve their predictive accuracy towards under-represented cases. However, as previously stated, the random case selection procedure is prone to several issues such as overfitting and the removal of relevant cases.

Data from the domain of web content popularity is commonly associated to contextual information, such as the platform in which it is published or its topic. Additionally, this type of data also contains a temporal dimension, denoted by time stamps, e.g. publication date. Concerning the predictive ability of models when focusing on examples of web content data such as online news feeds, a case can be made as to recent news items providing a better contribution to the predictive accuracy of models than older items: given the fast paced evolution of news stories, such items are constantly replaced by more recent events.

However, solely relying on the recency of events can also be misleading, because it could lead to a situation where normal cases are favoured solely because of their recency. Therefore, a second case can be made concerning recent and highly relevant news items providing a better contribution to the predictive accuracy of models rather than older and less relevant items.

Based on these claims regarding the impact of the temporal dimension of web content items

and their relevance scores in terms of the contribution to optimizing data-level methods, the following hypothesis are asserted:

Hypothesis 1 *The use of resampling strategies significantly improves the predictive accuracy of prediction models on imbalanced domains in comparison to the direct application of standard learning algorithms.*

Hypothesis 2 *The use of bias in case selection of resampling strategies is able to improve the predictive accuracy of prediction models in imbalanced domains, in comparison to non-biased strategies.*

Given such hypotheses, this section describes the proposal of context-bias resampling strategies. Such strategies are motivated by the claim that the temporal order of web content items should be taken into account when altering the distribution of training sets. By considering such order, it is possible to introduce a bias in the case selection process of resampling strategies. This proposal is presented as an extension of well-known resampling strategies: random undersampling, random oversampling and SMOTEr [222, 34]. For each strategy, two alternatives are proposed: undersampling, oversampling and SMOTEr with *i)* temporal bias, and *ii)* with temporal and relevance bias.

It should be reminded that these proposals are framed within the concept of utility-based regression. As such, it is assumed that users provide a relevance function $\phi()$ denoting their domain preferences or a distribution-driven procedure is applied to automatically define one. Also, it is assumed that users provide a relevance threshold t_R in order to define which target values are highly relevant. Given such information, it is possible to obtain two subsets of a target variable Y: *i)* a subset with all the highly relevant cases $Y_R = \{y \in Y : \phi(y) > t_R\}$ and ii) a subset with the remaining cases which are considered as having a normal relevance, $Y_N = Y \setminus Y_R$. The random undersampling, random oversampling and SMOTEr algorithms are described as follows.

5.3.1 Non-Biased Strategies

The random undersampling (**U_B**) strategy is described in Algorithm 1. This approach has the default behaviour of balancing the number of normal and rare values by randomly removing normal cases. The algorithm also allows the specification of a particular undersampling percentage by defining the parameter u. When the user sets this percentage, the number of cases removed is calculated w.r.t. to the amount of normal cases in the data. The percentage of undersampling $0 < u < 1$ defines the percentage of normal cases that are maintained in the new data set.

The random oversampling (**O_B**) approach is described in Algorithm 2. In this strategy, the default behaviour is to balance the number of normal and rare cases with the introduction

Algorithm 1 Random Undersampling (**U_B**).

function RANDUNDER($D, Y, \phi(Y), t_R, u$)
  // D - A data set
  // Y - The target variable
  // $\phi(Y)$ - User specified relevance function
  // t_R - The threshold for relevance on y values
  // u - (optional parameter) Percentage of undersampling

  $D_N \leftarrow \{D_i : \forall y_i \in Y, \phi(y_i) \leq t_R\}$ // Cases considered as normal
  $newData \leftarrow D \backslash D_N$ // Highly relevant cases are kept in the new data set
  if u **then**
    $TgtNr \leftarrow |D_N| \times u$
  else
    $TgtNr \leftarrow |D_N| \times \frac{|D \backslash D_N|}{|D_N|}$
  end if
  $selNormCases \leftarrow$ SAMPLE($tgtNr, D_N$) // randomly select a number of normal cases from D_N
  $newData \leftarrow c(newData, selNormCases)$ // add the normal cases to the new data set
  return $newData$
end function

of replicas of randomly chosen highly relevant cases. An optional parameter $o > 1$ allows the user to select a specific percentage of oversampling to apply to highly relevant cases.

Algorithm 2 The Random Oversampling algorithm (**O_B**).

function RANDOVER($D, Y, \phi(Y), t_R, o$)
  // D - A data set
  // Y - The target variable
  // $\phi(Y)$ - User specified relevance function
  // t_R - The threshold for relevance on y values
  // o - (optional parameter) Percentage of oversampling

  $D_R \leftarrow \{D_i : \forall y_i \in Y, \phi(y_i) > t_R\}$ // Cases considered as highly relevant
  $newData \leftarrow D \backslash D_R$ // Normal cases are kept in the new data set
  if o **then**
    $tgtNr \leftarrow |D_R| \times o$
  else
    $tgtNr \leftarrow |D_R| \times \frac{|D \backslash D_R|}{|D_R|}$
  end if
  $selRareCases \leftarrow$ SAMPLE($tgtNr, D_R$) // Randomly select a number of rare cases from D_R
  $newData \leftarrow c(newData, selRareCases)$ // Adds the replicas of rare cases to the new data set
  return $newData$
end function

The third strategy, SMOTEr (**SM_B**), is an adaptation of the original SMOTE [45] method for regression tasks, proposed by Torgo et al. [222]. It combines random undersampling with oversampling through the generation of synthetic cases. The random undersampling part is carried out through the process described in Algorithm 1. The oversampling strategy generates new synthetic cases by interpolating a seed example with one of its k-nearest rare case neighbours.

Algorithm 3 shows the process for generating synthetic examples and Algorithm 4 describes the **SM_B** strategy. By default, this strategy balances the number of normal and rare cases in a data set. Alternatively, the user may set the percentages of under/oversampling to be

applied using parameters u and o.

Algorithm 3 Generating synthetic cases.

function genSynthCases(D, ng, k)
  // D - A data set
  // ng - Number of synthetic cases to generate for each existing case
  // k - The number of neighbours used in case generation

  $newCases \leftarrow \{\}$
  for all $case \in D$ **do**
    if $|D \setminus \{case\}| < k$ **then** // Less examples than number of neighbours required
      $nns \leftarrow$ kNN($|D \setminus \{case\}|, case, D \setminus \{case\}$)
    else
      $nns \leftarrow$ kNN($k, case, D \setminus \{case\}$) // k-Nearest Neighbours of $case$
    end if
    for $i \leftarrow 1$ **to** ng **do**
      $x \leftarrow$ randomly choose one of the nns
      $new \leftarrow \{\}$ // A new synthetic case
      for all $a \in$ attributes **do** // Generate attribute values
        $diff \leftarrow case[a] - x[a]$
        $new[a] \leftarrow case[a] + \text{random}(0,1) \times diff$
      end for
      $d_1 \leftarrow$ dist($new, case$) // Decide the target value
      $d_2 \leftarrow$ dist(new, x)
      $new[y] \leftarrow \frac{d_2 \times case[y] + d_1 \times x[y]}{d_1 + d_2}$
      $newCases \leftarrow newCases \bigcup \{new\}$
    end for
  end for
  return $newCases$
end function

Algorithm 4 SMOTEr algorithm (**SM_B**).

function SMOTEr($D, Y, \phi(Y), t_R, k, o, u$)
  // D - A data set
  // Y - The target variable
  // $\phi(Y)$ - User specified relevance function
  // t_R - The threshold for relevance on y values
  // k - The number of neighbours used in case generation
  // o, u - (optional parameters) Percentages of over and undersampling

  $D_R \leftarrow \{D_i : \forall y_i \in Y, \phi(y) > t_R\}$ // Cases considered as highly relevant
  $D_N \leftarrow D \backslash D_R$ // Cases considered as normal
  $newData \leftarrow \{\}$
  if u **then** // Apply undersampling
    $TgtNr \leftarrow |D_N| \times u$
  else
    $TgtNr \leftarrow |D_N| \times \frac{|D_R|}{|D_N|}$
  end if
  $selNormCases \leftarrow$ sample($tgtNr, D_N$)
  $newData \leftarrow newData \bigcup selNormCases$
  if o **then** // Generate synthetic examples
    $TgtNr \leftarrow |D_R| \times o$
  else
    $tgtNr \leftarrow |D_R| \times \frac{|D_N|}{|D_R|}$
  end if
  $synthCases \leftarrow$ genSynthCases($D_R, tgtNr - |D_R|, k$)
  $newData \leftarrow newData \bigcup synthCases \bigcup D_R$
  return $newData$
end function

5.3.2 Resampling with Temporal Bias

In this section, variants of the random undersampling, random oversampling and SMOTEr methods are proposed. The main difference concerning the original proposals relate to the case selection procedures. Instead of randomly selecting cases, a biased procedure is applied. The idea implemented in these proposals is that, the older the example is, the lower is the probability of it being selected for the new training set. This provides a modified distribution which is balanced in terms of normal and rare cases, with a probabilistic preference towards the most recent cases. The integration of the temporal bias is performed as follows. Given a data set D,

- obtain the difference between the publication time stamps of cases and the most recent case, $t.dif = \{max(D[pubTime]) - D_i[pubTime], \forall i \in (1, \cdots, |D|)\}$;
- assign the preference of $p_i = 1 - \frac{t.dif_i}{max(t.dif)}$ for selecting each case in D, where $i \in (1, \cdots, |D|)$;
- select a sample from D where each case has a probability p_i of being selected.

The Undersampling with Temporal Bias (**U_T**) proposal is based on Algorithm 1. The main difference regarding the previously formalized random undersampling strategy, concerns the temporal bias implemented in the case selection procedure. This corresponds to substituting line 15 in Algorithm 1 by the lines 15 through 17 presented in Algorithm 5.

Algorithm 5 The Undersampling with Temporal Bias algorithm (**U_T**).

```
1: function UnderT(D, Y, φ(Y), t_R, u)
2:     // D - A data set
3:     // Y - The target variable
4:     // φ(Y) - User specified relevance function
5:     // t_R - The threshold for relevance on y values
6:     // u - (optional parameter) Percentage of undersampling
7: ···
15:    t.dif ← {max(D[pubTime]) − D_{N,i}[pubTime], ∀i ∈ (1, ··· , |D_N|)}  // Time difference w.r.t. most recent case
16:    prefs ← {1 − t.dif_i / max(t.dif), ∀i ∈ (1, ··· , |D_N|)}  // Higher preferences for most recent cases
17:    selNormCases ← sample(tgtNr, D_N, prefs)  // Sample normal cases from D_N based on prefs
       ···
20: end function
```

The second proposed strategy, oversampling with temporal bias (**O_T**), is based on Algorithm 2. This strategy performs oversampling giving a higher preference to the most recent examples. As such, the strategy incorporates a bias towards newer cases in the replicas selected for inclusion in the novel data set. The integration of the temporal bias in oversampling corresponds to replacing line 15 in Algorithm 2 (random oversampling) by the lines 15 through 17, presented in Algorithm 6.

Algorithm 6 Oversampling with Temporal Bias (**O_T**).

1: **function** OverT($D, Y, \phi(Y), t_R, o$)
2: // D - A data set
3: // Y - The target variable
4: // $\phi(Y)$ - User specified relevance function
5: // t_R - The threshold for relevance on y values
6: // o - (optional parameter) Percentage of oversampling
7: ...
15: $t.dif \leftarrow \{max(D[pubTime]) - D_{R,i}[pubTime], \forall i \in (1, \cdots, |D_R|)\}$ // Time difference w.r.t. most recent case
16: $prefs \leftarrow 1 - \frac{t.dif_i}{max(t.dif)}, \forall i \in (1, \cdots, |D_N|)\}$ // Higher preferences for most recent cases
17: $selNormCases \leftarrow$ SAMPLE($tgtNr, D_R, prefs$) // Sample rare cases from D_R based on $prefs$
...
20: **end function**

The third proposed strategy is SMOTEr with Temporal Bias (**SM_T**). This approach combines the application of undersampling with temporal bias in normal cases, and the application of an oversampling mechanism integrating temporal preferences. The undersampling with temporal bias strategy is the same as described in Algorithm 5. Regarding the oversampling strategy, a preference for the most recent cases is included in the SMOTEr generation of synthetic examples. This means that when generating a new synthetic case, after evaluating the k-nearest neighbours of the seed example, the neighbour selected for the interpolation process is the most recent case. Algorithm 7 shows the changes applied to the original SMOTEr method, described in Algorithm 4. To include the temporal bias, line 17 in Algorithm 4 referring to the undersampling step, is replaced by lines 17, through 19 in Algorithm 7. Concerning the oversampling step, line 24 in Algorithm 4 is replaced by line 27 in Algorithm 7.

Regarding the function for generating synthetic examples, Algorithm 8 describes the changes in Algorithm 3 for including the temporal bias. In this case, only line 13 of Algorithm 3 is altered, in order to consider the time factor, ensuring that the nearest neighbour selected is the most recent.

Algorithm 7 SMOTEr with Temporal Bias (**SM_T**).

1: **function** SMOTErT($D, Y, \phi(Y), t_R, k, o, u$)
2: // D - A data set
3: // Y - The target variable
4: // $\phi(Y)$ - User specified relevance function
5: // t_R - The threshold for relevance on y values
6: // k - The number of neighbours used in case generation
7: // o, u - (optional parameters) Percentages of over and undersampling
8: ...
17: $t.dif \leftarrow \{max(D[pubTime]) - D_{N,i}[pubTime], \forall i \in (1, \cdots, |D_N|)\}$
18: $prefs \leftarrow \{1 - \frac{t.dif_i}{max(t.dif)}, \forall i \in (1, \cdots, |D_N|)\}$
19: $selNormCases \leftarrow$ SAMPLE($tgtNr, D_N, prefs$)
...
27: $synthCases \leftarrow$ GENSYNTHCASEST($D_R, tgtNr - |D_R|, k$)
...
29: **end function**

Algorithm 8 Generating synthetic cases with temporal bias.

```
1: function genSynthCasesT(D, ng, k)
2:     // D - A data set
3:     // ng - Number of synthetic cases to generate for each existing case
4:     // k - The number of neighbours used in case generation
       ...
13:    x ← choose the nns most recent in time
       ...
26: end function
```

5.3.3 Resampling with Temporal and Relevance Bias

This section describes the final proposals of resampling strategies for imbalanced domain learning tasks. The idea of the approaches described in this section is to also include the relevance scores in the sampling bias, in addition to the temporal bias. The motivation is that while it is assumed that the most recent cases are relevant, as they better entail recent dynamics of popularity, it is questionable if older cases with considerable scores of relevance should be discarded. To combine the temporal and relevance bias three new methods are proposed: undersampling (Algorithm 9), oversampling (Algorithm 10) and SMOTEr with temporal and relevance bias (Algorithm 11). The integration of temporal and relevance bias is performed as follows. Given a data set D and a relevance function $\phi()$,

- obtain the difference between publication time stamps of cases, w.r.t. the most recent, $t.dif = \{max(D[pubTime]) - D_i[pubTime], \forall i \in (1, \cdots, |D|)\}$;
- assign the preference of $p_i = (1 - \frac{t.dif_i}{max(t.dif)}) \times \phi(D_i[y])$ for selecting each case in D, where $i \in (1, \cdots, |D|)$;
- select a sample from D where each case has a probability p_i of being selected.

The Undersampling with Temporal and Relevance Bias (**U_TPhi**) proposal is based on Algorithm 1. The main difference regarding the original formalization of the random undersampling strategy, concerns the temporal and relevance bias implemented in the case selection procedure. This corresponds to replacing the line 15 in Algorithm 1 by lines 15 through 17 in Algorithm 9.

The second proposed variant, oversampling with temporal and relevance bias (**O_TPhi**), is based on Algorithm 2. This strategy performs oversampling giving a higher preference to the most recent and most relevant cases. This corresponds to replacing line 15 in Algorithm 2 by lines 15 through 17 in Algorithm 10.

The same integration of time and relevance bias is also done in the SMOTEr algorithm. In this case, both the undersampling and oversampling steps of SMOTEr algorithm are altered. Algorithm 11 (**SM_TPhi**) shows the changes applied to Algorithm 4. Lines 17 and 24 of Algorithm 4 are replaced by lines 17 through 19, and by line 27 in Algorithm 11, respectively.

Algorithm 9 Undersampling with Temporal and Relevance Bias (**U_TPhi**).

1: **function** UNDERTPHI($D, Y, \phi(Y), t_R, u$)
2: // D - A data set
3: // Y - The target variable
4: // $\phi(Y)$ - User specified relevance function
5: // t_R - The threshold for relevance on y values
6: // u - (optional parameter) Percentage of undersampling
7: ...
15: $t.dif \leftarrow \{max(D[pubTime]) - D_{N,i}[pubTime], \forall i \in (1, \cdots, |D_N|)\}$ // Time difference w.r.t. most recent case
16: $prefs \leftarrow \{(1 - \frac{t.dif_i}{max(t.dif)}) \times \phi(D_{N,i}[y], \forall i \in (1, \cdots, |D_N|)\}$ // Preferences based on time and relevance
17: $selNormCases \leftarrow$ SAMPLE($tgtNr, D_N, prefs$) // Sample normal cases from D_N based on $prefs$
...
20: **end function**

Algorithm 10 Oversampling with Temporal and Relevance Bias (**O_TPhi**).

1: **function** OVERTPHI($D, Y, \phi(Y), t_R, o$)
2: // D - A data set
3: // Y - The target variable
4: // $\phi(Y)$ - User specified relevance function
5: // t_R - The threshold for relevance on y values
6: // o - (optional parameter) Percentage of oversampling
7: ...
15: $t.dif \leftarrow \{max(D[pubTime]) - D_{R,i}[pubTime], \forall i \in (1, \cdots, |D_R|)\}$ // Time difference w.r.t. most recent case
16: $prefs \leftarrow \{(1 - \frac{t.dif_i}{max(t.dif)}) \times \phi(D_{R,i}[y], \forall i \in (1, \cdots, |D_R|)\}$ // Preferences based on time and relevance
17: $selRareCases \leftarrow$ SAMPLE($tgtNr, D_R, prefs$) // Sample rare cases from D_R based on $prefs$
...
20: **end function**

These changes correspond to biasing the undersampling process to consider time and relevance of the cases considered as normal, as previously described: the most recent examples with higher relevance are preferred to others, for remaining in the new data set. Regarding the oversampling strategy, the generation of synthetic examples also assumes this tendency: new examples are generated using the function GENSYNTHCASESTPHI(), by prioritising the selection of highly relevant and recent cases. Algorithm 12 shows the changes made in Algorithm 3 (line 13 in Algorithm 3 is replaced by lines 13 through 16). The bias towards more recent and high relevance cases is achieved in the selection of the most recent and relevant nearest neighbour.

In summary, for each of the three resampling strategies considered (random undersampling, random oversampling and SMOTEr), two new variants are proposed, attempting to incorporate a form of context bias in order to improve predictive accuracy on imbalanced domain learning tasks. The baseline strategies (**U_B**, **O_B** and **SM_B**) carry out sampling in the data sub sets of cases considered as normal and those considered as highly relevant (i.e. rare), according to user preferences. The selection of cases when using such strategies is commonly carried out in a random manner. The first variants (**U_T**, **O_T** and **SM_T**) extend the baseline strategies by adding a preference toward the most recent cases within each subset, as these could provide a better depiction of the recent dynamics of popularity. Finally, the second variants (**U_TPhi**, **O_TPhi** and **SM_TPhi**) add another preference to

Algorithm 11 SMOTEr with Temporal and Relevance Bias (**SM_TPhi**).

function SMOTERTPHI($D, Y, \phi(Y), t_R, k, o, u$)
    // D - A data set
    // Y - The target variable
    // $\phi(Y)$ - User specified relevance function
    // t_R - The threshold for relevance on y values
    // k - The number of neighbours used in case generation
    // o, u - (optional parameters) Percentages of over and undersampling
...
    $t.dif \leftarrow \{max(D[pubTime]) - D_{N,i}[pubTime], \forall i \in (1, \cdots, |D_N|)\}$ // Time difference w.r.t. most recent case
    $prefs \leftarrow \{(1 - \frac{t.dif_i}{max(t.dif)}) \times \phi(D_{N,i}[y], \forall i \in (1, \cdots, |D_N|)\}$ // Preferences based on time and relevance
    $selNormCases \leftarrow$ SAMPLE($tgtNr, D_N, prefs$) // Sample normal cases from D_N based on $prefs$
    ...
    $synthCases \leftarrow$ GENSYNTHCASESTPHI($D_R, tgtNr - |D_R|, k$)
    ...
end function

Algorithm 12 Generating synthetic cases with temporal and relevance bias.

function GENSYNTHCASESTPHI($D, ng, k, \phi(Y)$)
    // D - A data set
    // ng - Number of synthetic cases to generate for each existing case
    // k - The number of neighbours used in case generation
    // $\phi(Y)$ - User specified relevance function
    ...
    $y.rel \leftarrow \phi(nns[y])$ // Relevance the target value of nns
    $t.dif \leftarrow \{max(D[pubTime]) - D_i[pubTime], \forall i \in (1, \cdots, |D|)\}$ // Time difference w.r.t. most recent case
    $y.time \leftarrow \{(1 - \frac{t.dif_i}{max(t.dif)}), \forall i \in (1, \cdots, |D|)\}$
    $x \leftarrow \underset{neig \in nns}{\text{argmax}} \; y.rel(neig) \times y.time(neig)$
    ...
end function

the sampling procedures, by also including the relevance scores of the cases, avoiding the disposal of cases that may not be the most recent, but are highly relevant for the user.

In order to understand the impact of applying each of the resampling strategies described in this section, a depiction of the effect on the distribution of data from imbalanced domains is illustrated in Figure 5.1. This data reports to the single-source data set of online news feeds, previously presented in Section 3.2.1, concerning all news published in the month of June (2015) in the topic "economy". To this data sample, the baseline resampling strategies (random undersampling, random oversampling and SMOTEr) and the proposed context biased variants of such strategies are applied. Regarding parametrization, the percentage of undersampling and oversampling are respectively defined as 0.5 and 5; relevance functions are obtained using the boxplot approach proposed by Ribeiro [196] (described in Section 4.3.1) with a relevance threshold of 0.9; and finally, concerning the number of k-nearest neighbours required for the application of the SMOTEr algorithm and its variants, this is defined as 3. This illustration is restricted to a target value of 200 for understandability purposes.

Results show that the impact of each of the resampling strategies is different, corresponding to their respective objectives. The random undersampling strategy clearly reduces cases considered as "normal". Its variant of temporal bias presents similar results. However, the

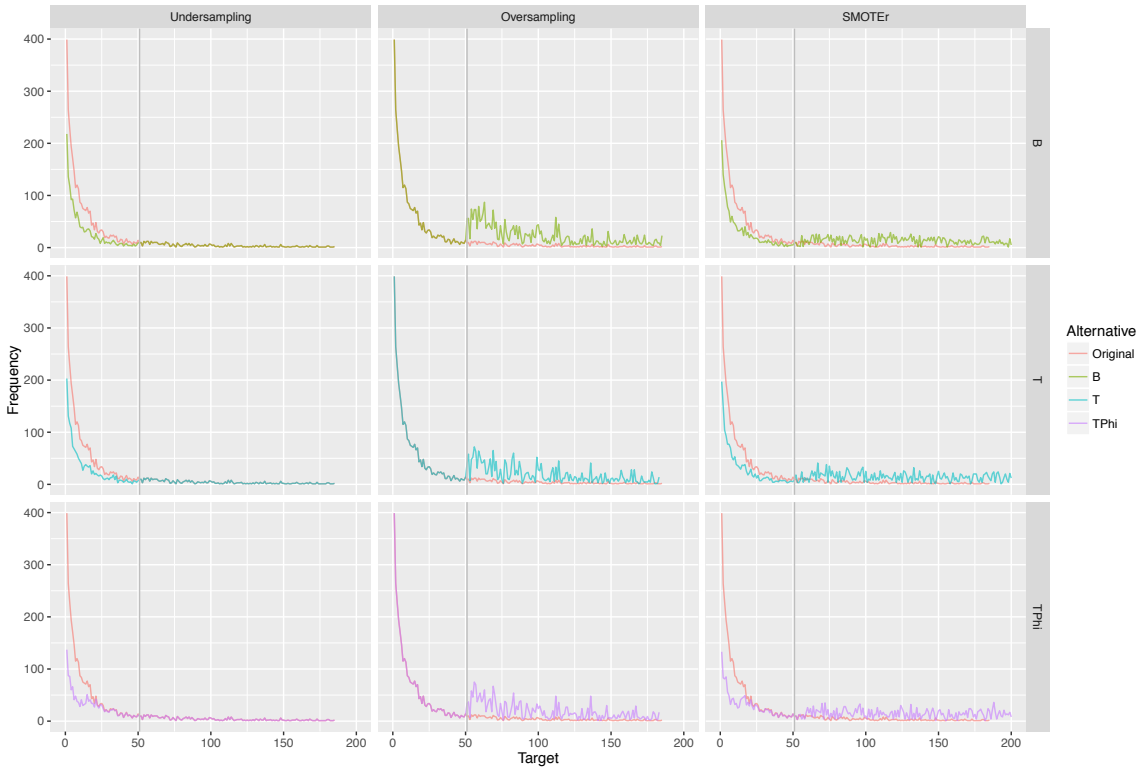


Figure 5.1: Distribution of the target variable of a data sample when resampling strategies are applied, in comparison to the original data (red). The grey line delimits the target variable as to "normal" or "rare" values, given a relevance threshold of 0.9.

variant of temporal and relevance bias shows a more severe reduction of near-zero target values, maintaining examples with a bigger relevance score (although still considered as "normal" cases). Concerning the random oversampling strategy, overall results do not show any consistent difference between the baseline proposal and both variants. It should be noted that this is not unexpected, as *i)* such cases have similar relevance (above the threshold of 0.9 and below 1, the maximum), and *ii)* the impact of the proposed variants are mostly related to the distribution of cases on a temporal dimension. Concerning the SMOTEr algorithm, the results of the baseline approach and its variants confirm the same conclusions w.r.t. to the application of undersampling and oversampling.

The impact of the context-bias variants also relates to the distribution of examples in a temporal dimension. To analyze the impact of applying such strategies in such dimension, the same data sample is used, as well as the previously described parametrization. In Figure 5.2 the impact of applying the described resampling strategies concerning the distribution of examples per day is illustrated. It should be noted that in this example day 1 concerns the less recent items, and day 30 concerns the most recent items.

As expected, by analysing the impact of the resampling strategies, and specifically the

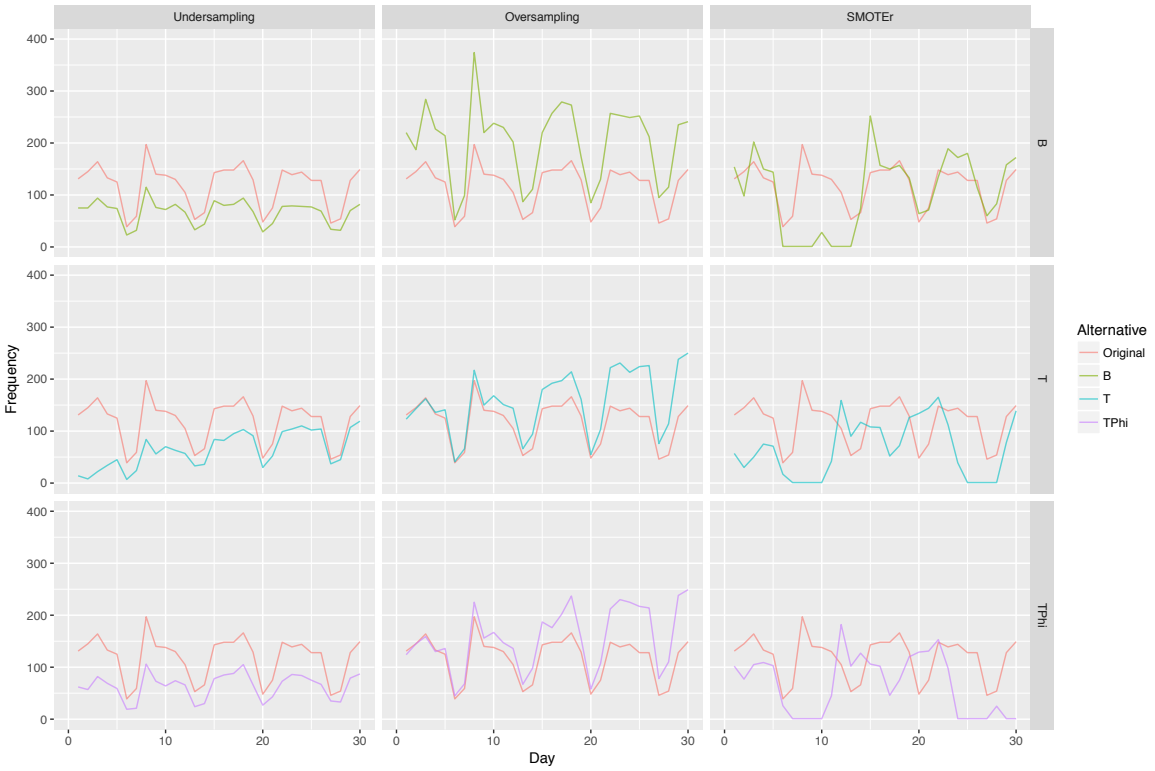


Figure 5.2: Distribution of the target variable of a data sample when resampling strategies are applied, concerning the number of cases per day. The original data is denoted in red.

proposed context-biased variants, it is observed that they favour more recent cases to the detriment of older ones. This impact is observed in both the temporal-bias variants and the temporal and relevance bias variants. Nonetheless, such impact is more evident concerning the temporal-bias variants.

5.4 Algorithm-Level Approaches

In this section, the focus is directed towards the problem of predicting the popularity of web content items after they are published, i.e. *a posteriori* prediction. In such tasks it is commonly assumed that one is in possession of a set of cases C describing the dynamics of their popularity evolution, and that the objective is to predict the final popularity values of a second set of cases depicting their popularity evolution until a given time slice t, P^t. In this thesis, time slices are defined as consecutive periods of 20 minutes, after the publication of a given web content item. The sets C and P are defined as follows, where t_f corresponds to the final time slice, i.e. prediction horizon.

$$C = \begin{bmatrix} c_1^1 & \cdots & c_1^{t_f} \\ \cdots & \cdots & \cdots \\ c_i^1 & \cdots & c_i^{t_f} \end{bmatrix} \quad (5.1) \qquad P = \begin{bmatrix} p_1^1 & \cdots & p_1^t \\ \cdots & \cdots & \cdots \\ p_j^1 & \cdots & p_j^t \end{bmatrix}, t < t_f \quad (5.2)$$

The study of web content popularity prediction after the items are published is the task that collected the most attention within this topic of research. Proposals to solve this task range from statistical approaches, to applying learning algorithms. Some of the most known solutions concern the proposals by Szabo and Huberman [212] and Pinto et al. [191], which are often used as experimental baselines. These are well representative of the types of approaches proposed over the years.

Szabo and Huberman [212] propose two statistics-based approaches, the constant scaling and the log-linear approaches. The constant scaling approach is based on the calculation of a factor α which is solely dependent on the reference time, i.e. time slice. It depicts a concept similar to a growth factor, which is multiplied by the popularity of all the items chosen for prediction, at a given prediction time. The log-linear approach explores the linear relationship of popularity values at a given time slice and their final value, using logarithmic transformations. Other authors have proposed modelling the dynamics of popularity according to a given type of distribution, such as the PCI [118] or the Poisson distributions [206, 87].

Concerning the application of learning algorithms, Pinto et al. [212] propose two approaches which attempt to take advantage of the dynamics of items' popularity, and as such are not solely based on the most recent account of popularity. The first proposal by the authors is a multivariate linear regression model, using sampling intervals (time slices) and denoting each interval as a popularity *delta*, i.e. the difference in popularity between consecutive intervals. A second proposal extends the multivariate linear model by accounting for the similarity of cases, using features obtained by the application of Radial Basis Functions [39]. Tatar et al. [217] and Asur and Huberman [18] also rely on linear models to learn the dynamics of popularity. The former proposes a direct approach, where the model represents the relation between the popularity of items in a training set w.r.t. a given time slice, and their final popularity. The latter extends this direct approach by including features related to sentiment analysis.

As previously described, one of the issues concerning standard prediction tools such as those described in the former paragraphs, is their focus on the average behaviour of the data. Given our predictive focus towards rare cases of highly popular items and the objective of obtaining accurate predictions early on, such approaches raise two caveats: *i)* in the first moments after an item is published its popularity may not be clearly distinguishable from known cases, and *ii)* using the dynamics of all known items in a training set will cause the predictions to be biased towards normal cases, due to the imbalanced distribution. Therefore, the standard statistical and learning methods commonly applied to solve *a posteriori* tasks, could lead to

under-performing prediction models concerning the objective of early and accurate forecast of highly popular web content.

The focus of algorithm-level methods is to address the shortcomings of standard approaches when these are biased towards covering the average behaviour depicted by the majority of cases in the data. This is commonly carried out by altering learning algorithms in order to correct such bias, and to focus on the intended target cases.

To improve the predictive accuracy of highly popular web content items in *a posteriori* tasks, two proposals are presented in this chapter. These proposals are built on the concept of algorithm-level methods, using kernel regression and k-nearest neighbour methods. The distinguishing characteristic of the proposed approaches concern the use of a biased case selection procedure. As such, instead of basing the prediction of cases on overall statistics (e.g. the average slope between the popularity of training cases at a given time slice and their final value), the objective of the proposed approaches is to implement such process locally.

5.4.1 Kernel-Based Approach

Kernel regression is a statistical-based approach for a non-parametric estimation of the conditional expectation regarding a given variable. A kernel is a non-negative weighting function based on the density of random variables. Kernels commonly include an ad-hoc definition of the bandwidth, a smoothing parameter. Seminal work on kernel regression by Nadaraya [173] and Watson [233], formalize the problem as follows:

$$\hat{h}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)}, \tag{5.3}$$

where $\hat{h}$ is the approximation (prediction) for a given set of predictors, x, with a corresponding target variable y; K_h denotes a kernel with a bandwidth h. According to this formulation, kernel regression corresponds to the estimation of a target variable as a locally weighted average, where weights are obtained by a kernel. The inclusion of a kernel allows to weigh the contribution of known examples x_i where $i \in (1, \cdots, n)$, according to their distance w.r.t. to the item being predicted.

The concept of kernel regression provides an interesting framework concerning the problem dealt in this thesis. As mentioned, one of the major issues of learning tasks in imbalanced domains is the focus of standard learning algorithms on the average behaviour of the train data. Kernel regression provides an approach to modelling and prediction tasks using locally weighted estimation. It should be noted that the definition of a kernel's bandwidth variable is crucial for the efficiency of kernel regression. Regarding such issue, previous work has

described the difficulties in defining this variable in heavy-tail distributions [40], such as that of the web content popularity domain.

Description

The first proposal in this thesis concerning algorithm-based methods is a kernel-based approach, building on the concept of kernel regression. The idea of this proposal is related to capturing the local dynamics of items' popularity, by using locally weighted averages instead of coarse-grain statistics concerning the data. This relates to the procedure of case selection, where distinct training cases should provide a differentiated contribution to the prediction of future values, depending on their distance w.r.t. to the target case.

To achieve such outcome, a distance factor is introduced in the training case selection process. This factor enables the determination of an interval around the popularity value of a given web content item. This allows the selection of cases that have similar levels of popularity w.r.t. the target prediction case, at a given time slice t. Concerning the definition of the interval, this proposal uses the interquartile range (IQR), considered to be a basic robust measure of scale. It is defined as $IQR = Q_3 - Q_1$, where Q_3 and Q_1 report to the third and first quantile of a given continuous target variable.

As depicted in Equations 5.1 and 5.2, the popularity of a given item may vary in consecutive time slices, and as such the distribution of the the target variable as well. Given this, the value of IQR is calculated for each time slice t and is denoted as IQR_t, where $t \in (1, \cdots, t_f)$.

Formalization

Given a case p_j for prediction at a given time slice t, the proposed kernel-based approach formulates the predictive problem as follow.

$$\hat{p}_j^{t_f} = f(p_j^t, C^t), \tag{5.4}$$

where $\hat{p}_j^{t_f}$ is the predicted value of popularity for the item p_j in the final time slice t_f, p_j^t is the level of popularity at the reference time (time of prediction), and C^t represent the popularity values of cases from a given training set C in time slice t.

By using the value of the target case at the reference time (p_j^t), a procedure is applied in order to obtain a set of cases that are within the interval of IQR_t, i.e. the maximum admissible distance for considering an example as similar. This is carried out by defining the lower and higher value thresholds of p_j^t concerning IQR_t, and retrieving the index of the items in C^t that present a value framed within the mentioned thresholds:

$$low_j^t = max(0, p_j^t - IQR_t) \tag{5.5}$$

$$high_j^t = p_j^t + IQR_t \tag{5.6}$$

$$A_j^t = \{i : c_i^t \in [low_j^t, high_j^t], c \in C^t\} \tag{5.7}$$

Given a set of indexes A_j^t representing the cases considered as similar to the target case p_j^t, it is necessary to calculate the weight that each of these train cases have when predicting the final popularity value of the target case p_j.

In kernel regression approaches, the common choices of kernels concern polynomial functions and Gaussian radial basis functions (described by Wu and Chang [243]). Such approaches attribute a weight to all known cases based on a given notion of similarity: the most similar obtain a value close to 1, and the most distant a value which tends to 0.

The motivation for the kernel-based approach is that the number of cases that are considered as the basis for the prediction should be restricted. As such, the weight of cases is defined as the inverse distance between the popularity at the time slice t of each case c_a^t where $a \in A_j^t$, and the popularity value of the target case p_j^t. These values are normalized into a $[0, 1]$ scale. The calculation of the weights is formalized in the following equation.

$$W_j^t = \{1 - \frac{|p_j^t - c_a^t|}{c_a^t}, \forall a \in A_j^t, c \in C^t\} \tag{5.8}$$

Using the train cases considered as being similar (in terms of popularity) to the target case, and their calculated weights, the prediction of the popularity value at the final time slice t_f for a given case p_j^t is carried out as follows:

$$\hat{p}_j^{t_f} = \frac{\sum_a w_a \times c_a^{t_f}}{\sum_a w_a}, a \in A, c \in C^t \tag{5.9}$$

5.4.2 kNN-Based Approach

The second proposal in this thesis concerning algorithm-based methods is based on the k-nearest neighbour algorithm [13] (kNN). This algorithm is a non-parametric method, considered to be one of the most simple algorithms in the field of machine learning. Concerning regression tasks, a typical setting of the method operates by deriving a subset of k train cases, those presenting the smallest distance to the target case. Using this subset, the predictions are given by the average of their target values.

Description

In comparison to the previously described kernel regression approach, the main difference between these two methods is related to the use of weights. While kernel regression is built on locally weighted averages, the kNN algorithm does not use weights. Therefore, it is based on predictions given by local averages, w.r.t. to the k-nearest neighbours of a given target case. In comparison to the original kNN algorithm, instead of providing a fixed number of neighbours k, in the proposed kNN-based approach this value is given the amount of cases in a training set that, in a time slice t, have a similar popularity value w.r.t. the target case, i.e. IQR_t.

Formalization

The formalization of this second proposal concerning algorithm-based methods, the kNN-based approach, is very similar to the formalization of the kernel-based approach. The main difference between these two approaches relates to the non-use of weights (Equation 5.8). Disregarding the influence of such factors, the formalization of the kNN-based approach is given by the following, concerning a given target case p_j and the prediction of its final popularity value w.r.t. to the reference time slice t:

$$\hat{p}_j^{t_f} = \frac{\sum_a c_a^{t_f}}{|A|}, a \in A_j^t, c \in C^T \tag{5.10}$$

where A_j^t (Equation 5.7) is an index set regarding cases in C^t (set of values at time slice t from the train set) with a popularity value at the reference time slice t framed within an interval of IQR_t (Equations 5.5 and 5.6).

5.5 Hybrid Methods

Hybrid methods consist of a strategy for tackling learning tasks with imbalanced domains where data-level and algorithm-level methods are combined. By exploring their main advantages, hybrid methods are considered to a robust and efficient solution in imbalanced domain learning tasks [131]. Related work shows that such methods are commonly formulated in two manners: *i)* through the proposal of algorithm-based methods which incorporate data-level methods, or *ii)* by combining both types of methods through ensemble approaches. However, after a careful review of previous work concerning the subject of imbalanced domain learning and the proposals for hybrid methods, no evidence was found regarding proposals of such methods in the context of regression tasks.

Motivation

Several shortcomings may be raised by the formalization of predictive tasks as either *a priori* or *a posteriori* prediction. Although predictive modelling approaches in *a priori* tasks are useful for predicting web content popularity when no social feedback is available, they are not designed to correct or update such predictions once observations of popularity become available. As for approaches in *a posteriori* scenarios, relying solely on a limited amount of data available shortly after publication might not be sufficient for an accurate prediction of highly popular items.

In this section, building on the concept of hybrid methods, two novel proposals for hybrid methods are introduced, concerning the accurate prediction of highly popular web content. The proposals consist of time-based ensembles, in order to combine the strengths of data- and algorithm-level methods, and as such, the predictive ability presented by approaches for both *a priori* and *a posteriori* prediction tasks. Despite using proposals concerning the task of *a priori* prediction, the proposed hybrid methods are targeted towards *a posteriori* prediction tasks.

5.5.1 Time-Based Ensembles

Ensemble methods train several learners to tackle the same problem and combine their outcome [258]. The most common combination methods are averaging and voting. The proposed hybrid methods focus on averaging methods which are more appropriate for regression tasks. A simple version of these methods consist of averaging the output of the learners directly. One may also use the weighted averaging method, where the combined output is obtained by averaging the outputs of each learner with different weights. This approach is common when the objective is to attribute different levels of importance for each learner. Given that the predictive ability of web content popularity models in *a posteriori* tasks is related to the level of the available data at a given time t, the proposed time-based ensembles resort to such weighted averaging methods.

The scarcity of social feedback is a major issue in both *a priori* and *a posteriori* prediction tasks. Such shortcoming is related to the recency of the events. As such, the alive-time (denoted by a time slice t) of web content items is a crucial factor in order to combine models designed for both predictive scenarios. In Figure 5.3, the evolution of the mean proportion of available social feedback in online news feeds data from both the single- and multi-source data sets (described in Chapter 3) is illustrated. These samples concern the topic "obama". It should be reminded that in the mentioned data sets the evolution of popularity is observed for a period of two days and the final time slice is 144. The blue (dashed) line shows the evolution of the mean proportion of available data for the rare cases of highly popular

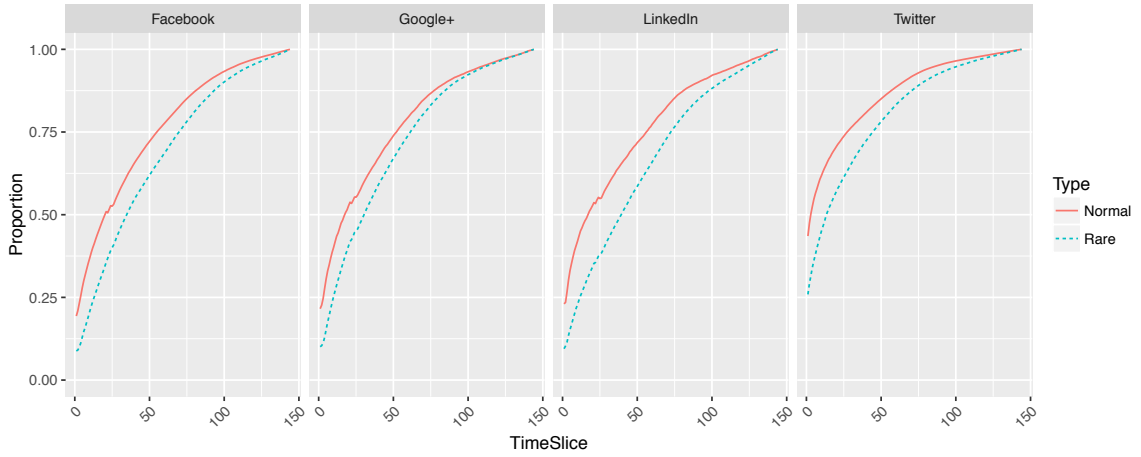


Figure 5.3: Example of the evolution in mean proportion of data concerning the topic "obama" in both the single- and multi-source data sets described in Chapter 3. The dashed line represents the evolution of the cases considered as highly popular.

news, according to the application of the boxplot approach for the automatic generation of relevance function, proposed by Ribeiro [196]. The relevance threshold is defined as 0.9. As expected, Figure 5.3 shows that, in comparison to cases considered as normal, the cases of highly popular items have a much slower evolution of popularity.

Given this, the proposed hybrid methods of time-based ensembles are based on the following assumptions:

1. When social feedback is unavailable, only *a priori* models are able to predict the popularity of web content;
2. When web content items are recent, the available social feedback may be insufficient to confirm *a priori* predictions or to accurately predict popularity using approaches focusing on *a posteriori* prediction;
3. As time passes since the publication of the web content items, the available social feedback increases the accuracy of *a posteriori* predictions.

Given that the ability of prediction models in accurately predicting highly relevant cases is related to the available data at the time of prediction, the proposed time-based ensembles relate the weights of each learner in an ensemble w.r.t. the evolution of the average proportion of available data at a given time t. The average proportion of available data concerning the final popularity values of cases is learned using the available training data. This characteristic of the proposed hybrid methods justifies its time-based property: weights attributed to models of each scenario (*a priori* and *a posteriori*) are dependent of the reference time slice t, in which the prediction occurs.

Proposals

Two time-based ensemble approaches are proposed, resorting to the weighted averaging method: *i)* by combining the numeric predictions of the models ($ENSt$), and *ii)* by combining the relevance of the models' predictions ($ENSphi$). As a reminder, the concept of relevance, as proposed by Ribeiro [196] and described in Section 4.3.1, is related to the rareness of the popularity of the web content items: higher the popularity, higher the relevance of the web content item.

When web content items are published, there is no related social feedback available. Therefore, the predicted popularity of items when $t = 0$ is solely based on the *a priori* models' predictions (see Assumption 1). As such, the weight of *a priori* models in the combination method of the ensemble is $w^0_{pr} = 1$, and the weight of the *a posteriori* models is $w^0_{po} = 0$. This is true for both proposed approaches of time-based ensembles ($ENSt$ and $ENSphi$) .

When $t \in (1, 2, \ldots, t_f)$, where t_f is the final time slice, the weight of models focusing on *a posteriori* tasks are associated to the average proportion of available data, k_t at a given time slice t, learned with data from a training set C,

$$k_t = \frac{\sum_{i=1}^{|C|} \frac{c_i^t}{c_i^{t_f}}}{|C|}, \tag{5.11}$$

where c_i^t is the popularity in the time slice t of each case in a training set C. Therefore, the weight of the *a posteriori* models is defined by $w^t_{po} = k_t$. Conversely, the weight of *a priori* models is given by $w^t_{pr} = 1 - k_t$ (see Assumption 2).

Combination of Predicted Values (ENSt)

The first proposed approach (ENSt) applies weighted averaging to the numeric predictions of *a priori* models, $\hat{y}_{pr}$, and of *a posteriori* models, $\hat{y}_{po}$, where weights are calculated as previously described (k_t). Therefore, the formalization of this time-based ensemble proposal is as follows:

$$\hat{y} = w^t_{po} \times \hat{y}_{po} + w^t_{pr} \times \hat{y}_{pr}. \tag{5.12}$$

One of the effects of applying data-level methods in *a priori* prediction tasks, is that a bias is introduced in the original learning data. By providing a more balanced distribution of data, under-represented cases with extreme values of popularity cause the average value of the target variable to increase, also influencing the predicted values. As such, for web content items with a low popularity level, the application of data-level methods may cause

an over-estimation of popularity and an increase of normal cases obtaining predictions with values considered as highly relevant. In classification tasks, such errors are known as false positives.

Combination of Relevance Scores (ENSphi)

To tackle these potential issues, a second approach (ENSphi) is proposed. In this proposal the weights used in time-based ensembles, $w_{po} = k_t$ and $w_{pr} = 1 - w_{po}$, are applied to the relevance of the predicted values of each model ($\phi(\hat{y}_{po})$ and $\phi(\hat{y}_{pr})$ respectively) instead of being applied to the predicted values. Using this combined relevance score, the predicted popularity value is given by the inverse of the previously mentioned relevance function (ϕ^{-1}). Formally, the second proposed approach for time-based ensembles (ENSphi) is described as such:

$$\hat{y} = \phi^{-1}(w_{po}^t \times \phi(\hat{y}_{po}) + w_{pr}^t \times \phi(\hat{y}_{pr})). \tag{5.13}$$

It should be noted that a rule is introduced in both the proposed time-based ensemble proposals: if the observed popularity level of a given web content item, at a given time slice t, is considered to be highly relevant (has a relevance score above the relevance threshold), the entire weight of the time-based ensemble is attributed to *a posteriori* models (see Assumption 3).

5.6 Experimental Evaluations

In this section an experimental evaluation of approaches for web content popularity prediction is presented and discussed. This section includes two sets of experiments using the online news feeds data sets presented in Chapter 3. The first is focused on the evaluation of prediction models in *a priori* prediction tasks. The second relates to the predictions models proposed for *a posteriori* prediction tasks.

In both of these predictive tasks, the problem of web content popularity prediction is formalized as an imbalanced domain learning task. Concerning the prediction horizon, the objective is to predict the final value of each news' popularity two days after their publication. These values may be different, depending on the social media source used to check the items' popularity. As such, the ground-truth values in both sets of experiments are given by the final popularity of the news items, according to each of the social media sources.

Data

The experiments carried out and analysed in this section are based on data concerning two data sets of online news feeds presented in Chapter 3. In comparison to other types of web content data, online news feeds have a much shorter life-span. This creates added difficulties for predictive tasks, since it requires models to enable accurate predictions shortly after the items are published.

The main difference between the data sets presented relate to the amount of official and social media sources. The first data set is a single-source collection of news items, collected from the news recommender system Google News, and their respective popularity was obtained by querying the social media source Twitter. As for the second data set, it uses data from multiple official and social media sources: news are collected from Google News and Yahoo! News, and their popularity is obtained from the social media sources Facebook, Google+ and LinkedIn. In both data sets, the data was obtained by querying both types of sources in 20 minutes intervals: from the official media sources the top-100 news of a given news topic is retrieved, and the popularity of all items with an alive time under two days (prediction horizon) is obtained from the social media sources.

Depending on the predictive task (*a priori* or *a posteriori*), the data used in the predictive modelling process is different. The former relies solely on descriptors of the items, excluding observations of the evolution of their popularity in social media sources. As for the latter, it is primarily based on the modelling of the popularity evolution of web content. The news items in both data sets belong to four different news topics: *economy*, *microsoft*, *obama* and *palestine*. As previously referenced, the criteria for the topics selected relates to them being very active topics and to report different types of entities.

Data for *A Priori* Tasks

Concerning the data used in *a priori* prediction tasks, this mainly relates to the information retrieved from official media sources. For each query, information concerning the top-100 recommended news is collected. This includes the title, headline, publication data, news outlet and the position that such news are presented in the respective official media source ranking. Based on the results obtained by the experimental analysis provided in Section 4.7 the predictors used by the predictive models tested in this experimental evaluation report to content and meta-data features: the sentiment scores of both in the titles and headlines of news items using the SentiWordNet 3.0 [21] sentiment lexicon, and the average popularity of the news outlets and of the entities mentioned in both the titles and headlines. The process for obtaining such features is described in Section 4.7.1. It should be reminded that the use of social network and external sources features is not considered. The issues concerning

social network features relate to privacy concerns and the problem of scability in accessing data [91]. The latter relates to the conclusions shown by Martin et al. [161].

Data for *A Posteriori* Tasks

For *a posteriori* prediction tasks, in addition to the data sets based on news descriptors, it is necessary to have data concerning the popularity evolution of the news. The evolution of the news popularity according to each social media source is used to construct additional data sets, which are used in approaches for *a posteriori* prediction tasks. The procedures to obtain such data from each social media source is thoroughly described in Section 3.2.

Table 5.1 illustrates the mentioned data sets, where T_t reports to a given time slice t (periods of 20 minutes), n_i represents the news items obtained from a given official media sources and $y_i^t \in Y$ describes the popularity of news item n_i in time slice t, where t_f represents the final time slice. Considering the timespan of two days, the final number of time slices is 144 ($t_f = 144$).

Table 5.1: Illustration of the data set used in *a posteriori* prediction tasks, encapsulating the evolution of popularity in a social media source.

News	$\mathbf{T_1}$	$\mathbf{T_2}$	$\cdots$	$\mathbf{T_{t_f}}$
n_1	y_1^1	y_1^2	$\cdots$	$y_1^{t_f}$
n_2	y_2^1	y_2^2	$\cdots$	$y_1^{t_f}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n_i	y_i^1	y_i^2	$\cdots$	$y_i^{t_f}$

Evaluation Metrics

The objective of the experimental analysis provided in this chapter is to assess the ability of prediction models in accurately predicting the popularity of highly popular items, concerning both *a priori* and *a posteriori* predictive settings. As thoroughly discussed in Chapter 4, standard evaluation metrics are focused on assessing the average error of predictions, assuming uniform domain preferences by users. This is not the case in this experimental evaluation, given that the predictive task is formalized as an imbalanced domain learning task.

In order to provide an appropriate evaluation of the prediction models employed in both experimental sets, the utility-based evaluation framework proposed in Section 4.6 is used. This framework includes three evaluation metrics: *i)* the root mean squared error $RMSE$, the relevance-weighted root mean squared error $RMSE_\phi$ and a utility-based F-Score F_β^u. Concerning the parametrization of the utility-based evaluation framework, the settings used in the experimental analysis described in Section 4.7 are applied.

Learning Algorithms

In order to build predictive models in *a priori* tasks, it is necessary to use learning algorithms. In order to test the hypothesis that data-level methods are able to improve the predictive accuracy of models concerning the target cases of the online news feeds domain, a diverse set of regression tools is tested. The goal is to ensure that the conclusions provided the experimental evaluation are not biased by a particular learning tool.

Table 5.2 presents the regression tools used in the experimental evaluation concerning *a priori* tasks. To allow for an easy replication, the tools used concern implementations in the free and open source **R** environment.

Table 5.2: Regression algorithms and respective R packages.

ID	Method	R package
LM	Multiple linear regression	stats [195]
MARS	Multivariate adaptive regression splines	earth [168]
RF	Random forests	randomForest [144]
SVM	Support vector machines	e1071 [167]

Concerning the parameter settings, a simple heuristic method is applied in order to discover the optimal parametrization for the models of each learning algorithm: given a parameter space of possible values, the optimal parametrization is the setting that obtains the best possible results, concerning the evaluation metric F_1^u. This choice is related to the fact that the main goal of the prediction tasks evaluated in this section is to accurately predict rare cases of highly popular items.

The following list describes the parameters tested. For SVM models, the *cost* (c) and *gamma* (g) parameters were test; for MARS models, the parameters nk, *degree* (d) and *thresh* (th); and for Random Forest models parameter *ntree* (nt) were also tested. These parameters correspond to the mentioned implementation of such models in **R**.

- **svm**: $c \in \{10, 150, 300\}$, $g \in \{0.01, 0.001\}$;
- **mars**: $nk \in \{10, 17\}$, $d \in \{1, 2\}$, $th \in \{0.01, 0.001\}$;
- **rf**: $nt \in \{500, 750, 1500\}$;

The optimal parametrization method was applied to each of the regression tools used in each combination of regression tool - data set. In addition, all of the variants using resampling strategies were also optimized. Concerning the under-sampling percentages, the following values were tested: 0.1, 0.2, 0.5, 0.6, 0.7, 0.8, 0.9. The over-sampling percentages tested include 1.1, 1.5 and 2. In resampling strategies combining both under-sampling and over-sampling techniques, all combinations of percentages were tested. Results are detailed in Appendix C.

Evaluation Methodology

One of the major concerns in performing experimental evaluations concerns the decision of the experimental methodology employed. These are applied in order to accurately assess the prediction error of the models tested. Given the implicit temporal order of web content data, it is necessary to choose an experimental methodology that guarantees that the original order of the data is maintained, i.e. models are trained on past data and tested on future data.

As such, the Monte Carlo simulation method is applied in both sets of experiments concerning *a priori* and *a posteriori* predictive tasks. This method randomly selects data points in the available data, selecting a certain past window as training data, and a subsequent window for test data. All alternative models use the same training and test sets in order to ensure a pairwise comparison of the estimates obtained.

5.6.1 Evaluation of A Priori Prediction Tasks

This section presents the results of the first set of experiments, concerning *a priori* prediction tasks. The data used in order to perform this experimental set is previously described in Section 5.6.

The objective of this experiment is to evaluate the contribution of data-level methods in terms of the predictive models accuracy, when focusing on under-represented cases (highly popular items). The data-level methods tested include the random undersampling and SMOTEr proposals by Torgo et al. [222], and the random oversampling method. This experiment also tests the proposed context-bias resampling strategies, described in Section 5.3. These include two variants of the random undersampling, random oversampling and SMOTEr methods: *i)* using a temporal bias, and *ii)* using a temporal and relevance bias in the respective case selection procedures.

In order to ensure that conclusions are not biased, several learning algorithms are employed in order to build predictive models. These are detailed in Section 5.6, as well as the method for optimal parametrization of the models w.r.t. each learning algorithm.

Baselines

As baselines for this experiment, in addition to the models produced by the learning algorithms without the application of resampling strategies (data-level methods), the approach proposed by Bandari et al. [24] is also tested.

The authors report that the best results concerning web content popularity prediction are

obtained by resorting to support vector machines (SVM) to model the data, using a different set of predictive features than the one used in the approaches proposed in this thesis. Bandari et al. propose the use of six predictive features: the density score of the each news source and category, a subjectivity score, the number of named entities, and the highest and average scores of popularity among named entities. The source and category scores report to the items' level of popularity. As for the subjectivity score, the authors examine if an article written in a more subjective voice can resonate stronger with the readers. Originally, the authors used a subjectivity classifier from LingPipe[2]. However, due to issues concerning its implementation in **R**, the approach to obtain subjectivity scores by Asur and Huberman [18] is applied. The authors state that the subjectivity of a given text is given by the fraction of sentiment words w.r.t. to the total number of words it contains.

Results

To evaluate the various combinations of learning algorithms and data-level methods (resampling strategies), the previously described utility-based evaluation framework (Section 4.6) is applied. Results are obtained through 20 repetitions of a Monte Carlo simulation process with 50% of the cases used as training set and the subsequent 25% as test set, using the infrastructure provided by R package **performanceEstimation** [220].

Figure 5.4 presents an illustration of the results concerning the combination of news data on each of the topics *economy*, *microsoft*, *obama* and *palestine* and the popularity of such news according to each of the social media sources available: Twitter in the single-source data set, and Facebook, Google+ and LinkedIn in the multi-source data set. Results are grouped by the respective learning algorithm. In each group, the results obtained by the baseline approach of Bandari et al. [24] is included, in order to provide an easier comparison.

This illustration presents the results concerning the evaluation metric F_1^u, which is the most robust metric concerning the objective of the task: the accurate prediction of highly popular items. In Annex D results concerning all metrics are described, where for each combination of social media source and news topic, the best result according to each evaluation metric is denoted in bold.

Results obtained by the standard evaluation metric *rmse* (Annex D) show that the best models are those which do not apply resampling strategies. This outcome is expected, as the imbalanced distribution of web content popularity causes the predictive focus of such models to target the cases considered to be normal. Nonetheless, it should be noted that SVM and Random Forest models using resampling strategies are capable of improving results concerning the *rmse* evaluation metric, in some of the social media source-news topic contexts.

[2]Lingpipe 4.1.0: `http://aliasi.com/lingpipe`

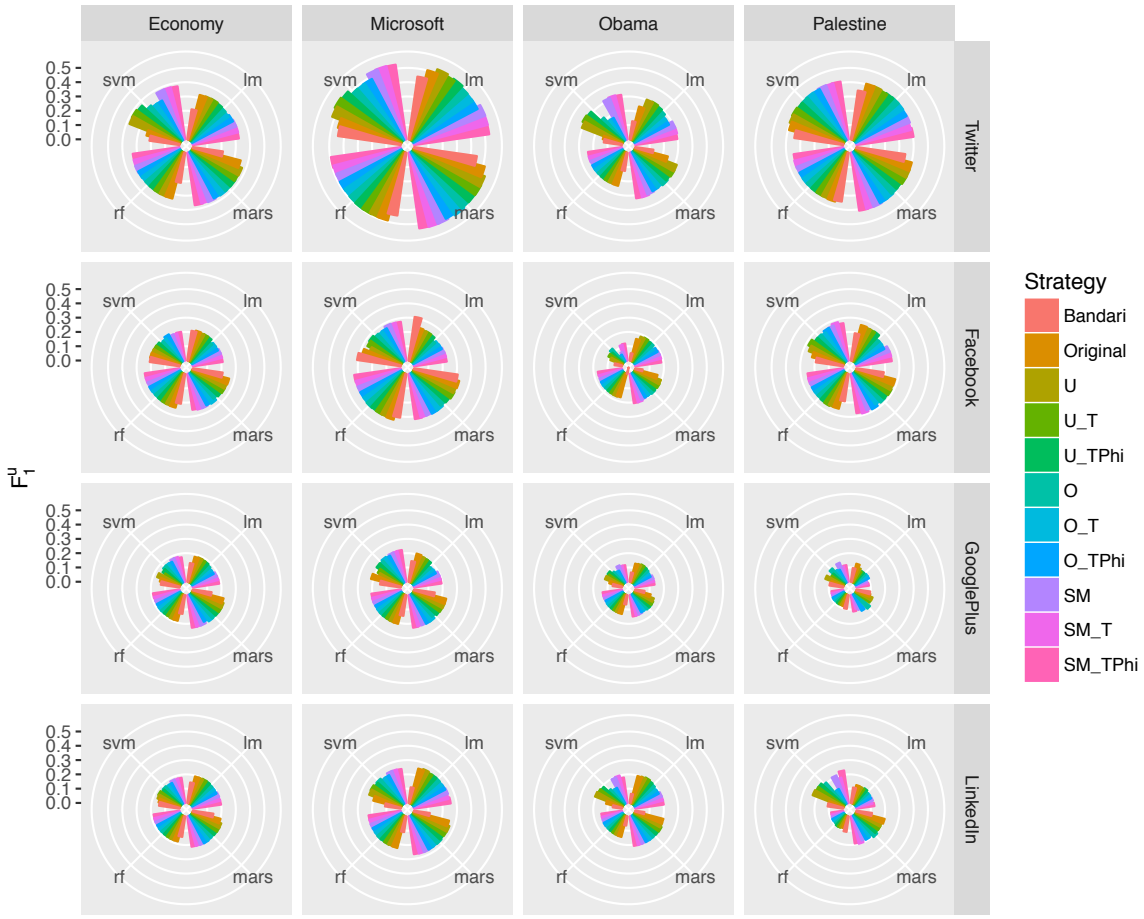


Figure 5.4: Evaluation results of prediction models for *a priori* prediction tasks, concerning the F_1^u metric, for all combinations of the available news topics and the popularity scores given by all social media sources.

By comparing the results of this evaluation metric and the outcome according to the $rmse_\phi$ metric, the best overall results show the impact of accounting for relevance when assessing the numeric prediction error of highly popular news: the best results considering the $rmse_\phi$ metric are given by models which apply resampling strategies. It is noted that concerning this metric, the best overall models are obtained by combining the MARS models and the resampling strategy SMOTE.

Regarding the utility-based evaluation metric F_1^u, which combines the utility-based precision ($prec_\phi^u$) and recall (rec_ϕ^u) metrics, results show that the best results are given by models where

resampling strategies are applied. This confirms the claim set forth in this thesis concerning the issues raised by standard evaluation metrics, in the context of *a priori* prediction tasks. Concerning the regression algorithms employed, results show that the best overall results are obtained when using the regression algorithm MARS. As to the best combination of a regression algorithm and a given resampling strategy, results show that the best overall combination is provided by the use of the MARS algorithm, and the application of the resampling strategy undersampling with temporal bias.

Concerning the baseline approach of Bandari et al. [24], tested in this experimental set, results show that such prediction models do not present an improvement over the approach (feature set of content and meta-data) proposed in this thesis, concerning both the $rmse_{\phi}$ and F_1^u utility-based metrics, without the use of resampling strategies. This is observed in the vast majority of combinations concerning social media source-topic-regression algorithm settings. However, it should be noted that concerning the standard evaluation metric $rmse$, the approach by Bandari et al. shows a consistent advantage over the approach proposed in this thesis, when models are built using the Random Forest learning algorithm.

Given the outcome of this experimental evaluation, it is observed that the models where resampling strategies are applied provide a considerable advantage in web content popularity prediction tasks, when focusing on the accurate prediction of highly popular items. However, the contribution of such resampling strategies in comparison to not using such strategies, and between the non-biased and the proposed context-bias strategies, is still unclear. In terms of comparison with its non-use, and concerning the context-bias resampling strategies proposed in this thesis. To assert the statistical significance of such methods, the Wilcoxon signed rank tests are applied. These are used to test the hypothesis that the performance of the resampling strategies provide significant accuracy improvements over a given baseline, concerning the utility-based evaluation metric F_1^u. The objective is to infer the statistical significance (with p-value < 0.05) of the paired differences in the outcome of each approach.

The statistical tests concerning the results of models with and without the application of resampling strategies show that the former provides a significant performance improvement in every comparison. However, concerning the comparison between the baseline resampling strategies and the context-bias strategies proposed in this thesis, the results are more diverse. These are shown in Table 5.3, by aggregating their outcome by each regression algorithm. Such tests are carried out separately concerning the random undersampling (U), random oversampling (O) and SMOTEr (SM) strategies, which are compared to their respective context-bias variants: *i)* temporal bias ($\_T$), and *ii)* temporal and relevance bias ($\_TPhi$).

Given the outcome of the significance tests carried out, results show that the proposed context-bias resampling strategies are capable of providing significant advantages over the baseline strategies used in this thesis, in the majority of cases. It is observed that, depending on the regression algorithm employed and the baseline resampling strategy used, results vary.

Table 5.3: Number of significant ($p-$value < 0.05) wins/ties/losses according to Wilcoxon signed rank tests, concerning the F_1^u evaluation metric, for models with and without the application of resampling strategies.

	LM		**MARS**		**SVM**		**RF**	
Strategy	Wins (Sig)	Losses (Sig)	Wins (Sig)	Losses (Sig)	Wins (Sig)	Losses (Sig)	Wins (Sig)	Losses (Sig)
U	**4** (**3**)	4 (2)	3 (**2**)	5 (3)	4 (**2**)	4 (2)	2 (2)	6 (0)
U_T	**4** (2)	4 (1)	**6** (**2**)	2 (0)	**6** (1)	2 (1)	**6** (**5**)	2 (0)
U_TPhi	**4** (**3**)	4 (1)	3 (**2**)	5 (3)	2 (**2**)	6 (4)	4 (4)	4 (1)
O	3 (2)	5 (2)	2 (2)	6 (1)	**6** (1)	2 (1)	3 (3)	5 (0)
O_T	**5** (**4**)	3 (0)	**6** (**6**)	2 (0)	3 (1)	5 (3)	3 (3)	5 (0)
O_TPhi	4 (3)	4 (1)	4 (3)	4 (0)	3 (**2**)	5 (4)	**6** (**6**)	2 (0)
SM	**7** (**5**)	1 (1)	**7** (**6**)	1 (0)	**7** (**3**)	1 (0)	2 (2)	6 (0)
SM_T	1 (0)	7 (2)	3 (2)	5 (1)	4 (**3**)	4 (3)	4 (4)	4 (0)
SM_TPhi	4 (2)	4 (2)	2 (2)	6 (1)	1 (1)	7 (2)	**6** (**6**)	2 (0)

Concerning the baseline version of the random undersampling strategy, the proposed variants provide a significant advantage in most of the cases. Regarding the random oversampling strategy, results show that both the temporal bias and the temporal and relevance bias variants of the strategy provide a significant advantage. As for the SMOTEr method, the proposed variants only show a significant advantage when using the Random Forest learning algorithm.

An analysis of the predictive ability of models w.r.t. the their data setting (combination of topic and social media source), shows that such ability is divergent. Results show that the social media source Twitter (single-source data set) is more "predictable" than the remainder of the social media sources. Also, concerning the predictive ability of models concerning the topic of news and the social media source used, these show clear oscillations of evaluation results. This demonstrates the difficulty of obtaining prediction models that are capable of accurately forecasting the popularity of highly popular news items in different settings, providing an explanation for the varying results when resampling strategies are applied.

The outcome of this experimental evaluation set concerning *a priori* prediction tasks shows that the application of resampling strategies significantly improves the ability to accurately predict the popularity of the highly popular items. Concerning both the baseline resampling strategies and the proposed variants, results show that they are able to overcome the bias of learning algorithms towards the average behaviour of the data, to a significant degree. However, despite their significant advantage in comparison to previous work and baseline models, the models evaluated in this experimental set obtain scores of the utility-based evaluation metric F_1^u varying between 0.2 and 0.4. Such outcome is not optimal, and its impact on the ability to provide rankings of news items which are fast and accurate in suggesting highly popular news is still unclear.

In the following section, results of the second experimental evaluation set are described. The experimental set concerns web content popularity prediction tasks in an *a posteriori* setting,

i.e. after the news items are published.

5.6.2 Evaluation of A Posteriori Tasks

Given the differences between both predictive scenarios, the data used in each of them is different: *a priori* tasks do not use behavioural features whilst *a posteriori* tasks are primarily based on modelling such features. Therefore, it should be made clear that the data used in this experimental set is different from the former. Such data is described in Section 5.6. It consists of a data set where the evolution of news items' popularity is described in intervals of 20 minutes.

In *a posteriori* prediction tasks the news items may have different levels of popularity data available at a given moment. As an example, for a news item n_i framed within the time slice t_5 (between 80 and 100 minutes after it is published), the available data consists of consecutive measurements of popularity in time slices $t_1, \cdots, t_5$. Using such data, the objective is to accurately predict the amount of popularity for this given item in its final time slice t_f. Given the prediction horizon of 2 days defined in the online news feeds data used in these experimental evaluations sets, the final time slice is 144.

This experimental set is also formalized as an imbalanced domain learning task. The main objective is to evaluate the proposals previously presented and formalized in Sections 5.4 and 5.5. These concern algorithm-level methods (kernel- and kNN-based methods) and hybrid methods (*ENSt* and *ENSphi* variants of time-based ensembles) used to tackle imbalanced domain learning tasks.

The methodology applied to carry out this set of experiments is described as follows. First, an experimental evaluation is carried out concerning the proposed approaches for algorithm-level methods, including several well-known approaches as baselines. The outcome of such experimental evaluation is presented and analysed. Then, using the approaches that present the best results concerning *a priori* prediction (described in the previous section), and the best performing algorithm-level method, a second evaluation is carried out in order to assess the predictive accuracy of the proposed hybrid methods concerning the prediction of highly popular news items.

Baselines

In this experimental set several methods presented in previous work are tested as baselines. These include the proposals made by Szabo and Huberman [212], Pinto et al. [191] and Asur and Huberman [18].

Based on the observation that early and future log-transformed values of popularity show a

high correlation, Szabo and Huberman [212] propose two prediction models of web content popularity: *i)* the constant scaling model, and *ii)* linear-log model.

Using the correspondence of the popularity values in a logarithmic scale, the constant scaling (*ConstScale*) model is expressed as

$$\hat{y}_j^{t_f} = \alpha_2\left(t, t_f\right) \times y_j^t, \tag{5.14}$$

where y_j^t is the popularity value of a given item n_j, received in time slice t; the final time slice is described as t_f; and α_2 is a factor which is independent of the item being predicted, defined as

$$\alpha_2\left(t, tf\right) = \frac{\sum_a \frac{p_a^t}{p_a^{t_f}}}{\sum_a \left[\frac{p_a^t}{p_a^{t_f}}\right]^2}, \tag{5.15}$$

where a is an index of all items in the training set.

As for the proposal of linear log (*Linear-log*) models, this approach is based on a linear regression procedure using the values of popularity on a logarithmic scale. This approach is described as

$$\hat{y}_j^{t_f} = exp\left(ln\left(y_j^t\right) + \beta_0\left(t, t_f\right) + \frac{\sigma_0^2\left(t, t_f\right)}{2}\right), \tag{5.16}$$

where y_j^t is the popularity of a given item n_j, received in time slice t; the parameter β_0 is computed using maximum likelihood parameter estimation given the regression fuction $ln\left(y_j^{t_f}\right) = \beta_0\left(t, t_f\right) + ln\left(y_j^t\right)$, on the training set; and the estimate of the variance of residuals on a logarithmic scale is given by σ_0^2.

These proposals are based on the value of popularity at a given time slice t, discarding past values of popularity and therefore its evolution pattern. Conversely, Pinto et al. [191] propose two linear modelling approaches which attempt to predict future values of popularity by focusing on such patterns of popularity. In the first proposal (*ML*) the authors sample the popularity of items in regular intervals (i.e. time slices) until the present time t. However, instead of using the cumulative value of popularity, the authors use the popularity of the items in each timeslice, i.e. popularity deltas.

The second proposal by the authors (*MRBF*) extend the former, by introducing a factor of similarity between the target case for prediction and training cases. This is carried out by adding feature to the learning process regarding a measure of similarity, using a Gaussian

Radial Basis Function [99]. A radial basis function is a real-valued function where values solely depend on the distance between inputs and a given point.

Finally, Asur and Huberman [18] propose the combination of behavioural features (popularity) and sentiment scores of the given texts in order to construct prediction models using linear regression (LM). Instead of using the accumulated values of items' popularity, the authors propose the use of a rate, entailing the difference between equal time intervals, i.e. time slices. In addition, the authors also use a distribution parameter as a predictive feature. Originally, this proposal is focused on the prediction of box office revenues, using social media data. Given this scope, the authors used as a distribution parameter the number of theaters a given movie is presented in. In our case of web content popularity prediction, the distribution parameter is defined as the accumulated popularity of the web content item, until the moment of prediction.

Results

In this section, results of the experimental evaluation concerning the imbalanced domain learning task of predicting web content popularity in an *a posteriori* setting are presented. The approaches tested in this evaluation include the previously described baselines, and the algorithm-level methods proposed in this thesis: the kernel-based and the kNN-based approaches (Section 5.4). These will be denoted as *kernel* and *knn*, respectively. The constant scaling and linear logarithm models proposed by Szabo and Huberman [212] are mentioned as *ConstScale* and *Linear-log*; the proposals by Pinto et al. [191] are referred to as *ML* and *MRBF*; and the approach proposed by Asur and Huberman is denoted as *LM*.

To evaluate the performance of all the approaches in *a posteriori* prediction tasks of web content popularity, the utility-based evaluation framework described in Section 4.6 is applied. Concerning the evaluation methodology used in this experiment, the Monte Carlo simulation method is applied. Results are obtained through 20 repetitions of the evaluation methodology process, using 50% of cases as training set, and the subsequent 25% as test set. This procedure is carried out using the infrastructure provided by the R package **performanceEstimation** [220].

Given the context of *a posteriori* prediction tasks, and the data used in this experimental evaluation (online news feeds), the prediction horizon is established at two days. For a given item, the ability of models in predicting its final popularity may be carried out using different levels of available data, i.e. depending on the time slice of the prediction. The objective of this evaluation is to assess the ability of prediction models in accurately predicting the popularity of highly popular news. In addition, the overall goal is also to enable such accurate prediction as early as possible. Therefore, our focus is on the first moments after the publication of the items, given that after a certain period of time the prediction becomes

obvious. In this evaluation the focus is directed towards the accurate prediction of highly popular news within the first hour of it being published. As such, results report the predictive ability of models in the first three time slices.

Figure 5.5 presents an illustration of the evaluation results concerning the combination of news in each of the topics *economy*, *microsoft*, *obama* and *palestine* and the popularity of such news according to each of the social media sources available: Twitter in the single-source data set, and Facebook, Google+ and LinkedIn in the multi-source data set. The depicted results concern the evaluation metric F_1^u. In Annex E results concerning all metrics are described, where for all combinations of social media sources and news topics, the best result of each evaluation metric is denoted in bold.

Results obtained with the *rmse* evaluation metric show that the best predictive approach is the *ConstScale* model, proposed by Szabo and Huberman [212], concerning all the first three time slices. This outcome is relatively unexpected since most of the remaining baseline models in this experimental evaluation have used such model as a baseline, concluding that their proposals provided an increased predictive ability. Regardless, it is also observed that the algorithm-based methods proposed in this thesis are capable of obtaining considerable results, presenting the best overall outcome when excluding the constant scaling model proposal.

When comparing the top performing approaches in this experimental evaluation concerning the $rmse_\phi$ evaluation metric, it is observed that the same three approaches provide the best overall results in all three initial time slices: the *ConstScale*, *kernel* and *knn* models. However, unlike the outcome concerning the standard evaluation metric *rmse*, results concerning this metric show that the best overall results are given by the predictions of the *kernel* approach, in the time slices pertaining to the first hour after the items are published.

Focusing on the utility-based evaluation metric F_1^u, results clearly show an advantage concerning the methods proposed in this thesis, the *kernel* and *knn* approaches. As such, this outcome confirms the intuition motivating such proposed methods, which denotes the possible issues concerning previously proposed approaches and the influence of the imbalanced distribution of web content popularity. Nonetheless, it should be noted that *Linear-log* models show a poor predictive performance concerning the standard evaluation metric *rmse* and the utility-based $rmse_\phi$ metric. However, concerning the metric F_1^u, such models show a considerable advantage in comparison to other baselines. This shows that *Linear-log* models, although presenting non-optimal performance concerning cases with a level of popularity considered as normal, it shows a considerable ability to predict highly relevant cases in comparison to other baselines.

Given the outcome provided by the experimental evaluation performed, results show that regarding the objectives of accurately predicting highly popular news, and doing so in the

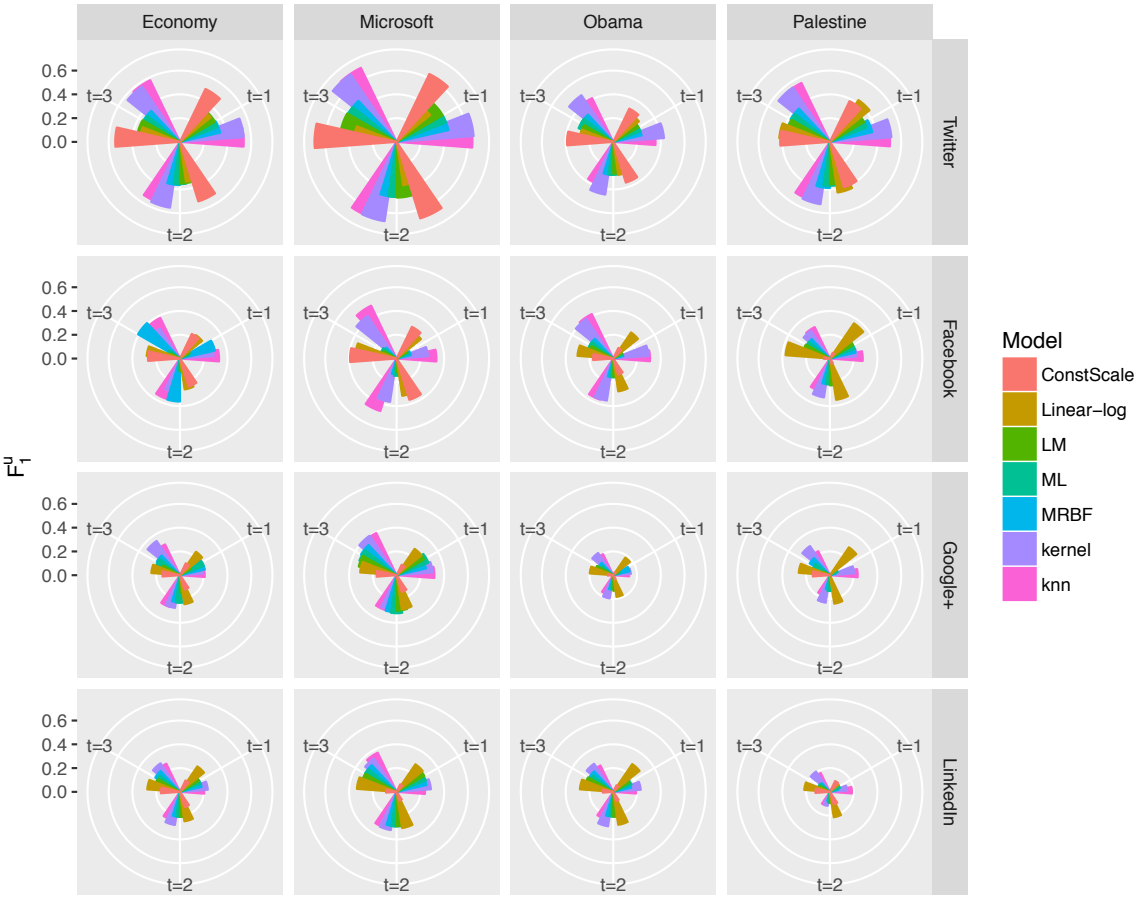


Figure 5.5: Evaluation results of prediction models for *a posteriori* prediction tasks, concerning the F_1^u metric, for all combinations of the available news topics and the popularity scores given by all social media sources, in the first three time slices.

shortest amount of time after items are published, the proposed algorithm-based methods present the best overall results. In order to confirm such claim, a study of the statistical significance of their predictive ability is provided, using critical difference diagrams [61]. The objective is to test if the improvements shown by the proposed algorithm-based methods are statistically significant ($p-$values < 0.05) in comparison to previous work approaches. Results are presented in Figures 5.6, 5.7 and 5.8.

An analysis of the results obtained by the application of critical difference diagrams show that the proposed algorithm-level methods are capable of providing a significant predictive

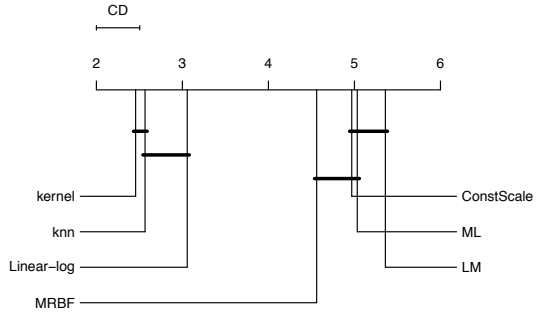


Figure 5.6: Critical difference diagram concerning the results of the evaluation metric F_1^u for models in *a posteriori* prediction at timeslice 1.

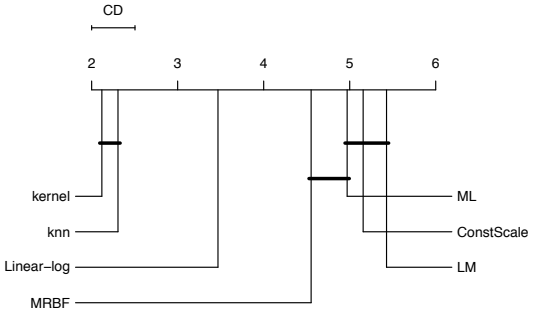


Figure 5.7: Critical difference diagram concerning the results of the evaluation metric F_1^u for models in *a posteriori* prediction at timeslice 2.

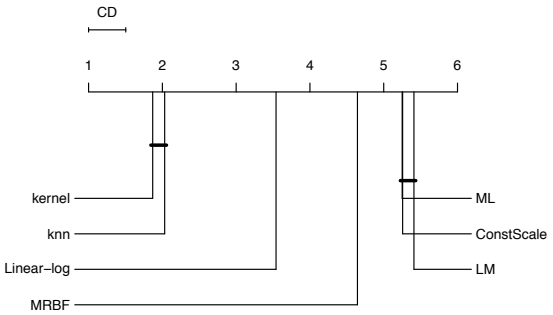


Figure 5.8: Critical difference diagram concerning the results of the evaluation metric F_1^u for models in *a posteriori* prediction at timeslice 3.

advantage in comparison to all baselines used in this experimental evaluation, and that the *kernel* method obtains the best overall performance. It is also observed that this conclusion holds true for all time slices analysed: the first three time slices, concerning the first hour of news alive-time.

Evaluation of Time-Based Ensembles

Given the analysis of the predictive ability concerning both *a priori* and *a posteriori* prediction tasks of web content popularity, an evaluation of the hybrid methods proposed in this thesis is provided. The proposed time-based ensembles are hybrid methods that combine data-level and algorithm-level methods in order to improve predictive accuracy in imbalanced domain learning tasks. This proposal defines the combination of prediction models with a weighted averaging approach. Both types of methods are combined with the following time-sensitive process: the weight of predictions concerning approaches using data-level methods and those using algorithm-level methods are dependent on the amount of time passed since

the publication of the items and the proportion of available data observed in the training sets at a given time. The proposal for time-based ensembles is presented and described in Section 5.5.

Concerning the data-level methods used, this evaluation resorts to the results presented by the evaluation of *a priori* prediction tasks, illustrated in Section 5.6.1. For each combination of social media source (Twitter for the single-source data set, and Facebook, Google+ and LinkedIn for the multi-source data set) and the topic of news items (*economy*, *microsoft*, *obama* and *palestine*), this evaluation of hybrid methods uses the approach which obtained the best score concerning the utility-based evaluation metric F_1^u. As for the algorithm-level methods employed in this experimental evaluation, given the clear advantage concerning predictive accuracy towards highly popular news by the proposed kernel- and kNN-based methods (denoted as *kernel* and *knn*), these will also be tested in the evaluation of the time-based ensembles.

Using the same settings as in the previous evaluation, the objective of this second evaluation within the scope of *a posteriori* prediction tasks is to evaluate the predictive accuracy of the proposed time-based ensembles concerning highly popular news items, in comparison to the sole use of *kernel* and *knn* methods. The proposal of time-based ensembles includes two approaches for the combination of predictions from data-level and algorithm-level methods: *i)* using the predicted values of each type of method, and *ii)* using the relevance of their predicted values. The former is denoted with the prefix *ENSt*, and the latter with the prefix *ENSphi*, e.g. *ENSt_kernel ENSphi_kernel*. This nomenclature is used throughout this section, in order to refer to each of the time-based ensemble alternatives, and the algorithm-level method used.

Results concerning the utility-based evaluation metric F_1^u are illustrated in Figure 5.9 and the results of all evaluation metrics employed are detailed in Annex F, where the best result according to each evaluation metric is denoted in bold. Such results show that the proposals of time-based ensembles present distinct results regarding the various evaluations metrics, and that this is mainly related to the data setting concerning the news topic and the social media source in question.

A careful analysis of the results presented in Annex F shows that the proposed time-based ensembles are capable of improving the predictive performance of models in comparison to the use of the proposed *kernel* and *knn*-based methods.

Concerning both the *rmse* and the $rmse_\phi$ evaluation metrics, overall results show that both time-based ensemble proposals are capable of improving the results of the base models obtained by *kernel* and *knn* methods. The same conclusion is obtained when concerning the utility-based evaluation metric F_1^u: in all combinations of social media sources data and each news topic tested, for all the first three time slices, the time-based ensemble proposals

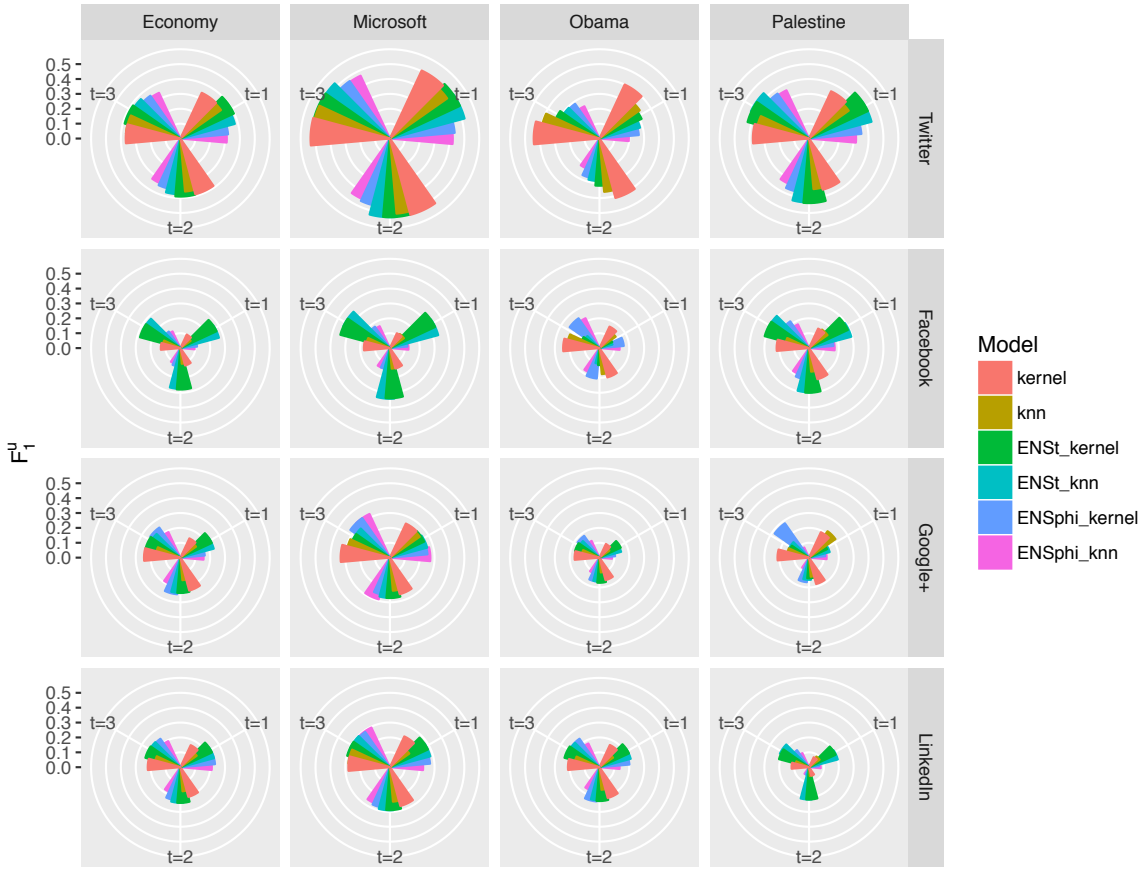


Figure 5.9: Evaluation results of prediction models concerning algorithm-based and hybrid methods in *a posteriori* prediction tasks, regarding the F_1^u metric, for all combinations of the available news topics and the popularity scores given by all social media sources, in the first three time slices.

are capable of improving the predictive performance of both *kernel* and *knn*-based models.

In the following section, the results obtained in the experimental evaluation sets provided in this chapter are discussed. Several issues presented in the analysis of results are addressed, and further insights concerning the ability of the proposed approaches in this thesis are provided, regarding the prediction of highly popular news.

5.6.3 Discussion

The experimental evaluation sets provided in the previous sections are focused on evaluating the impact of using standard learning algorithms in web content popularity predictions tasks. As stated, this task is considered to be an imbalanced domain learning task.

Experiments are divided in two groups concerning the type of prediction task: *i) a priori* or *ii) a posteriori* prediction tasks. The former is focused on predicting the final values of popularity for news items before or when these are published. As such, it is assumed that observations of the dynamics of news popularity are unavailable. Concerning the latter, these are mainly focused on modelling the dynamics of news popularity, in order to predict their final values. In this latter case, the prediction is dependent of the amount of time passed since the publication of the item: as time passes, more observations of popularity are available. This presents a paradox, given that as time passes the final values of popularity also become considerably evident. Therefore, the objective in *a posteriori* tasks is not only to accurately predict highly popular news, but to do so in the shortest amount of time possible.

A Priori Prediction Tasks

Concerning the first set of experiments focusing on *a priori* tasks, the objective is to evaluate if by resorting to data-level methods it is possible to improve the accuracy of prediction models, concerning on highly popular news. In order to do so, several standard learning algorithms are used, in order to ensure that conclusions are not biased. Results show that data-level methods are indeed capable of significantly improving the predictive accuracy of prediction models towards under-represented cases, when compared to models where such methods are not applied. This conclusion also holds when comparing the results with the work of Bandari et al. [24], a well-known approach in the context of *a priori* prediction tasks.

Furthermore, regarding the proposals made in this chapter concerning context-bias resampling strategies, results show that these are capable of providing an advantage over non-biased strategies: the best overall results are obtained by models combining the application of the undersampling strategy with temporal bias, and the use of the standard learning algorithm MARS to learn the models. However, this conclusion varies when addressing each of the baseline resampling strategies: *i)* random undersampling, *ii)* random oversampling and *iii)* SMOTEr. Results show that concerning the random undersampling and oversampling, the advantage of their context-bias variants is observed in most cases. However, concerning the SMOTEr strategy, the proposed variants only show a significant advantage when models are built using the Random Forest learning algorithm

In order to provide additional insights concerning the impact of applying data-level methods (resampling strategies), the following test is carried out. Three of the best performing

and worst performing prediction models are chosen according to the results of the critical difference diagrams applied in the respective experimental evaluation. These relate to the evaluation of the prediction models using the utility-based evaluation metric F_1^u and a p-value < 0.05. The critical difference diagrams are again used concerning the components of F_1^u: the utility-based precision ($prec_\phi^u$) and recall (rec_ϕ^u) metrics. The objective is to better understand the predictive accuracy improvement towards under-represented cases, when data-level methods are applied. Results are depicted in Figures 5.10 and 5.11. The best performing models consist of the application of the MARS learning algorithm with the undersampling strategy and its respective variants. The worst performing models include those built with the approach proposed by Bandari et al. [24] and the SVM algorithm without the application of resampling strategies and by applying oversampling with temporal bias.

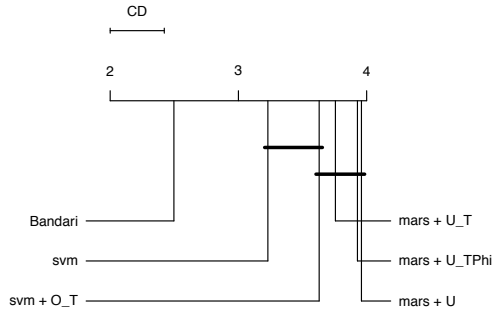


Figure 5.10: Critical difference diagram concerning the results of the evaluation metric $prec_\phi^u$ for the three best and three worst *a priori* models according to the F_1^u metric.

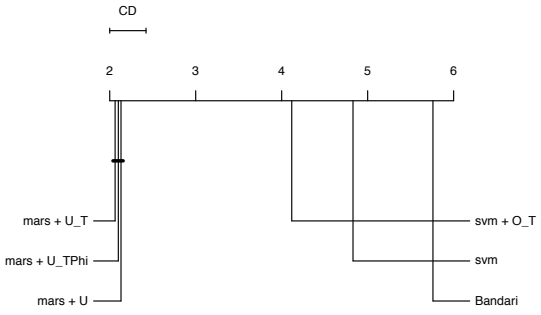


Figure 5.11: Critical difference diagram concerning the results of the evaluation metric rec_ϕ^u for the three best and three worst *a priori* models according to the F_1^u metric.

Results provide an illustrative example of how data-level methods operate, as well as the impact that they have on imbalanced domain learning tasks. The utility-based precision metric ($prec_\phi^u$) is focused on assessing the utility of models' predictions in cases considered as highly relevant, i.e. the relevance score of the predicted values is higher than the relevance threshold. Results show that the best outcome comes from prediction models which obtain poor F_1^u results. This means that one of the problems associated with the application of data-level methods is that prediction models become more prone to forecasting items with highly relevant popularity values, when their true value is considered to be of normal relevance.

On the other hand, the objective of the utility-based recall metric (rec_ϕ^u) is to measure the utility of predictions concerning cases where their true target value is considered to be highly relevant. In this case, results show that the best outcome is given by the models presenting the best evaluation concerning the F_1^u. This shows that such prediction models are able to accurately predict a greater number of highly relevant cases. Also, in comparison to the worst three predictions models (w.r.t. F_1^u), results show that the gap is more significant. This demonstrates the reason for their poor predictive accuracy towards highly popular

cases: although accurately predicting a small amount of highly popular cases, the majority of such cases are predicted with low levels of popularity.

Given such results, one is able to conclude that the application of data-level methods presents an interesting trade-off, in terms of their ability to predict highly relevant cases. These methods cause prediction models to incur in a greater number of cases where the predicted value is considered to be highly relevant, but the true target value is considered to be normal. However, they are also much more able to accurately predict the rare cases of highly popular web content.

Given the results provided by the utility-based evaluation metric F_1^u in the experimental evaluation of *a priori* models, it may be concluded that the increased predictive accuracy towards highly relevant cases associated to the application of data-level methods, outweighs the impact of cases affected by misleading predictions.

A Posteriori Predictions Tasks

Regarding the second set of experiments, these are focused on *a posteriori* tasks. The objective of these experimental evaluations is to assess the predictive ability of several proposals concerning algorithm-based and hybrid methods, focusing on highly popular news. First, the predictive accuracy of algorithm-based methods is compared to several state-of-the-art approaches for web content popularity prediction. Results show that the proposed algorithm-based methods provide the best overall results concerning predictive accuracy towards highly popular cases. Given such results, a second experimental evaluation is carried out concerning the proposed hybrid methods (time-based ensembles), in comparison to the proposed algorithm-based methods. Results show that the hybrid methods are capable of improving the predictive ability of models, in comparison to algorithm-based methods.

The algorithm-based methods proposed in this chapter focus on the concept of local averages. Both kernel- and kNN-based approaches build on such concept, and their main difference relates to the former, which is based on the concept of local weighted averages. These proposals show a clear improvement of predictive accuracy towards highly popular news, in comparison to all state-of-the-art baseline approaches. Regardless, the focus of the experimental evaluation presented concerns the first moments after the publication of news, i.e. the first hour. In order to further understand the impact of the proposed algorithm-based models, in Figure 5.12 the evolution of the metric F_1^u for all 144 time slices (i.e. two days) concerning the constant scaling [212] and the proposed kernel-based method is illustrated. This illustration uses data from the social media source Google+ and the topic "microsoft".

Results clearly show that the proposed kernel-based method provides a significant advantage concerning the first hours of news' alive-time. In this case, results show that the kernel-based

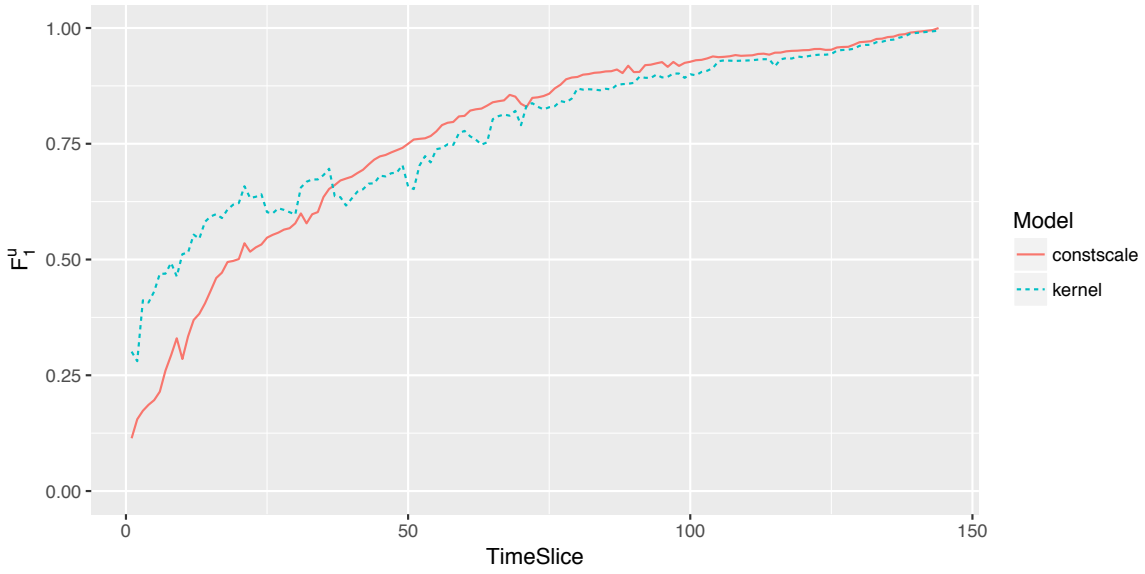


Figure 5.12: Evaluation results of ConstScale and Kernel models regarding the F_1^u metric, using data from social media source Google+ and topic "microsoft", for all time slices.

method provides such advantage for roughly the first 10 hours after news are published. Such outcome is also observed in the majority of combinations of social media sources and news topics data, for both the proposed methods, in comparison to the baselines used in the experimental evaluation. It should be noted that after a considerable period, the majority of baseline methods tested are capable of achieving or providing better results than the proposed methods. Nonetheless, after such amount of time, the relevance of news items is most probably already obvious.

Concerning the proposed hybrid methods, in comparison to the proposed algorithm-based methods, results show that the hybrid methods are capable of providing a considerable improvement in terms of predictive performance. Although this observation is visible in terms of the difference of results w.r.t. data from different social media sources, this claim should be confirmed statistically. To achieve such objective, Wilcoxon signed rank tests are employed, in order to assess the statistical significance (p-value < 0.05) of the results obtained by the hybrid methods using each of the proposed algorithm-based methods as a baseline, concerning the utility-based evaluation metric F_1^u. These results are aggregated by the respective social media source, and by each of the first three time slices in which the predictions models were tested, i.e. concerns the ability of models to predict the final values of news popularity within the first hour after they are published. The outcome of the statistical tests is described in Table 5.4.

By observing the outcome of the statistical tests, results confirm the previously stated conclusions: the use of the hybrid methods proposed in this thesis (time-based ensembles) allows for a significant increase in predictive accuracy over the best performing algorithm-

Table 5.4: Number of significant ($p-$value < 0.05) wins/ties/losses according to Wilcoxon signed rank tests, concerning the F_1^u evaluation metric, for the proposed hybrid methods using the proposed algorithm-based methods as baseline, aggregated by social media source and the first three time slices.

		t = 1			**t = 2**			**t = 3**		
	Model	Wins (Sig)	Ties	Losses (Sig)	Wins (Sig)	Ties	Losses (Sig)	Wins (Sig)	Ties	Losses (Sig)
Twitter	ENSt_kernel	**4 (4)**	0	0 (0)	**4 (4)**	0	0 (0)	**4 (4)**	0	0 (0)
	ENSphi_kernel	**4 (4)**	0	0 (0)	3 (3)	0	1 (1)	4 (3)	0	0 (0)
	ENSt_knn	**4 (4)**	0	0 (0)	**4 (4)**	0	0 (0)	**4 (4)**	0	0 (0)
	ENSphi_knn	**4 (4)**	0	3 (3)	3 (3)	0	1 (0)	**4 (4)**	0	0 (0)
Facebook	ENSt_kernel	3 (3)	0	1 (1)	**3 (3)**	0	1 (1)	3 (3)	0	1 (1)
	ENSphi_kernel	**4 (4)**	0	0 (0)	2 (2)	0	2 (2)	**4 (3)**	0	0 (0)
	ENSt_knn	3 (3)	0	1 (1)	3 (3)	0	1 (1)	3 (3)	0	1 (1)
	ENSphi_knn	**4 (4)**	0	0 (0)	**4 (4)**	0	0 (0)	**4 (4)**	0	0 (0)
Google+	ENSt_kernel	**3 (3)**	0	1 (0)	3 (2)	0	1 (0)	1 (0)	0	3 (1)
	ENSphi_kernel	2 (2)	0	2 (1)	**4 (3)**	0	0 (0)	**3 (3)**	0	1 (1)
	ENSt_knn	**4 (4)**	0	0 (0)	3 (3)	0	1 (0)	3 (2)	0	1 (1)
	ENSphi_knn	**4 (4)**	0	0 (0)	**4 (4)**	0	0 (0)	**4 (4)**	0	0 (0)
LinkedIn	ENSt_kernel	**4 (4)**	0	0 (0)	**4 (3)**	0	0 (0)	**4 (4)**	0	0 (0)
	ENSphi_kernel	3 (3)	1	0 (0)	3 (3)	1	0 (0)	4 (3)	0	0 (0)
	ENSt_knn	**4 (4)**	0	0 (0)	**4 (4)**	0	0 (0)	**4 (4)**	0	0 (0)
	ENSphi_knn	**4 (4)**	0	0 (0)	3 (3)	1	0 (0)	**4 (4)**	0	0 (0)

based methods. Results also confirm that this observation is transverse to all combinations of social media sources and news topics, as well as in all the first three time slices tested (the first hour of news' alive-time).

In summary, this extensive experimental evaluation shows that it is possible to significantly improve the predictive accuracy of models, when focusing on rare cases of highly popular items. The first set of experiments concerning *a priori* prediction tasks shows that the application of data-level methods is able to significantly improve the prediction of under-represented items in the data (highly popular news), in comparison to its dismissal. This conclusion is also observed concerning the state-of-the-art approach proposed by Bandari et al. [24]. Furthermore, the impact of applying data-level methods is studied, and results show that such methods present a trade-off: more cases with popularity values considered as highly relevant are correctly predicted, but more cases are erroneously predicted as highly relevant. Notwithstanding, results show that the former significantly outweighs the latter. As for the second set of experiments, concerning *a posteriori* prediction tasks, results show that the proposed algorithm-based methods significantly improve the ability to accurately predict highly popular news in comparison to several state-of-the-art approaches. Concerning the proposed hybrid methods (time-based ensembles), results show that this approach is capable of further improving predictive accuracy in comparison to the proposed algorithm-based models.

5.7 Conclusions

In this chapter, the use of standard learning algorithms in imbalanced domain learning tasks is addressed. Previous work shows that such algorithms are prone to several issues such as *i)* the optimization of models towards the well-represented cases of the domains when using standard evaluation metrics, *ii)* the disregard of under-represented cases due to their reduced coverage, and *iii)* the interpretation of such rare cases as noise, causing them to be discarded.

To tackle this issue, distinct approaches have been proposed in previous work, mainly concerning classification tasks. Given the scope of the web content popularity domain, several approaches to each of these types of methods are proposed in this chapter. Concerning data-level methods, the concept of context-bias resampling strategies is introduced. This type of data-level methods differs from traditional strategies due to their guided case selection procedure, instead of the commonly employed random selection. Two variants of the random undersampling, random oversampling and SMOTEr [222] strategies are proposed: *i)* with temporal bias, and *ii)* with temporal and relevance bias.

Regarding algorithm-level methods, two novel approaches are proposed: *i)* a kernel-based approach, and *ii)* a kNN-based approach. The motivation for these proposals is based on the fact that the standard learning algorithms used in previous work concerning the prediction of web content popularity are prone to the previously mentioned issues. As such, these proposed approaches for prediction of highly popular items are based on the notion of local averages, building on the concept of kernel regression, and the k-nearest neighbour algorithm, respectively.

Finally, concerning the proposed hybrid methods, an ensemble approach is detailed. This approach is based on the concept of time-based ensembles and the limitations of approaches for both *a priori* and *a posteriori* tasks. In the former, approaches are not able to update their predictions when social feedback from users becomes available. As for the latter, their predictive accuracy may be affected in the first moments after the publication of web content items, due to the lack of sufficient feedback from users. In order to overcome such issues, the time-based ensembles propose the combination of approaches employed both data-level methods (in *a priori* prediction tasks) and algorithm-level methods (in *a posteriori tasks*). This is carried out in a time-sensitive manner w.r.t. the data available at a given prediction time, using a weighted averaging approach. Two proposals for time-based ensembles are presented: *i)* by combining the predicted values of approaches using data- and algorithm-level methods, and *ii)* by combining the relevance of their predicted values.

An extensive experimental evaluation concerning online news feeds data shows that the approaches proposed in this chapter for tackling the problem predicting highly popular

items provide a considerable advantage in comparison to state-of-the-art approaches. In *a priori* prediction tasks, results show that the application of resampling strategies significantly improves results over several baselines, and that the context-bias variants are capable of further improving such results. The algorithm-level methods also show that they are capable of providing a significant advantage over several state-of-the-art approaches concerning the ability to predict rare cases of highly popular items in the first moments after their publication. As for the hybrid methods, in comparison to the proposed algorithm-level models, results show that the time-based ensembles proposal is capable of further improving predictive accuracy in all scenarios tested (i.e. combination of social media sources and news topics).

The outcome of the sets of experimental evaluations carried out in this thesis show that the proposed methods are capable of improving the predictive accuracy obtained by state-of-the-art approaches, concerning the rare cases of highly popular items. Regardless, it is not clear that such improvement can translate to an improved user experience. This relates to the fact that the main motivation for the use of predictive modelling in web content popularity prediction tasks is to enable a faster recommendation of highly relevant items to users and to aid in a more accurate promotion of such content. In the following chapter, the ability of the approaches proposed in this chapter in properly ranking web content items is studied.

Chapter 6

Single and Multi-Source Ranking

In previous chapters, the problems of predicting highly popular web content and of correctly evaluating their predictive performance is studied. The main distinction of this study in comparison to previous work is that the prediction task is considered to be an imbalanced domain learning task. However, one of the major challenges concerning web content is the ability to provide meaningful suggestions of such content in a timely manner. In this chapter, the ability of the previously proposed and evaluated prediction models in providing accurate and timely suggestions of web content items is assessed. This includes the evaluation of rankings provided by the predicted values of modelling approaches in a single-source and a multi-source context.

6.1 Introduction

One of the most distinctive aspects of web content data is that it is pervasive. Accounting for the most known types of web content data, such as online videos or news feeds, the massive amount of data generated on a daily basis provides a difficult setting for the process of searching for relevant items. This is one of the main problems that recommendation systems face today.

Concerning previous work, one of the most popular approaches to deal with this problem concerns collaborative filtering techniques [1]. Such techniques are based on using the interaction of users and items in order to capture preferences and trends. Nonetheless, as previously discussed (Section 3.1), the use of collaborative filtering approaches raise several caveats. For example, in settings where data is massively and constantly generated, systems based on such techniques may not provide the best solution. This is related to collaborative filtering requiring the existence of pre-existing data concerning the interaction between users and items: for very recent items, no such data is obtainable. In addition to

this, DeChoudhury et al. [59] have also demonstrated that traditional information retrieval and ranking approaches are prone to severe issues when attempting to identify the most relevant information for users.

In order to overcome such issues, one of the most explored solutions concerns the use of predictive modelling tools. The objective of such approach is to allow for an anticipation of the items' relevance to users, allowing a more preemptive recommendation of relevant content. The previous chapters studied this problem, where the ability to accurately predict highly relevant content is addressed in both the aspects of evaluation and predictive modelling (Chapters 4 and 5).

Results show that the predictive approaches proposed in this thesis are capable of providing a significant improvement in the accurate prediction of highly popular web content. Regardless, it is unclear if their predictive accuracy translates to an improvement concerning the rankings it generates. This is the focus of this chapter, where the ability of prediction models in providing accurate rankings of web content items is studied. This study evaluates such rankings in two scenarios: *i)* single-source and *ii)* multi-source rankings.

6.2 Single-Source Ranking

From a statistics point of view, ranking tasks concern the simple transformation of numerical or ordinal target values into a sorted rank. The sorting order (ascending or descending) of the rank depends on the context of the data and the objectives defined for the task. Simply put, a rank is an ordered rearrangement of a list of items w.r.t. a given criteria. For example, in the case of web content popularity, it is expected that the most popular items are positioned in the first positions. As such, the ranking of a given set of items depicts the rearrangement of the set in a descending order of popularity, i.e. the most popular are placed first, the less popular are placed last.

Regarding the most common representation of rankings, these are usually depicted as ordinal numbers. As such, given a set of items which are properly rearranged according to a given order, their target values are replaced by an ordinal number $1, \cdots, n$, where n is the number of items in the list. In an ordinal ranking, all items receive a distinct number, including the items that have an equal target value. In order to solve ties, the random attribution of a number can be applied.

The main assumption of ordinal rankings is that, for any given pair of items, one must be ranked higher or lower than the other. Given this, rankings provide a useful tool in order to evaluate a given data set, according to a given criteria. This is the case for many tasks involving web content data. For example, consider the case of online news feeds and the task of providing a list of suggested items, given the overwhelming amount of news published

each day. In order to provide such suggestions, a ranking approach is applied, in order to sort the items as to their estimated relevance, enabling an easy selection of the content by the user. Formally, a single-source ranking task in the context of web content data may be described as follows.

Definition 6.2.1. Single-Source Ranking Task. Let X be a set of n items and Y the respective set of popularity values associated to $x_1, \ldots, x_n \in X$. Given a ranking function f that orders X according to the respective Y scores in a decreasing fashion, the resulting ranking list $x^f_{(1)}, \ldots, x^f_{(n)}$ satisfies the condition $y^f_{(1)} \geq \ldots \geq y^f_{(n)}$, where $x^f_{(n)}$ represents the $n-$th item in the ranked list and $y^f_{(n)}$ the respective popularity.

The usefulness of ranking tasks concerning web content data has evolved according to the demands of users. Initially, the main goal of rankings was to enable the development of efficient algorithms in order to provide complete sets of items in a timely manner. In comparison, today, the usefulness of rankings relates to enabling a short and accurate set of items which attempt to match as well as possible the preferences of user queries [204].

In this thesis the problem of ranking web content data is approached by solely resorting to publicly available data, and therefore, as independent as possible from any given user profile, or group of profiles. The main goal is to identify the top-ranking web content items concerning a given query, in order to provide a concise set of the most relevant items. In this context, the relevance of items is related to their popularity (i.e. amount of attention received) in social media sources such as Twitter, Facebook, Google+ or LinkedIn.

Therefore, concerning single-source ranking tasks in the scope of this thesis, the objective is to use predictive modelling tools (described in the previous chapters) in order to anticipate the popularity of web content items, enabling the derivation of rankings. The usefulness of the rankings produced by predictive modelling tools is evaluated concerning its ability to accurately suggest the most popular items in the top positions of the ranking, in a timely manner.

6.3 Multi-Source Ranking

A similar problem to the task of single-source ranking concerns the application of ranking approaches using multiple sets of preferences concerning a given list of items. In the domain of web content popularity, this translates to the task of predicting the popularity of a set of items according to data from l multiple social media sources. Such prediction sets are used to generate l rankings of the items, which are then combined in order to provide a unique final multi-source ranking.

The combination of preference lists into a single ranking is commonly referred to as rank

aggregation [125]. This is a well known problem in the field of information retrieval [68]. In such context, the main focus of rank aggregation approaches is to optimize the "correctness' of the final ranking. However, this task of rank aggregation (i.e. multi-source ranking) may be affected by Arrow's impossibility theorem [17] when the number of preference lists is equal or larger than 3. Based on social choice theory, this theorem states that when voters have three or more distinct options, no ranked voting electoral system can convert the individuals' ranked preferences into a global ranking, while also meeting a pre-specified set of criteria. Based on the domain of web content popularity, and given a setting where three lists of preferences are available for a set of web content items, Arrow's criteria may be restated as follows:

- If every social media source prefers item A over alternative B, then the group prefers A over B;
- If the popularity of items A and B remain unchanged in all social media sources, then the group's preference between A and B will also remain unchanged (despite changes in popularity of other pairs of items);
- There is no "dictator": no single social media source possesses the power to always determine the group's preference.

As such, in a multi-source ranking setting where 3 or more social media sources are considered, a unique top-k ranking may not be possible to obtain [204][1]. Therefore, the objective of multi-source ranking tasks (also known as rank aggregation), is to derive a ranking of a given set of items, which maximizes the coherence of the final ranking with all of the individual sources' rankings [125]. Formally, the multi-source ranking task may be defined as follows, for the domain of web content popularity.

Definition 6.3.1. Multi-Source Ranking Task. Let X be a set of n items, S a set of social media sources' preference lists, and Y^1, Y^2 and Y^3 the respective sets of popularity values associated to items $x_1, \ldots, x_n \in X$ and to preference lists $s^1, s^2, s^3 \in S$. Given the ranking functions f, g and h that order X according to the respective Y^1, Y^2, Y^3 scores and in a decreasing fashion, the resulting ranking list $x_{(1)}^{f,g,h}, \ldots, x_{(n)}^{f,g,h}$ maximizes the condition $y_{(1)}^{f,g,h} \geq \ldots \geq y_{(n)}^{f,g,h}$, where $x_{(n)}^{f,g,h}$ represents the $n-$th item in the aggregated list.

Concerning the domain of web content popularity, the objective of the multi-source ranking task is to combine the rankings of a set of items according to their popularity in multiple social media sources, in a manner that the amount of items which are presented in each sources' top positions is maximized in the final aggregated ranking.

[1]It should be noted that several proposals have provided solutions to scenarios where some of Arrow's criteria are relaxed [69]

Several approaches to the problem of multi-source ranking tasks in top-k lists have been proposed in related work (see [147]). Some of the most popular approaches are Borda-inspired methods and Markov Chain based methods. The original method proposed by Borda [112] stated that the aggregation of ranks could be carried out by an arithmetic average. Other approaches have proposed to use other statistical methods such as the median, the geometric mean, and the l-2 norm [147]. Concerning Markov Chain based methods, Lin also proposes an approach using ergodic Markov Chains. In such proposal, a larger probability in the stationary distribution will correspond to a higher rank of the corresponding item. The author proposed three variants of this approach: *i)* spam sensitive, *ii)* majority rule, and *iii)* proportional.

In order to evaluate the ability of predictive modelling approaches in deriving accurate and timely rankings of web content items, the following section provides a thorough experimental evaluation concerning both single- and multi-source ranking tasks.

6.4 Experimental Evaluation

In this section an experimental evaluation is carried out in order to assess the ability of predictive modelling tools in providing accurate and timely rankings of web content items, based on their levels of popularity. Two sets of experiments are presented, using the online news feeds data presented in Chapter 3. The first concerns the ability of prediction models in generating single-source rankings. The second set aims at evaluating their ability in generating multi-source rankings. Concerning the predictive modelling tools employed in these experiment sets, such tools are described in Chapter 5.

The data used in these experiments concerns the rankings provided by official media sources (Google News and Yahoo! News). Given that news items provided in such official media rankings concerns items that have already been published, the experiments will be mainly focused on the ability of *a posteriori* approaches in generating rankings for both single- and multi-source rankings.

Data

In these sets of experiments the data used concerns the two data sets of online news feeds presented in Chapter 3. However, unlike the experimental evaluation of prediction models provided in the previous chapter which concerns news items, in this evaluation the train and test cases are rankings provided by the official media sources Google News in the single-source data set, and by Google News and Yahoo! News in the multi-source data set.

As previously stated, the main difference between these data sets concerns the cardinality

of official and social media sources. The single-source data set contains data regarding news items obtained from the news recommender system Google News, and their respective observations of popularity from the social media source Twitter. In the case of the multi-source data set, data concerns news items obtained from both Google News and Yahoo! News, and the observations of the news items popularity from the social media sources Facebook, Google+ and Linkedin.

The data acquisition process for both data sets is the same. The top-100 news rankings from the official media sources are retrieved in intervals of 20 minutes (time slices). These news rankings concern four different news topics: *economy*, *microsoft*, *obama* and *palestine*. For each retrieval of the top-100 news rankings, the popularity of all known items with an alive time under 2 days (prediction horizon) is queried with all social media sources.

Given this description of the data used in the experiments described in this experimental evaluation, the objective of these experiments may be further specified: given a train set of rankings containing news items, the aim of this experimental evaluation is to assess the ability of web content popularity prediction approaches in accurately ranking news items presented in rankings provided by official media sources. In order to build both *a priori* and *a posteriori* prediction models, their respective training sets include all news items presented in known news rankings, and their respective format is described in Sections 5.6 and 5.6.

It should be stressed that the ground-truth values for the each predictive modelling approach is the popularity of news according to each of the social media sources. Also, in assessing the ground-truth ranking order for a given set of news, the ground-truth (preferences) are obtained by a descending order of popularity from the news item. However, it should be noted that in a given ranking news items have differentiated alive times. This may lead to two types of issues. First, as previously stated concerning *a posteriori* prediction tasks, after a considerable amount of time passes since the publication of a given news item, its popularity becomes obvious. Second, given the diversity of alive times of news, it is possible to have a majority of news that were not published recently, for which their popularity is already known.

Given this, in order to ensure that the evaluation of the single-source and multi-source ranking tasks is capable of assessing the ability of predictive modelling tools in *i)* detecting news cases which are considered to be highly relevant, and *ii)* to do so in the shortest amount of time possible, a time-sensitive weight is applied to the final (accumulated) values of news popularity. It is defined that the ground-truth popularity of items is given by the product of the final popularity of each news item, and a discount factor that emulates the loss of novelty and interest in a news items as time passes. Given a prediction horizon of 2 days in the data sets used in this experimental evaluation, the final time slice (consecutive periods of 20 minutes) is 144 ($t_f = 144$), and the ground-truth popularity values are formulated as follows:

$$y_i^t = y_i^{t_f} \times \frac{t_f - t}{t_f}, \tag{6.1}$$

where y_i^t is the ground truth-popularity of a given item i in time slice t for ranking evaluation, and $y_i^{t_f}$ is the true amount of popularity accumulated by a given item. It should be stressed that when describing *a posteriori* tasks the nomenclature y_i^t denotes the amount of popularity accumulated by a given item until a time slice t. In these experiments, such notation represents the ground-truth popularity of a given item i and an alive time discount. By applying this procedure, this evaluation ensures that no advantage is given to news items which are no longer recent, and for which their magnitude of popularity already became obvious.

Finally, concerning the use of the data sets presented in Chapter 3 (single- and multi-source data sets): for single-source ranking tasks, the evaluation is carried out for each of the social media sources in both data sets, using ranking cases from the official media source Google News; as for the multi-source ranking tasks, these are tested using data from the social media sources in the multi-source data set, resorting to rankings data from both the Google News and Yahoo! News official media sources.

Baselines

Concerning the single-source ranking task, the effectiveness of the proposed prediction models in generating news rankings is compared to three baseline strategies:

- *Time*: news are ranked by time of publication with the most recent first;
- *Live*: news are ranked by the amount of popularity accumulated until the reference time (time of prediction);
- *Source*: news are ranked by the average popularity of other known news items from their news outlet.

The first two baselines are simple heuristic strategies, usually employed by news aggregators to promote popular content [215]. As for the third, although no reference of its use is found in previous work, it is an acceptable hypothesis that there are news outlets that gather more attention than others. Therefore, this baseline is introduced in order to assess if knowledge of the news outlet is enough to provide optimal news recommendations.

As for the multi-source ranking tasks, according to the previously described Arrow's impossibility theorem, there is no perfect ordering of rankings capable of converting the preferred items (i.e. the most popular news items) by each of the social media sources (when 3 or

more), into a group-wide unique ranking while maintaining the fairness criteria described in Section 6.3. Regardless, the objective is to apply rank aggregation methods in order to maximize the match between the aggregated rankings and each of the social media sources. As such, as baselines, the rankings produced by rank aggregation methods are compared to each of the single-source rankings derived from predictive modelling tools. In addition, the average rank (AR) of single-source rankings is also employed as a baseline.

Evaluation Metrics

Given the importance of the ranking order to news recommendation, the following metrics are applied: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain ($NDCG@k$) metrics. The first metric is focused on the global outcome, and the remaining metrics consider the position of a given item in the ranking. These are thoroughly described in Section 2.4.

Average Precision computes the average precision for all values of k where k is the rank, n is the number of retrieved items and Rel_k is a binary function evaluating the relevance of the k^{th} ranked item, attributing 1 to the relevant items at rank k and 0 otherwise. The Mean Average Precision (MAP) metric computes the fraction of relevant documents retrieved over a set of queries. It should be noted that in this experimental evaluation it is established that the relevant items are those which belong to the top-10 positions of the ranking.

Taking into account the ranking positions of items, the Reciprocal Rank is defined as the inverse of the rank at which the first relevant item is retrieved. As in MAP, the Mean Reciprocal Rank (MRR) is defined as the average of the reciprocal ranks over all queries.

The Normalized Discounted Cumulative Gain ($NDCG@k$) measures the search result quality of the ranking function by assigning high weights to documents in highly ranked positions and reducing the ones found in lower ranks. Unlike the previous ranking evaluation metrics which use a binary definition of items relevance, the Its uses multiple ad-hoc relevance judgments which are associated to ranked items in a given query. The normalization of this metric is obtained by using the ideal ordering of the ranking.

The evaluation of the single-source ranking tasks resort to all of the mentioned evaluation metrics. As for the multi-source ranking tasks, the experimental set is focused on the outcome concerning the $NDCG@k$ metric, which is the most robust [203].

Prediction and Ranking Approaches

In the sets of experiments presented in this experimental evaluation, two types of approaches are tested which relate to each type of ranking task. Concerning the single-source ranking

tasks, the alternatives evaluated concern predictive modelling approaches. As for the multi-source ranking tasks, the approaches tested concern rank aggregation methods, using the predictive modelling tool which presented the best overall results in the single-source ranking task evaluation.

The rankings provided by official media sources such as Google News concern items that have already been published. As such, the predictive modelling tools tested in the single-source ranking task evaluation concern *a posteriori* approaches. Namely, this evaluation includes the following approaches, which are thoroughly described in the previous chapter (Sections 5.6.2, 5.4 and 5.5): the constant scaling and linear-log models proposed by Szabo and Huberman [212], the multi-linear regression with and without the use of RBF functions proposed by Pinto et al. [191], the linear regression approach with sentiment analysis features by Asur and Huberman [18], the algorithm-level (*kernel* and *knn*) methods proposed in this thesis, as well as the proposed approaches for hybrid methods (time-based ensembles *ENSt* and *ENSphi*). It should be noted that the latter (hybrid methods) also uses approaches focusing on the *a priori* task of web content popularity prediction. The best performing method in each combination of social media source and news topic is also evaluated as to its ability to provide accurate and timely rankings of news, along with the baseline approach tested in their respective experimental evaluation, the proposal presented by Bandari et al. [24]. However, the *a priori* approaches proposed in this thesis will not be tested individually since their objective is to accurately predict highly popular news before (or when) the items are published.

In the single-source ranking task, the ranking approach is simply the application of a re-ordering of the predicted cases in a descending fashion. As for the multi-source ranking task, several approach for rank aggregation are tested in order to assess if they provide a cumulative advantage in comparison to the rankings derived for each social media source using predictive tools. These include Borda methods [112] and Markov Chain approaches [148], a Cross Entropy Monte Carlo approach [149] and the Rank Product method [37].

The Borda methods tested [148] in the evaluation of multi-source ranking tasks concern the combination of Borda counts using the statistical methods *i)* average (*Borda_m*), *ii)* median (*Borda_M*), *iii)* geometric mean (*Borda_geo*), and *iv)* the l-2 norm (*Borda_l2n*). For any given setting, the Borda count attributes a number of points to each item which denote the number of items which are ranked lower. The authors also provide several variants of Markov Chains to tackle the problem of rank aggregation where a larger probability in the stationary distribution corresponds to a higher rank of the corresponding element. These variants are the denoted as spam sensitive (*MCspam*), majority rule (*MCmaj*) and proportional (*MCprop*) Markov Chains algorithms. In a different work, Lin also proposes the Cross Entropy Monte Carlo approach [149]. This approach selects random samples of data using an iterative importance sampling technique, which is then optimized by minimizing its

cross-entropy [201, 58]. Finally, the Rank Product approach proposed by Breitling et al. [37] is a non-parametric statistics approach based on ranks of fold changes. The aggregated rank is given by sorting the product of ranks in different preference lists via a geometric mean. As such, given n items and k lists of preferences (i.e. popularity) from social media sources, let $r_{i,s}$ be the rank of item i in the preference list of the social media source s:

$$RP(i) = (\prod_{s=1}^{k} r_{i,s})^{1/k} \tag{6.2}$$

All of these methods are tested in the evaluation of multi-source ranking tasks, using the rankings produced by predictions of items' popularity according to each of the social media sources.

Evaluation Methodology

Similarly to all experimental evaluations carried out in the previous chapters, the data used in the experiments detailed in this section have an implicit temporal order. As such, this requires evaluation methodologies which are capable of accurately assessing the prediction error of the approaches, while maintaining such temporal order. This is important in order to guarantee predictions models solely resort to past data in order build the models, and that predictions concern future data.

Given this, the Monte Carlo simulation method is again applied in both sets of experiments in this section: evaluation of single-source and multi-source ranking tasks. This method guarantees that all the alternative models are trained and tested with the same pair of train and test sets. The Monte Carlo simulation method randomly selects points in the data set, which are then used to select a given past window for training data, and a subsequent window for test data.

6.4.1 Evaluation of Single-Source Ranking Tasks

This section describes an experimental evaluation of single-source ranking tasks, with the objective of assessing the ability of prediction models in forecasting the popularity of news items in such a manner that the most relevant items are accurately positioned in the top positions of the ranking, in a timely manner. For such endeavour, in the scope of this evaluation, the data used concerns the top-100 rankings proposed by the official media sources Google News and Yahoo! News in both the single-source and multi-source data sets. The news items presented in the rankings of the training set are used for the learning processes of the predictive approaches proposed in the previous chapter, and enumerated in Section 6.4. It should be noted that the popularity of items is given by the social media sources presented

in the data sets. As such, this means that the number of models learned in this experimental evaluation is singular to each combination of social media sources and news topics. The social media sources used in this evaluation are Twitter (concerning the single-source data set), and Facebook, Google+ and LinkedIn (multi-source data set). The topics studied concern the terms *economy*, *microsoft*, *obama*, and *palestine*.

For each ranking in the test set, the prediction models are used in order to forecast the final popularity of the items according to each available social media source. Using such predictions, a ranking is derived and evaluated w.r.t. the ground-truth. Given that news have a very short lifetime in comparison to other types of web content, mainly due to factors such as habituation or the diversion of attention towards other news [242], the decision concerning the definition of ground-truth values (and therefore the order of the ranking considered as correct) is debatable. As time passes, and the level of news popularity becomes evident, the usefulness of anticipating the future popularity of the items becomes irrelevant. That is the main reason for the objective of this task to be focused not only on accurately predicting the highly relevant cases, but also to do that in the shortest amount of time possible after the publication of the item.

Therefore, in order to correctly assess rankings concerning these two conditions *(i)* highly relevant items are positioned in the top of the ranking, and *ii)* this prediction is only relevant if provided in a timely manner) the ground-truth value of the items popularity in a ranking is weighted by a temporal factor. Using a linear decay factor, the ground-truth values of popularity in a given ranking is given by the product of the final popularity of the item and a decay factor $\frac{t_f - t}{t_f}$, where t_f is the final time slice (the prediction horizon is two days, and therefore the final time slice is 144), and t is the time slice of prediction. This procedure is also applied to the predictions of models, in order to ensure fairness in the evaluation of their predictive and ranking ability.

Results

Experimental results are obtained through 20 repetitions of the Monte Carlo simulation process with 50% of cases (ranking cases from Google News and Yahoo! News) used as training set and the following 25% as test set, using the R package **performanceEstimation** [219]. Results are given by applying the ranking evaluation metrics MAP, MRR and $NDCG@k$, described in Section 6.4. Concerning the definition of k, given that the objective is to assess the quality of the ranking in its top positions, this value is defined as 10 ($NDCG@10$): this evaluation focuses on the quality of the ranking in its top 10 positions. The $NDCG@k$ evaluation metric also requires the ad-hoc definition of relevance judgments concerning the items' importance. Given that the rankings evaluated in this section have 100 items, it is defined in an ad-hoc manner that the judgments of relevance for an instance space of

$\{0, 1, 2, 3\}$ are as such: items with the top-10 popularity values have a relevance of 3, the remaining items in the top-25 popularity values a relevance of 2, the remaining items in the top-50 popularity values a relevance of 1, and the following values a relevance of 0.

Concerning the remaining evaluation metrics used in this experiment, it should be reminded that their objective is different. The goal of the Mean Average Precision MAP is to assess the amount of relevant cases (those in the top-10 positions of the ground-truth ranking) that are presented in the rankings proposed by the tested predictive modelling tools. As for the Mean Reciprocal Rank MRR, its objective is to assert the ability of the proposed rankings in presenting a relevant case in the top positions of the ranking.

Figures 6.1 and 6.2 illustrate the results of the metric $NDCG@10$ obtained, concerning the evaluation of the rankings obtained by the prediction models tested in this experiment, in each combination of news topics and social media sources using ranking data from the official media sources Google News and Yahoo! News[2]. Results are grouped accordingly. The choice of $NDCG@10$ relates to it being considered as a robust metric, due to the use of multi-level relevance judgments and discount factors. Results concerning the remaining ranking evaluation metrics are described in Annexes G and H, for rankings of Google News and Yahoo! News, respectively, where the best result according to each evaluation metric is denoted in bold.

From a general standpoint, the results illustrated in Figures 6.1 and 6.2 show that predictive modelling tools are capable of providing news rankings which obtain a considerable level of accuracy, according to the ranking evaluation metric $NDCG@10$. Also, by analysing the outcome of the various approaches tested in this evaluation, results from all the evaluation metrics employed show a considerable coherence as to which prediction models and strategies obtain the best and worst outcome. Results show that in the majority of cases, the best performing approaches concern the proposals of algorithm-based models (*kernel* and *knn*) and the proposed time-based ensembles. Specifically, in most cases, the evaluation of the rankings proposed by these approaches shows that the time-based ensembles approach *ENSt*, using the algorithm-level method *knn*, show the best overall results. As for the worst performing approaches, results show that the baseline strategy *Time* and the *a priori* approaches evaluated present the worst news rankings.

However, some noticeable performances are worthy of mentioning. Although presenting a low predictive accuracy in terms of predicting highly popular news, the proposal by Asur and Huberman [18] indicates a considerable ability for positioning relevant items in the top positions of their proposed rankings, as shown by the evaluation metric MRR. Nonetheless,

[2]The evaluation results obtained by the original ranking of Google News and Yahoo! News are also depicted, denoted with the tag *Official*. However, it should be stressed that this is solely for illustration purposes, since there is no evidence that the objective of the ranks proposed by these official media sources is to provide the news items that are the most popular in social media.

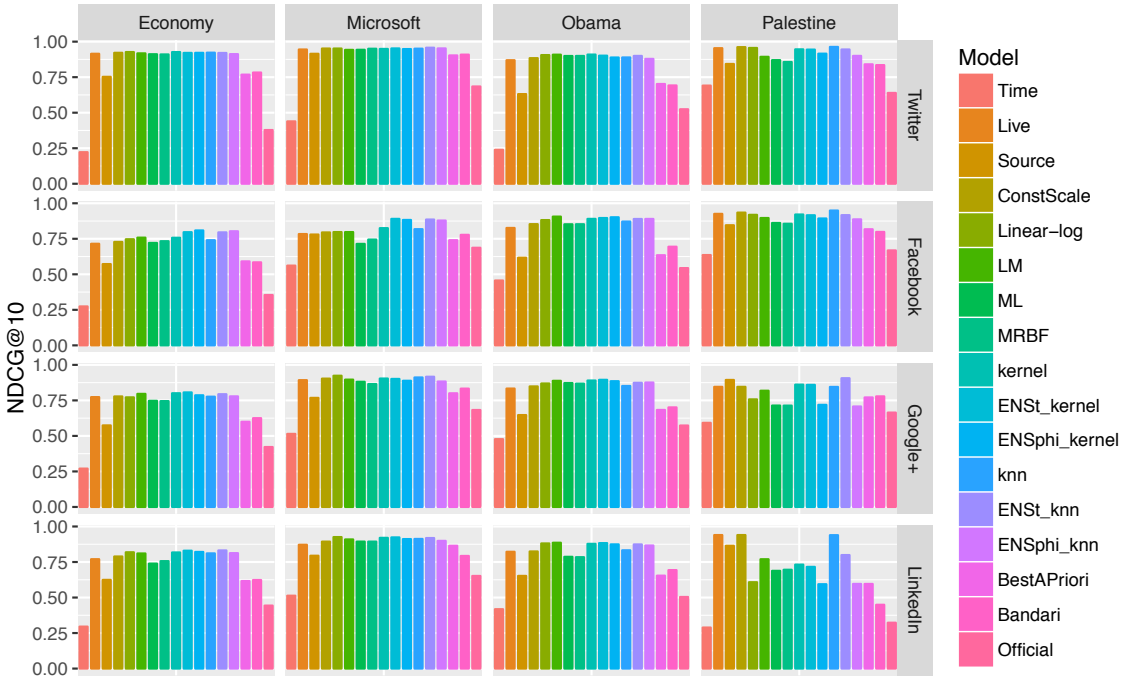


Figure 6.1: Evaluation results of single-source ranking tasks, using the $NDCG@10$ metric, concerning all *a posteriori* prediction approaches, for all combinations of the available news topics and the popularity scores given by all social media sources, using data from Google News.

the remaining evaluation metrics MAP and $NDCG@10$ show that this is not generalized to the items in the top-10 of their proposed rankings. Also, it should be noted that the rankings proposed by prediction models concerning the social media source Twitter obtain a very high score according to $NDCG@10$. In this context, it is also worth noticing that the linear-log approach proposed by Szabo and Huberman [212] are among the best overall approaches for providing news rankings, concerning the amount and accuracy of relevant items positions in the top-10 positions.

In order to confirm the overall observation and analysis of the results obtained concerning the ability of prediction models in providing news rankings that are capable of positioning highly relevant news at the top of the suggestions in a timely manner, a study of the statistical significance of the predictive modelling tools in providing such rankings is carried out. This is done by applying critical difference diagrams [61] with the objective of assessing the statistical significance ($p-\text{value} < 0.05$) regarding the pairwise comparison of all the tested approaches and their respective evaluation concerning the metric $NDCG@10$. Results are presented in Figures 6.3 and 6.4, concerning the experiments using rankings data from Google News and Yahoo! News, respectively.

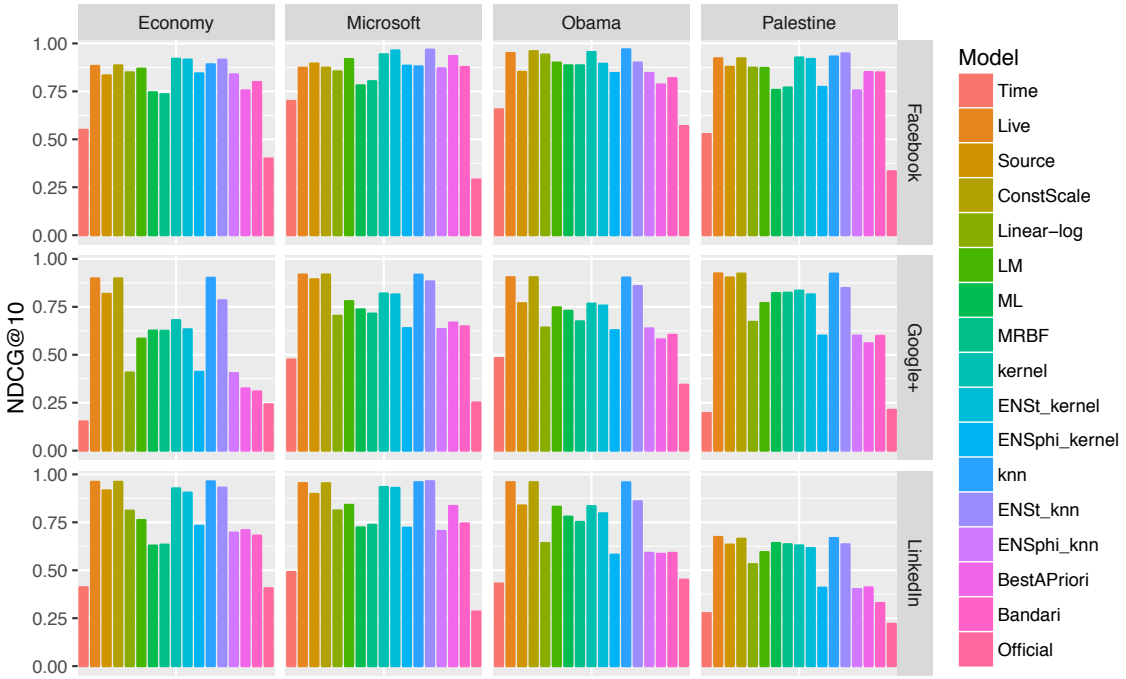


Figure 6.2: Evaluation results of single-source ranking tasks, using the $NDCG@10$ metric, concerning all *a posteriori* prediction approaches, for all combinations of the available news topics and the popularity scores given by all social media sources, using data from Yahoo! News.

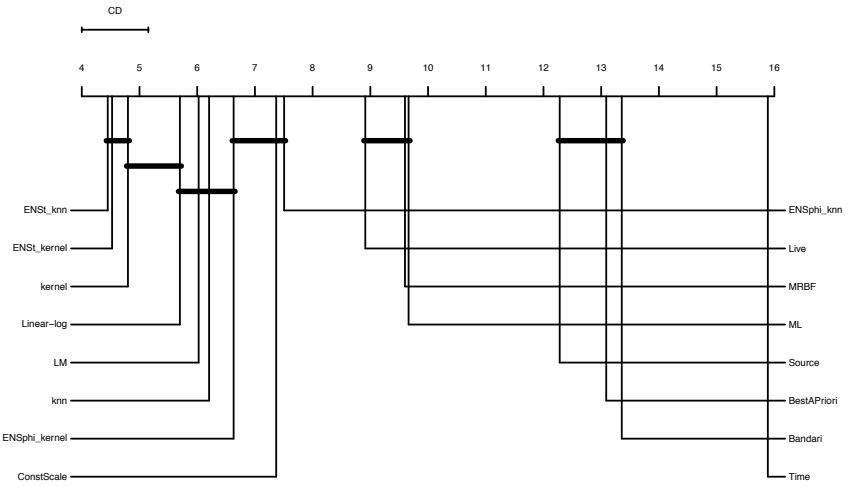


Figure 6.3: Critical difference diagram concerning all single-source ranking approaches tested, using Google News rankings, according to the $NDCG@10$ evaluation metric.

The outcome of the statistical tests applied to the predictive approaches tested in this evaluation of single-source ranking tasks confirms the overall analysis of results: in all

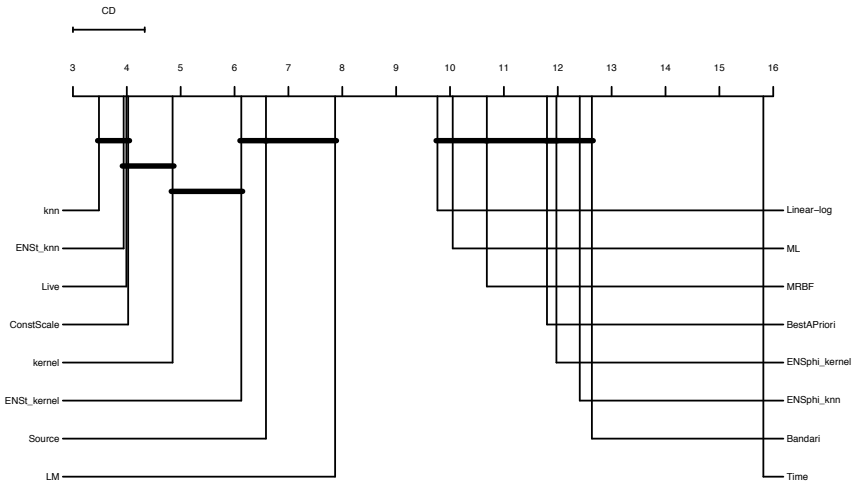


Figure 6.4: Critical difference diagram concerning all single-source ranking approaches tested, using Yahoo! News rankings, according to the $NDCG@10$ evaluation metric.

combinations concerning official and social media sources, the algorithm-based and hybrid methods proposed in this thesis provide approaches that are the most able in generating news rankings based on their predictions, given the criteria of accurately predicting highly popular cases and doing so in a timely manner. Nevertheless, despite statistical tests showing that in the case of Google News the proposed approaches are capable of providing a statistically significant advantage, in the case of Yahoo! News this is not verified. In the latter, results show that the proposed approaches do not provide a significant advantage over the baseline *Live* and the constant scaling approach proposed by Szabo and Huberman [212]. Regardless, both statistical tests confirm that the time-based ensemble approach *ENSt* using the algorithm-level method *knn* provides the best overall solution.

6.4.2 Evaluation of Multi-Source Rankings Tasks

In this section the results of the experimental evaluation concerning multi-source ranking tasks is presented. The aim of the previous evaluation concerning single-source ranking tasks is to use predictive modelling tools in order to provide news rankings which are capable of anticipating the most relevant news in a timely manner, and to position them in the top positions of the ranking. In this experimental evaluation the objective is to combine such single-source approaches in order to provide aggregated rankings concerning the popularity of news items in multiple social media sources.

The problems associated to this task have been previously presented. The main issue relates to Arrow's impossibility theorem [17]. However, several approaches for rank aggregation have been presented which relax some of the conditions presented by Arrow's theorem.

These approaches are described in Section 6.4.

The methodology of this experimental evaluation consists of evaluating the rankings obtained by single-source ranking approaches and by rank aggregation methods, having as a ground-truth the true values of each social media source, using the ranking evaluation metric $NDCG@10$. The objective is to assert if by combining the rankings provided by predictive modelling tools concerning each available social media source, these could obtain a cumulative advantage in comparison to the rankings provided by single-source ranking approaches. In this context, cumulative advantage is translated as an improvement concerning the evaluation of the aggregated rank w.r.t. to each social media source.

Results

The experimental setting of the former evaluation of single-source ranking tasks is again applied in this evaluation, with one exception. In order to proceed with the evaluation of such task it is required that the data should concern multiple social media sources. As such, this experimental evaluation is solely based on the multi-source data set. Results are obtained through 20 repetitions of the Monte Carlo simulation process with 50% of ranking cases used as training set and the following 25% as test set, using the R package **performanceEstimation** [219].

Figure 6.5 illustrates the results obtained concerning the metric $NDCG@10$, as to the evaluation of the rankings approaches tested in this experiment. The results are grouped by each scenario combining news topics and social media sources.

A thorough analysis of the obtained results clearly shows that some of the rank aggregation approaches enable a cumulative effect. This is observed when comparing the results of top performers such as *MCmaj*. Results show that these are, in the majority of cases, capable of providing a better ranking than the single-source ranking approach specific to each of the ground-truths, i.e. ranks are provided by predictions of forecasting models based on popularity data specific to the social media source. To illustrate, the *MCmaj* rank aggregation approach presents a better outcome in the context of the news topic *economy*, when compared to the ranking provided by single-source ranking approach specific to that same ground-truth (*Facebook**), as illustrated in the upper left scenario of Figure 6.5.

Concerning the overall outcome, results show that most accurate type of rank aggregation approaches is the Markov Chain methods proposed by Lin [147] with a specific emphasis on the majority voting variant (*MCmaj*), followed by the Cross-Entropy Monte Carlo method, also proposed by Lin [149].

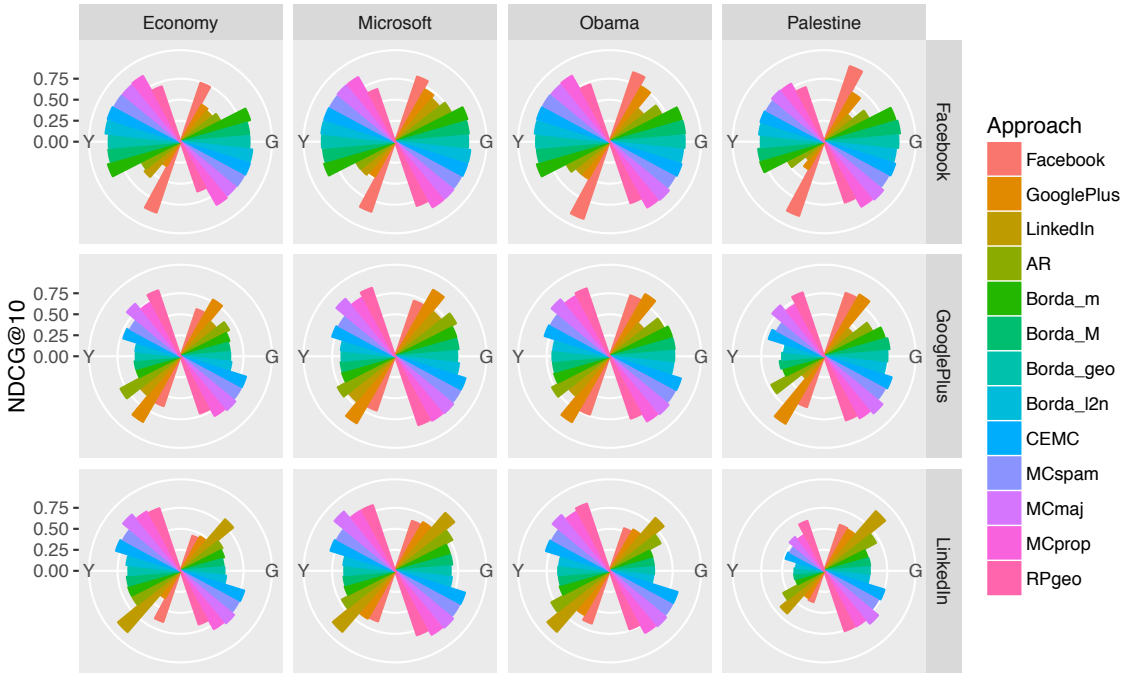


Figure 6.5: Evaluation results of multi-source ranking tasks, using the $NDCG@10$ metric, concerning all multi-source ranking approaches, for all combinations of the available news topics and the popularity scores given by all social media sources, using data from the multi-source data set.

6.5 Discussion

In this section a discussion concerning the results obtained in both sets of experimental evaluations is provided. The first set concerns single-source ranking tasks, which are focused on leveraging predictive modelling tools in order to rank news items by their popularity, regarding each of the social media sources used. As for the second set of experiments, these are focused on the task of providing aggregate rankings and to assess if such approach is capable of providing a cumulative advantage over single-source ranking approaches.

Evaluation results concerning single-source ranking tasks show that the predictive modelling approaches proposed in this thesis, namely the algorithm-based and the hybrid methods (see Chapter 5), obtain the best overall results in all combinations of news topics and social media sources, using rankings from Google News and Yahoo! News. Specifically, results show that the rankings which present the highest scores of the most robust metric used ($NDCG@10$)

is the time-based ensemble *ENSt* using the *knn* algorithm-level method[3].

Such results are relevant in terms of determining which evaluation metrics in popularity predictions tasks provide a better insight concerning the models' ability to provide accurate rankings of the most relevant items in a given set. Concerning such subject, results concerning both the web content popularity prediction tasks and the ranking tasks provided in this chapter, provide empirical evidences confirming the claims that standard evaluation metrics (e.g. $rmse$) are prone to misleading conclusions when dealing with imbalanced domain learning tasks. In addition, results also show that the outcome of single-source ranking tasks is very similar to the conclusions obtained in the experimental evaluations concerning the prediction of highly popular web content, when focusing on the utility-based evaluation metric F_1^u. As such, results provided in the first set of this experimental evaluation show that the optimization of prediction models using the metric F_1^u instead of standard evaluation metrics, is more accurate in determining the ability of such models in deriving rankings which are capable of positioning highly relevant items in the top positions of such rankings.

Regarding the multi-source ranking task evaluation, these are based on the results obtained by employing the predictive modelling tool that obtained the best outcome in the previous set of experiments: the time-based ensemble approach *ENSt* using the algorithm-level method *kernel*. The objective of multi-source ranking tasks is to provide aggregate rankings which are capable of maximizing the match between the preferences of multiple sources. In this evaluation such sources pertain to social media sources: Facebook, Google+ and LinkedIn.

Results show that several rank aggregation approaches are capable of providing a cumulative advantage in terms of their ability to accurately position highly relevant items in the top positions of the proposed rankings. In comparison to the best single-source ranking approach for each social media source, the rank aggregation approaches using Markov Chains (specifically the *MCmaj*) show an overall improvement. This advantage is also verified when focusing on the Cross-Entropy Monte Carlo approach (*CEMC*).

The results presented in the experimental evaluation of multi-source ranking tasks are grouped by social media source. In order to provide additional insights concerning the ability of rank aggregation methods in providing a cumulative advantage over single-source ranking tasks, Figure 6.6 illustrates the official media source-wide results (concerning the evaluation metric $NDCG@10$) obtained by each rank aggregation approach tested in the related experimental evaluation.

Results show that an overview of the rank aggregation approaches concerning its ability to provide better rankings of highly popular news is even more evident. Results concerning both the official media sources (Google News and Yahoo! News) reinforce the previously

[3]This approach also uses the *a priori* approach that obtain the best average outcome in the evaluation presented in Section 5.6.1.

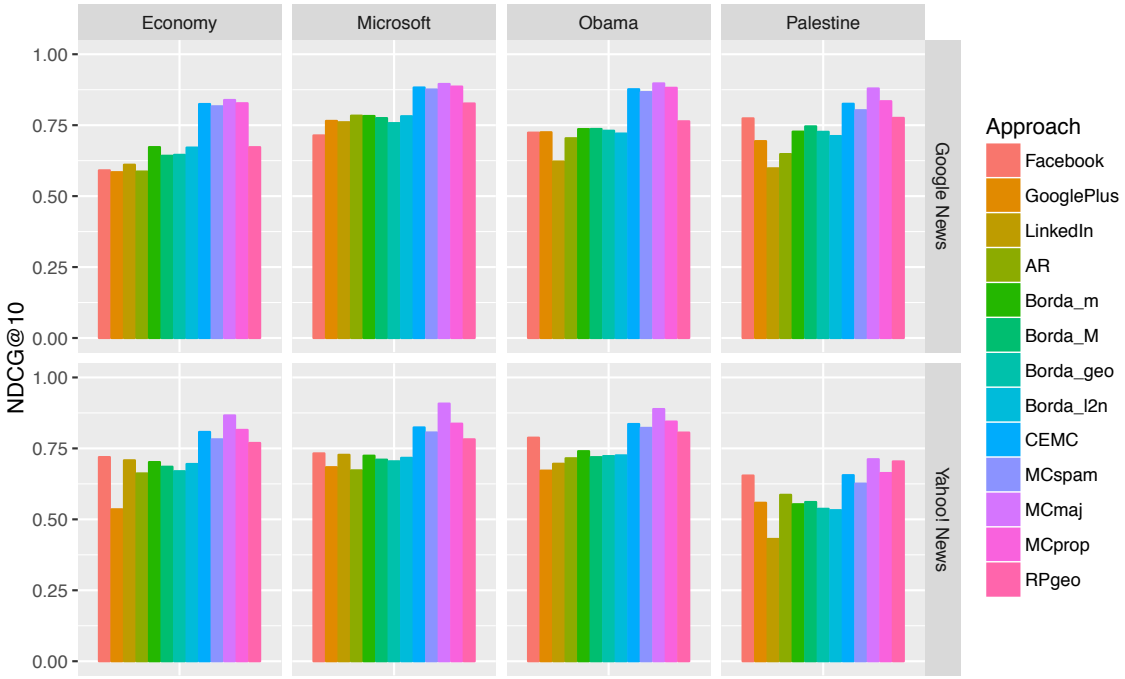


Figure 6.6: Evaluation results of multi-source ranking tasks, using the $NDCG@10$ metric, concerning all multi-source ranking approaches, for all available news topics and the average behaviour over all social media sources, using data from the multi-source data set.

mentioned conclusions, showing that in most cases the rank aggregation approaches based on Cross-Entropy Monte Carlo and Markov Chain methods are capable of providing a considerable advantage over other rank aggregation approaches, and all the single-source ranking approaches tested. Finally, it should be noted that in this evaluation perspective (official media source), the rank aggregation approach of Rank Products [37] also shows a consistent advantage over the single-source ranking approaches.

6.6 Conclusions

In this chapter the ability of web content popularity prediction models is assessed in terms of their ability to generate accurate and timely rankings of news. These comprehend two types of ranking tasks: single- and multi-source ranking. The first task concerns the simple task of ordering the predictions of models in a descending order, i.e. the most popular first. The second task is focused on maximizing the matching of preferences between an aggregated rank and the ground-truth ranks regarding each of the social media sources tested.

Experimental results show that, concerning the single-source ranking task, the algorithm-level and hybrid methods proposed in the previous chapter are the most accurate in deriving

rankings which are capable of positioning the most relevant news items in its top positions. Concretely, results show that the best single-source ranking approach is based on using the predictions of the time-based ensemble approach *ENSt* using the *knn* algorithm-level method and the predictive modelling approach which obtain the best average result in the evaluation of *a priori* prediction tasks (see Section 5.6.1).

Concerning the task of multi-source ranking, results show that the ranking ability provided by predictive modelling tools can be augmented by applying rank aggregation methods. This shows that it is possible to obtain a cumulative advantage through such combinatorial approaches, and that these are capable of providing a considerable improvement over the best overall approaches in single-source ranking tasks.

Finally, a comparison between the conclusions provided by this experimental evaluation of single- and multi-source rankings, and the outcome of the evaluation concerning the ability of prediction models in predicting highly popular news, provides interesting insights as to the impact of using standard evaluation metrics in the latter task.

The single- and multi-source source ranking evaluation results provide evidence to support the claim that standard evaluation metrics are not the best option when concerning tasks where the objective is to accurately predict target values which are under-represented in the domain: the best prediction models according to the standard evaluation metric $rmse$ are not amongst the top performing approaches in deriving rankings. Conversely, this experimental evaluation of ranking tasks shows that the utility-based evaluation metric F_1^u is more appropriate when optimizing models for such imbalanced domain learning tasks: results from both the web content popularity prediction tasks and these experimental evaluations concerning rankings tasks show evidence of agreement as to the conclusions provided by the F_1^u and $NDCG$@10 evaluation metrics.

Chapter 7

Conclusions

The possibility of every user to generate publicly available content provoked an explosion of available data and a growing demand concerning the computation capability to analyse, interpret and act upon this referred data. The exponential growth of available information to users poses many questions and challenges to researchers, which is the main motivation driving the work presented in this thesis.

Given such profusion of web content one of the main challenges concerns the ability to correctly identify/anticipate which of the content is relevant. This is of the most importance when concerning tasks such as suggesting or promoting web content. However, a careful analysis of the related work shows that a distinctive characteristic web content has been considerably neglected: the distribution of web content popularity. This distribution is highly skewed, where only a few items receive relevant levels of attention. This introduces the problems addressed in this thesis.

In this thesis, the problems of predicting and ranking highly popular web content are addressed. Unlike previous work, such tasks are solely focused on the accurate and timely anticipation of the levels of popularity (attention) that web content items receive, and their evaluation is mostly focused on the ability to accurately identify the rare cases of highly popular items. Given the imbalanced domain of web content popularity, previous work shows how the application of standard learning and evaluation tools are prone to under-performing models and erroneous conclusions.

The objectives of this thesis include *i)* the study of the web content popularity domain; *ii)* the formalization of the popularity prediction task as an utility-based regression task; *iii)* the proposal of evaluation approaches in order to accurately assess the ability of prediction models in detecting highly relevant cases; *iv)* the proposal of predictive modelling tools focusing on the accurate prediction of highly popular web content items and *v)* their ability to provide useful rankings.

Contributions

This thesis has produced the following contributions:

Literature Review of Web Content Popularity Prediction and Ranking. In Chapter 2 a thorough review of previous work concerning both prediction and ranking tasks using web content data is provided. A discussion is provided regarding several issues that may hinder the ability of previous proposals in accurately predicting and ranking highly popular web content items.

Study of Online News Feeds Data. A study of online news feeds data, a type of web content, is provided in Chapter 3. Its characteristics are thoroughly analysis by means of an exploratory study concerning two new data sets which are proposed in this thesis: a *i)* single-source, and a *ii)* multi-source online news data set.

Popularity Prediction as an Imbalanced Domain Learning Task. Based on the study concerning online news feeds data, it is concluded that the distribution of popularity values is highly skewed. Considering that one of the main goals when using such tasks is to correctly anticipate the cases of highly popular items in order to suggest or promote them, in Chapter 4 the predictive modelling problem of web content popularity prediction is formalized as an imbalanced domain learning task.

Utility-based Evaluation. Also in Chapter 4, the appropriateness of standard evaluation metrics is analysed as to its ability to correctly assess predictive modelling approaches focused on predicting uncommon cases of highly popular web content items. Such analysis shows that the assumption of uniform domain preferences in standard evaluation metrics does not provide the best solution for correctly evaluating approaches for imbalanced domain learning tasks. The concept of utility in regression is introduced and extended. A utility-based evaluation framework is proposed, including a new utility-based evaluation metric, and tested in *a priori* prediction tasks, i.e. predicting the popularity of items before or when they are published.

Popularity Prediction Approaches. Chapter 5 presents several proposals for tackling the imbalanced domain learning task of predicting highly popular web content items. These proposals are based on the concepts of data-level, algorithm-level and hybrid methods, which are focused on tackling the caveats raised by standard learning approaches when applied to imbalanced domain learning tasks. Such proposals are extensively evaluated in several scenarios, and compared to state-of-the-art approaches. Results show that the proposed approaches are capable of significantly improving the predictive ability of models in accurately anticipating the popularity of highly popular items. These results apply to both *a priori* and *a posteriori* tasks.

Single- and Multi-Source Ranking. Finally, the ability of prediction models in generating rankings which are capable of accurately identifying highly relevant items in a timely manner is addressed, in Chapter 6. An extensive evaluation is carried out concerning two types of tasks: single-source and multi-source ranking tasks. Results show that the best overall ranking approaches to both types of tasks concerns the use of the prediction methods proposed in this thesis. In addition, results also show that several rank aggregation methods are capable of obtaining a cumulative advantage effect, i.e. consistently provide better results that any single-source ranking approach.

7.1 Limitations and Future Work

Considering the results provided by the several experimental analysis provided in this thesis, it is possible to conclude that the task of web content popularity prediction is not trivial. In both predictive scenarios (*a priori* and *a posteriori*) results show that, regardless of predictive accuracy improvements, the overall results are still far from optimal. This shows that there is still a great amount of effort required concerning the prediction of highly popular web content items. For tasks with such goal, this thesis provides a starting point.

Nevertheless, as it would be expected, the work presented in this thesis also provides some limitations, and therefore, suggestions for future work.

Concerning imbalanced domain learning tasks, these were formalized as a numeric prediction task, for which previous work is negligible. Therefore, solving several issues concerning the proper evaluation of models' performance, and approaches for tackling imbalanced domains are still required. Furthermore, in this thesis the problems of prediction and ranking of web content popularity are evaluated in an offline environment. Given the growing interest in field such as data streams [130], it would be interesting to explore these problems within such scope.

In Chapter 4 the problem of *a priori* prediction is evaluated and a thorough study is provided concerning the best predictive features to tackle the web content popularity prediction task. As noted in such chapter, such conclusions are solely based on the outcome of the applying linear models as a learning tool. As such, as future work, a comprehensive study using multiple learning tools is required, with the objective of arriving at a conclusion concerning the best predictive features for such tools. Also, other learning algorithms should be evaluated, such as neural networks, as well as ensemble learning approaches, e.g. bagging and boosting methods.

Additionally, concerning predictive features, the study presented in this thesis points to features based on sentiment analysis providing an important predictive contribution. However, some questions may be raised for future endeavours. These are mainly related to

the static nature of sentiment lexicons, i.e. a fixed polarity strength value is attributed to sentiment words. Given the dynamic of web content data, it would be interesting to explore the possibility of inferring such polarity strength values from the feedback of users in social media platforms. This could lead to more precise sentiment scores of news items' titles and headlines and, ultimately, to a better contribution to the predictive accuracy of models. In addition, concerning the case of online news feeds, the impact of using the full-text of the items should be studied, including the evaluation of other representations of text (e.g. distributional semantic vectors [98]) and sentiment analysis approaches (i.e. model-based instead of lexicon-based sentiment analysis).

Regarding the evaluation of imbalanced domain learning tasks, further work is mostly needed regarding proper evaluation metrics, and the derivation of utility surfaces, i.e. continuous versions of cost-matrices [72]. As for strategies to tackle imbalanced domain learning tasks, this thesis presents proposals for data-level, algorithm-level and hybrid methods. Nevertheless, the motivation for applying such methods in each of the described scenarios is somewhat naïve: an attempt to match the strengths of each method to the specific prediction problem. Results show that these are capable of providing better results. However, these are solely a depiction of the predictive potential that could be provided by such methods. Concretely, many proposals can be explored by using hybrid methods. In this thesis, a solution proposed for such methods is based on combining two models of different characteristics. However, many other solutions can be explored. For example, the exploration of diversity, by combining models with different sampling percentages, e.g. [76].

Concerning the ability of the proposed predictive modelling approaches in generating accurate and timely rankings, results show that it is possible to improve such ability. However, the most noticeable part of such study concerns the multi-source ranking task. Results show how prediction models for each source of popularity observations regarding a set of items, i.e. social media sources, can be combined in order to provide an increased ability in timely suggestions of highly popular items, according to multiple sources. However, the implications of heavy-tail distributions should be studied in order to assess if such results could be further improved. Furthermore, the performance of the methods proposed in this thesis should be further evaluated by means of a human evaluation. This would provide increased confidence in the improved ability of the proposed approaches.

Annexes

A Evaluation Results of Feature Sets

Table 1: Evaluation results concerning the evaluation metrics $rmse$, $rmse_\phi$ and F_1^u, for all combinations of social media sources and news topics in *a priori* prediction tasks, regarding models with different sets of features.

		Economy			Microsoft			Obama			Palestine		
	Features	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u
Twitter	C	125.349	258.309	0.207	174.216	371.341	0.174	321.497	705.956	0.059	156.136	371.269	0.218
	T	125.540	261.475	0.000	174.406	377.722	0.000	322.252	710.022	0.000	156.668	376.245	0.000
	M	120.365	228.714	**0.376**	**128.910**	272.707	**0.581**	296.735	644.208	**0.313**	137.659	325.991	**0.444**
	CT	125.315	258.248	0.209	174.161	371.385	0.179	321.445	705.819	0.061	156.016	370.909	0.226
	CM	120.819	228.684	0.373	129.680	272.826	0.572	296.840	643.914	0.311	138.219	325.809	0.431
	TM	**120.363**	228.708	**0.376**	128.913	**272.679**	0.578	**296.648**	644.048	0.311	**137.657**	325.976	0.442
	CTM	120.819	**228.681**	0.372	129.683	272.808	0.578	296.758	**643.769**	0.311	138.210	**325.767**	0.431
Facebook	C	352.783	1049.817	0.121	411.477	1331.273	0.104	728.522	2321.233	0.012	289.507	880.689	0.120
	T	**345.760**	1058.186	0.000	**406.280**	1333.496	0.000	**720.947**	2333.649	0.000	**275.944**	891.975	0.000
	M	346.759	1038.391	0.246	439.445	1320.443	**0.330**	764.006	2240.831	0.156	317.719	886.859	**0.285**
	CT	352.797	1049.855	0.120	411.452	1331.132	0.105	728.623	2321.226	0.012	289.419	**880.237**	0.121
	CM	346.657	**1033.100**	0.224	443.701	1319.528	0.232	766.677	2235.903	**0.165**	322.675	882.358	0.265
	TM	346.841	1038.577	**0.247**	439.471	1320.431	0.322	764.115	2240.856	0.156	317.797	886.801	0.284
	CTM	346.742	1033.284	0.223	443.708	**1319.493**	0.235	766.708	**2235.620**	**0.165**	322.730	882.336	0.265
Google+	C	10.175	36.920	0.034	26.578	107.348	0.086	21.237	79.389	0.006	14.425	50.525	0.021
	T	**9.482**	36.678	0.000	**25.085**	108.885	0.000	**20.635**	79.392	0.000	**11.572**	52.449	0.000
	M	10.662	35.423	0.177	26.558	**101.498**	**0.237**	22.029	77.322	0.105	15.239	48.524	**0.099**
	CT	10.179	36.926	0.033	26.581	107.375	0.085	21.238	79.382	0.006	14.433	50.556	0.021
	CM	10.773	35.481	**0.186**	27.752	101.818	0.209	22.190	77.378	**0.135**	16.492	**48.247**	0.098
	TM	10.660	**35.413**	0.177	26.568	101.560	**0.237**	22.032	**77.320**	0.107	15.242	48.542	**0.099**
	CTM	10.770	35.472	**0.186**	27.754	101.862	0.209	22.192	77.374	0.132	16.495	48.266	0.098
LinkedIn	C	120.205	372.956	0.058	371.987	1195.104	0.100	82.255	231.651	0.086	59.863	100.127	0.117
	T	118.827	375.236	0.000	371.288	1200.371	0.000	80.679	232.456	0.000	**28.981**	**98.028**	0.000
	M	**108.418**	326.308	0.189	**363.499**	**1166.255**	0.239	**80.303**	225.443	0.188	33.403	102.454	0.184
	CT	120.225	372.978	0.058	372.009	1195.077	0.101	82.248	231.615	0.086	59.876	100.127	0.116
	CM	108.593	**324.575**	**0.192**	364.641	1167.213	0.246	80.599	**224.519**	**0.195**	46.930	102.554	0.142
	TM	108.425	326.315	0.189	363.539	1166.333	0.240	**80.303**	225.445	0.186	33.436	102.485	**0.185**
	CTM	108.604	324.578	0.191	364.661	1167.226	**0.251**	80.600	224.524	**0.195**	46.950	102.550	0.141

B Evaluation Results of Best Feature Set and Sentiment Scores

Table 2: Evaluation results concerning the evaluation metrics $rmse$, $rmse_\phi$ and F_1^u, for all combinations of social media sources and news topics in *a priori* prediction tasks, regarding models with meta-data features and sentiment scores.

		Economy			Microsoft			Obama			Palestine		
	Features	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u
Twitter	M	120.365	228.714	0.376	128.910	272.707	0.581	296.735	644.208	0.313	137.659	325.991	0.444
	MS (Baseline)	110.144	221.697	**0.424**	128.702	272.302	0.585	290.036	626.010	0.338	137.631	324.590	**0.450**
	MS (AFINN)	**110.095**	**221.599**	**0.424**	**128.678**	**272.222**	0.585	**289.889**	**625.612**	0.337	137.626	324.519	0.448
	MS (SentiStrength)	110.169	221.731	0.423	128.710	272.320	**0.586**	289.974	625.830	0.336	**137.480**	**324.059**	0.448
	MS (SWN)	110.151	221.688	**0.424**	128.702	272.302	0.585	290.035	625.783	**0.339**	137.631	324.590	**0.450**
Facebook	M	346.759	1038.391	0.246	439.445	1320.443	0.330	764.006	2240.831	0.156	**317.719**	**886.859**	0.285
	MS (Baseline)	303.711	737.001	0.307	391.701	948.464	**0.355**	**660.035**	**1662.589**	0.320	589.332	1513.117	0.308
	MS (AFINN)	303.765	736.975	0.307	**391.689**	**948.421**	**0.355**	660.076	1662.694	**0.324**	589.207	1512.749	**0.312**
	MS (SentiStrength)	303.741	736.879	0.307	391.817	948.632	0.346	660.052	1662.633	0.318	589.268	1511.711	0.306
	MS (SWN)	**303.670**	**736.811**	**0.308**	391.693	948.484	0.349	660.036	1662.598	0.322	589.047	1511.726	0.303
Google+	M	10.662	35.423	0.177	26.558	101.498	0.237	22.029	77.322	0.105	**15.239**	**48.524**	0.099
	MS (Baseline)	5.816	12.333	0.337	**18.624**	46.715	0.361	18.071	36.932	**0.386**	26.745	76.550	0.311
	MS (AFINN)	**5.815**	12.329	0.337	18.626	**46.689**	0.361	**18.068**	**36.921**	0.384	26.752	76.571	**0.319**
	MS (SentiStrength)	5.816	**12.328**	0.336	18.628	46.708	0.355	18.069	36.925	0.384	26.749	76.571	0.314
	MS (SWN)	**5.815**	12.331	**0.338**	18.628	46.719	**0.362**	**18.068**	36.925	**0.386**	26.750	76.591	0.316
LinkedIn	M	108.418	326.308	0.189	363.499	1166.255	0.239	**80.303**	**225.443**	0.188	**33.403**	**102.454**	0.184
	MS (Baseline)	75.231	164.683	0.335	301.541	**744.935**	0.361	103.702	227.002	0.376	104.202	236.683	0.354
	MS (AFINN)	**75.203**	**164.606**	0.335	301.544	744.990	0.362	103.700	226.999	0.376	104.195	236.639	0.354
	MS (SentiStrength)	75.220	164.663	**0.336**	**301.539**	744.971	0.361	103.705	227.008	0.374	104.211	236.690	**0.355**
	MS (SWN)	75.241	164.729	0.335	301.554	745.031	**0.366**	103.690	226.981	**0.378**	104.193	236.672	0.354

C Optimal Parametrization for Learning Algorithms in A Priori Experiments

Table 3: Results concerning the application of the optimal parametrization method in *a priori* prediction tasks based on each combination of news topics and social media source's data, concerning parameters *cost* and *gamma* for SVM's, *nk*, *degree* and *thresh* for MARS's, and *ntree* for Random Forests, with the addition of under-sampling and/or over-sampling percentages where applicable.

		Economy				Microsoft			
	Strategy	**lm**	**mars**	**svm**	**rf**	**lm**	**mars**	**svm**	**rf**
Twitter	Original	x	10,1,0.01	10,0.001	5,750	x	10,2,0.001	10,0.01	5,750
	U	0.7	17,1,0.001,0.2	300,0.01,0.1	7,500,0.5	0.1	10,1,0.01,0.1	150,0.01,0.1	7,1500,0.8
	O	2	17,1,0.001,1.5	10,0.01,2	5,750,1.1	1.5	17,1,0.001,1.1	150,0.01,2	7,750,1.5
	SM	0.8,2	17,1,0.001,0.5,2	300,0.01,0.2,1.5	7,750,0.8,1.5	0.1,2	10,20.01,0.2,2	300,0.01,0.1,2	7,500,0.2,1.1
	U_T	0.7	17,1,0.001,0.5	300,0.01,0.2	5,750,0.8	0.5	10,1,0.001,0.5	150,0.01,0.2	5,1500,0.7
	O_T	1.5	17,1,0.001,2	300,0.001,2	5,500,1.1	2	17,1,0.001,2	300,0.01,2	5,750,1.1
	SM_T	0.8,1.1	17,2,0.01,0.8,2	150,0.01,0.2,1.1	7,750,0.6,1.5	0.1,1.1	10,2,0.01,0.7,2	10,0.01,0.2,2	7,500,0.5,1.5
	U_TPhi	0.7	17,1,0.001,0.2	10,0.01,0.1	7,750,0.2	0.2	17,2,0.01,0.1	150,0.01,0.1	7,500,0.1
	O_TPhi	1.5	17,2,0.01,2	300,0.001,2	7,750,1.1	2	10,1,0.001,1,5	300,0.01,2	5,750,1.5
	SM_TPhi	0.8,1.5	17,2,0.001,0.7,2	10,0.01,0.2,1,5	7,500,0.5,1.1	0.1,1.1	17,2,0.01,0.2,1.5	150,0.001,0.1,2	7,1500,0.8,1.1
Facebook	Original	x	17,2,0.001	10,0.001	500	x	10,1,0.01	10,0.001	750
	U	0.9	17,2,0.001,0.9	10,0.001,0.9	500,0.9	0.9	10,2,0.001,0.9	150,0.001,0.1	500,0.2
	O	1.1	17,1,0.001,1.1	10,0.001,1.5	500,2	1.1	17,1,0.01,1.5	10,0.001,1.1	1500,2
	SM	0.9,1.1	17,2,0.001,0.9,1.1	10,0.001,0.7,1.5	500,0.7,1.1	0.9,1.1	17,2,0.001,0.9,1.5	10,0.001,0.9,1.1	500,0.1,1.5
	U_T	0.9	17,2,0.001,0.9	10,0.001,0.9	500,0.8	0.9	17,2,0.01,0.9	10,0.001,0.8	1500,0.7
	O_T	1.1	17,2,0.001,1.1	10,0.001,2	750,1.5	1.1	17,2,0.01,1.1	10,0.001,1.1	500,1.1
	SM_T	0.9,1.1	17,2,0.001,0.9,1.1	10,0.001,0.7,1.5	500,0.9,1.1	0.9,1.1	17,2,0.001,0.9,1.5	10,0.001,0.9,1.1	500,0.5,2
	U_TPhi	0.9	10,1,0.01,0.7	10,0.001,0.9	500,0.7	0.9	10,2,0.001,0.9	10,0.001,0.9	500,0.8
	O_TPhi	1.1	17,2,0.001,1.5	10,0.001,1.1	500,1.5	1.1	10,1,0.01,1.1	10,0.001,1.1	1500,2
	SM_TPhi	0.9,1.1	17,2,0.001,0.7,1.1	10,0.001,0.9,2	750,0.2,2	0.9,1.1	17,2,0.01,0.9,1.1	10,0.001,0.9,1.1	500,0.1,1.1
Google+	Original	x	10,1,0.001	150,0.001	500	x	10,1,0.01	10,0.001	750
	U	0.8	17,1,0.001,0.7	10,0.001,0.2	1500,0.5	0.9	10,2,0.001,0.9	150,0.001,0.1	1500,0.1
	O	1.1	10,1,0.01,1.1	10,0.001,2	750,1.1	1.1	10,2,0.01,1.1	10,0.001,2	1500,1.1
	SM	0.8,1.1	17,1,0.001,0.7,1.1	10,0.001,0.6,2	750,0.7,1.1	0.9,1.1	17,1,0.01,0.9,1.1	10,0.001,0.9,2	1500,0.8,2
	U_T	0.8	17,1,0.001,0.6	300,0.001,0.2	500,0.7	0.9	10,2,0.001,0.9	150,0.001,0.1	1500,0.6
	O_T	2	17,2,0.001,2	10,0.001	1500,1.5	1.1	10,2,0.01,1.1	10,0.001,1.5	500,1.1
	SM_T	0.8,1.1	17,1,0.01,0.5,1.1	10,0.001,0.6,2	1500,0.7,1.1	0.9,1.1	10,2,0.01,0.9,1.1	10,0.001,0.8,1.5	500,0.9,1.5
	U_TPhi	0.8	17,1,0.001,0.7	150,0.001,0.1	750,0.7	0.9	17,2,0.001,0.8	10,0.001,0.9	1500,0.2
	O_TPhi	1.5	10,2,0.001,2	10,0.001,2	750,1.5	1.1	10,2,0.001,1.1	10,0.001,2	750,1.5
	SM_TPhi	0.9,1.5	17,1,0.001,0.7,1.1	10,0.001,0.7,2	1500,0.9,2	0.9,1.1	17,1,0.01,0.9,1.1	10,0.001,0.9,1.5	1500,0.1,1.5
LinkedIn	Original	x	10,2,0.001	10,0.001	1500	x	10,1,0.001	150,0.001	750
	U	0.5	10,1,0.001,0.5	10,0.001,0.2	1500,0.7	0.6	17,2,0.001,0.6	10,0.001,0.5	500,0.9
	O	1.1	17,1,0.001,1.5	10,0.001,2	1500,2	1.1	10,1,0.01,1.1	10,0.001,2	500,1.5
	SM	0.8,1.1	10,2,0.001,0.6,1.5	10,0.001,0.6,2	500,0.8,1.5	0.9,2	17,2,0.001,0.5,1.1	10,0.001,0.6,2	500,0.8,1.1
	U_T	0.7	10,1,0.01,0.5	10,0.001,0.2	1500,0.8	0.9	17,1,0.001,0.5	10,0.001,0.5	500,0.7
	O_T	1.1	17,1,0.01,2	10,0.001,2	500,1.5	1.5	17,2,0.01,2	10,0.001,2	500,1.5
	SM_T	0.8,1.1	17,1,0.001,0.6,1.5	10,0.001,0.8,2	750,0.7,1.1	0.9,1.1	17,1,0.001,0.5,1.1	10,0.001,0.5,1.1	500,0.6,1.5
	U_TPhi	0.7	10,2,0.001,0.5	10,0.001,0.5	1500,0.9	0.6	10,1,0.01,0.5	10,0.001,0.5	500,0.5
	O_TPhi	1.1	10,1,0.01,2	10,0.001,2	1500,1.5	1.5	17,2,0.001,2	10,0.001,2	1500,2
	SM_TPhi	0.8,1.5	10,2,0.001,0.6,1.1	10,0.001,0.6,2	1500,0.5,1.1	0.8,1.5	10,1,0.001,0.8,2	10,0.001,0.6,1.5	1500,0.6,2

Continued on the next page

Table 3 – continued from the previous page

		Obama				Palestine			
	Strategy	**lm**	**mars**	**svm**	**rf**	**lm**	**mars**	**svm**	**rf**
Twitter	Original	x	17,1,0.001	150,0.01	7,500	x	10,2,0.01	10,0.001	5,500
	U	0.2	17,1,0.01,0.2	300,0.01,0.1	5,750,0.9	0.6	10,2,0.01,0.6	300,0.01,0.1	7,500,0.1
	O	2	10,1,0.01,2	150,0.001,2	5,750,1.5	1.1	17,2,0.001,2	300,0.01,2	7,750,1.5
	SM	0.2,1.1	17,1,0.01,0.5,2	150,0.01,0.1,1.5	7,500,0.5,1.1	0.8,2	10,2,0.001,0.5,1.5	10,0.01,0.5,2	5,750,0.2,1.1
	U_T	0.2	10,1,0.01,0.2	150,0.01,0.1	7,500,0.5	0.2	17,2,0.01,0.6	300,0.001,0.2	5,500,0.8
	O_T	2	17,1,0.001,2	300,0.01,2	5,500,2	2	10,2,0.001,1.5	150,0.001,2	5,750,1.1
	SM_T	0.2,1.1	17,1,0.01,0.6,2	300,0.01,0.2,2	5,500,0.9,1.5	0.2,1.1	10,2,0.001,0.6,1.5	10,0.01,0.5,2	5,500,0.8,2
	U_TPhi	0.2	17,1,0.01,0.2	300,0.01,0.1	5,500,0.1	0.9	10,2,0.01,0.1	300,0.01,0.2	7,1500,0.1
	O_TPhi	2	17,1,0.001,2	300,0.01,2	7,750,1.1	1.1	17,2,0.01,2	150,0.001,2	7,750,1.1
	SM_TPhi	0.2,1.1	10,1,0.01,0.5,2	300,0.01,0.2,2	5,1500,0.9,1.1	0.6,1.5	17,1,0.01,0.7,2	150,0.001,0.5,2	7,500,0.8,2
Facebook	Original	x	10,1,0.01	10,0.001	1500	x	10,2,0.01	150,0.001	1500
	U	0.5	17,1,0.01,0.5	150,0.001,0.5	500,0.9	0.9	17,1,0.01,0.6	150,0.001,0.5	1500,0.6
	O	1.5	17,2,0.01,1.5	300,0.01,2	1500,1.1	1.1	10,1,0.01,1.5	300,0.01,1.5	1500,1.1
	SM	0.9,2	17,1,0.01,0.7,1.5	10,0.001,0.9,2	1500,0.9,1.5	0.9,1.1	17,2,0.001,0.6,1.1	150,0.001,0.6,1.5	500,0.2,1.1
	U_T	0.5	10,1,0.01,0.5	300,0.01,0.5	1500,0.1	0.9	17,2,0.01,0.6	150,0.001,0.5	750,0.8
	O_T	1.5	17,1,0.01,2	300,0.001,2	1500,1.5	1.5	17,2,0.001,1.1	150,0.001,1.5	500,1.5
	SM_T	0.5,1.1	10,1,0.001,0.6,1.5	150,0.01,0.8,2	500,0.2,1.1	0.9,1.1	17,1,0.01,0.7,1.1	150,0.001,0.5,1.5	500,0.7,2
	U_TPhi	0.5	17,1,0.01,0.5	10,0.001,0.6	500,0.2	0.9	17,2,0.001,0.6	150,0.001,0.2	750,0.1
	O_TPhi	2	17,1,0.01,2	10,0.001,1.5	750,1.5	1.5	17,1,0.01,2	150,0.001,2	1500,1.1
	SM_TPhi	0.7,1.5	17,1,0.001,0.7,2	300,0.001,0.8,2	1500,0.2,1.5	0.9,1.1	10,1,0.001,0.7,1.5	300,0.01,0.9,2	500,0.1,1.5
Google+	Original	x	17,1,0.001	300,0.01	750	x	17,1,0.01	10,0.001	750
	U	0.8	17,2,0.01,0.2	10,0.001,0.1	750,0.8	0.8	10,2,0.01,0.1	300,0.001,0.2	750,0.2
	O	1.1	17,1,0.01,2	10,0.001,2	500,1.5	1.1	17,1,0.01,1.5	150,0.001,2	1500,1.1
	SM	0.9,1.1	17,1,0.01,0.6,2	10,0.001,0.1,1.1	500,0.6,1.5	0.6,1.5	10,2,0.01,0.1,2	150,0.001,0.5,1.5	1500,0.8,1.1
	U_T	0.8	17,1,0.001,0.5	10,0.001,0.1	750,0.1	0.6	17,2,0.001,0.2	150,0.01,0.1	1500,0.9
	O_T	1.5	17,1,0.01,2	300,0.001,2	750,2	2	17,2,0.001,2	150,0.01,2	750,2
	SM_T	0.9,1.1	17,1,0.001,0.9,2	10,0.001,0.1,1.1	500,0.6,1.1	0.5,2	10,2,0.01,0.2,2	150,0.01,0.2,1.5	750,0.9,2
	U_TPhi	0.7	17,1,0.001,0.5	10,0.001,0.2	750,0.1	0.7	10,2,0.01,0.6	150,0.01,0.1	750,0.1
	O_TPhi	1.5	10,1,0.01,2	150,0.001,2	1500,1.5	1.5	10,1,0.01,2	300,0.01,1.5	1500,1.5
	SM_TPhi	0.9,1.1	17,1,0.001,0.9,2	150,0.001,0.1,1.1	500,0.6,2	0.6,1.5	10,2,0.01,0.2,2	10,0.01,0.2,2	750,0.8,1.1
LinkedIn	Original	x	10,1,0.01	300,0.01	750	x	17,2,0.01	300,0.01	1500
	U	0.6	17,2,0.01,0.6	10,0.001,0.1	500,0.2	0.9	10,2,0.01,0.7	10,0.01,0.1	500,0.2
	O	1.1	10,2,0.001,2	150,0.01,2	500,2	1.5	10,1,0.001,2	10,0.01,1.5	750,1.1
	SM	0.9,1.5	10,1,0.001,0.5,2	150,0.001,0.1,1.1	500,0.2,1.5	0.8,1.5	10,2,0.01,0.7,1.1	10,0.01,0.6,1.5	750,0.1,1.5
	U_T	0.7	10,1,0.01,0.6	150,0.001,0.1	500,0.7	0.8	10,2,0.01,0.7	10,0.01,0.2	750,0.9
	O_T	1.5	17,1,0.01,2	300,0.01,2	1500,2	1.5	17,1,0.01,2	300,0.01,1.5	500,1.5
	SM_T	0.6,1.1	17,1,0.01,0.8,2	300,0.001,0.1,1.1	500,0.6,1.1	0.9,2	10,2,0.01,0.6,1.1	10,0.01,0.5,1.5	500,0.9,1.5
	U_TPhi	0.8	10,2,0.001,0.6	300,0.001,0.1	500,0.2	0.9	17,2,0.01,0.9	10,0.01,0.2	500,0.9
	O_TPhi	2	10,1,0.001,2	300,0.01,2	750,1.1	1.1	17,1,0.001,1.5	10,0.001,1.5	500,2
	SM_TPhi	0.9,1.1	10,2,0.001,0.7,1.1	10,0.001,0.1,2	500,0.2,2	0.9,1.1	17,2,0.01,0.5,1.1	10,0.01,0.5,1.5	500,0.7,1.5

D Evaluation Results of A Priori Prediction

Table 4: Evaluation results concerning the evaluation metrics $rmse$, $rmse_\phi$ and F_1^u, for all combinations of social media sources and news topics in *a priori* prediction tasks, regarding several regression algorithms and the proposed data-level methods.

			Economy			**Microsoft**			**Obama**			**Palestine**		
		Approach	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u
Twitter	**lm**	Original	**131.913**	314.052	0.323	**127.999**	339.848	0.502	311.500	860.664	0.239	**150.719**	433.922	0.404
		U	134.620	313.186	0.326	130.190	330.890	0.530	311.592	820.409	0.302	150.776	431.814	0.405
		O	139.297	311.845	0.319	128.175	336.708	0.511	309.771	832.106	0.290	150.948	431.296	**0.406**
		SM	135.726	**310.308**	**0.329**	132.725	327.772	**0.538**	312.329	818.244	0.303	150.886	430.867	0.405
		U_T	133.707	312.090	0.328	128.296	336.741	0.516	311.078	811.920	**0.305**	151.231	**426.340**	0.399
		O_T	139.819	315.742	0.314	128.017	337.579	0.508	**309.761**	844.144	0.273	151.896	434.690	**0.406**
		SM_T	133.587	312.240	**0.329**	132.283	328.164	0.535	311.758	**808.704**	0.304	151.482	426.426	0.396
		U_TPhi	134.699	314.030	0.323	129.047	332.987	0.524	315.193	816.752	0.299	150.722	433.630	0.403
		O_TPhi	136.817	313.766	0.323	128.013	337.463	0.509	309.866	844.280	0.273	151.981	437.636	0.402
		SM_TPhi	136.116	312.114	0.325	133.629	**326.949**	0.532	316.259	813.527	0.298	150.784	431.225	0.404
	mars	Original	126.192	309.113	0.348	**127.538**	337.789	0.513	310.025	858.127	0.237	**151.281**	438.703	0.402
		U	132.368	**297.629**	0.373	130.280	328.089	0.535	309.962	809.870	0.315	152.681	439.003	0.404
		O	129.174	302.059	0.371	128.451	335.527	0.532	**308.002**	824.126	0.316	157.672	451.079	0.398
		SM	132.128	298.165	**0.374**	130.465	328.614	0.537	309.090	816.290	0.317	153.040	434.893	**0.410**
		U_T	129.626	302.512	0.369	128.418	333.480	0.537	310.416	799.823	0.316	153.880	442.508	0.404
		O_T	133.031	303.543	0.373	128.532	333.618	0.541	308.525	836.375	0.307	152.768	442.508	0.403
		SM_T	124.505	302.291	0.369	129.280	330.512	**0.538**	308.117	814.272	**0.323**	154.319	441.557	0.408
		U_TPhi	131.525	298.217	0.372	130.789	**327.864**	0.535	311.327	**809.135**	0.308	154.788	**435.674**	0.396
		O_TPhi	**123.813**	305.036	0.367	128.226	336.637	0.529	308.411	836.154	0.311	152.698	436.856	0.405
		SM_TPhi	135.078	302.932	0.371	129.800	329.457	0.535	309.037	813.410	0.322	149.573	427.103	**0.410**
	svm	Original	126.561	329.303	0.241	129.042	350.945	0.456	322.952	911.970	0.140	151.463	443.457	0.393
		U	134.079	306.141	0.379	128.563	337.540	0.518	311.248	835.321	0.304	151.567	**431.206**	0.404
		O	**125.819**	317.073	0.330	128.243	343.552	0.485	313.544	872.436	0.218	154.992	443.379	0.404
		SM	131.936	306.526	0.378	129.473	333.802	0.529	312.293	821.983	0.314	150.380	435.110	**0.411**
		U_T	130.934	306.403	**0.383**	**128.096**	335.868	0.528	310.331	820.568	0.318	150.220	431.294	0.410
		O_T	127.527	320.159	0.314	128.341	345.027	0.482	316.079	883.215	0.178	151.068	438.557	0.406
		SM_T	131.316	305.110	0.378	128.618	**333.037**	**0.532**	**310.266**	**813.899**	**0.320**	**149.982**	431.592	0.407
		U_TPhi	132.731	**303.151**	0.373	129.600	335.417	0.518	312.377	825.348	0.309	151.552	434.172	0.403
		O_TPhi	128.781	320.287	0.314	128.412	345.350	0.485	316.057	882.702	0.177	150.688	437.900	0.404
		SM_TPhi	137.012	304.686	0.375	131.898	331.081	0.529	310.853	820.746	0.315	150.339	434.237	0.409
	rf	Original	**128.232**	316.704	0.333	**133.343**	**351.424**	0.491	321.700	873.604	0.241	152.831	442.788	**0.355**
		U	128.506	316.856	0.333	133.414	351.534	0.490	321.629	873.442	0.239	152.803	442.769	0.351
		O	128.426	316.947	0.332	133.452	351.627	0.490	321.868	874.092	0.239	152.800	442.914	0.351
		SM	128.358	316.874	**0.337**	133.439	351.640	0.487	321.822	873.786	0.238	152.875	443.167	0.353
		U_T	128.292	316.692	0.334	133.415	351.568	0.492	321.756	873.570	0.240	152.874	443.046	0.352
		O_T	128.308	316.962	0.331	133.385	351.531	0.486	321.496	873.105	0.240	152.908	443.105	0.353
		SM_T	128.355	316.916	0.333	133.330	351.460	0.488	**321.294**	**872.675**	**0.242**	152.973	443.366	**0.355**
		U_TPhi	128.379	316.907	0.332	133.435	351.522	0.493	321.595	873.583	0.240	152.870	442.942	0.352
		O_TPhi	128.262	**316.675**	0.334	133.385	351.448	0.488	321.597	873.440	0.241	152.934	443.104	0.351
		SM_TPhi	128.337	316.883	0.331	133.368	351.497	**0.494**	321.567	873.384	0.240	**152.744**	**442.666**	0.350
		Bandari	**128.923**	337.282	0.213	**131.978**	360.814	0.444	324.279	912.643	0.129	**151.417**	444.768	0.346

Continued on the next page

Table 4 – continued from the previous page

			Economy			Microsoft			Obama			Palestine		
		Approach	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u
Facebook	lm	Original	**300.764**	882.266	**0.218**	**453.783**	1378.938	0.241	**754.767**	2209.789	0.159	**324.116**	**905.497**	**0.259**
		U	302.270	881.736	0.214	456.513	1379.666	0.235	790.584	2130.905	0.184	326.458	906.297	0.254
		O	309.281	**874.864**	0.195	477.781	1388.157	0.211	804.590	**2109.642**	0.185	339.674	908.311	0.226
		SM	302.474	879.559	0.213	457.968	1380.119	0.231	794.956	2116.163	0.188	327.680	906.266	0.252
		U_T	301.301	880.910	0.217	455.061	**1378.906**	0.237	793.969	2128.887	0.188	327.019	905.971	0.254
		O_T	304.636	881.952	0.207	499.262	1399.496	0.224	784.289	2175.396	0.168	346.860	916.169	0.229
		SM_T	301.988	879.792	0.213	460.667	1379.926	0.233	801.859	2116.973	**0.189**	331.385	910.183	0.253
		U_TPhi	302.148	881.054	0.216	455.278	1378.919	0.234	791.567	2132.389	0.182	326.022	906.012	0.257
		O_TPhi	306.478	883.915	0.204	495.459	1400.538	0.226	804.010	2140.885	0.183	340.068	912.768	0.227
		SM_TPhi	302.630	880.003	0.211	465.481	1380.976	0.231	796.481	2123.296	0.184	327.746	905.524	0.250
	mars	Original	296.762	883.065	**0.264**	**451.423**	1379.953	**0.328**	**756.740**	2193.037	0.163	**303.041**	**892.726**	0.282
		U	297.172	885.793	0.260	725.500	**1372.088**	0.321	786.356	2107.161	0.197	333.124	925.651	0.278
		O	303.540	879.815	0.245	483.138	1391.301	0.310	896.966	2109.794	0.200	331.064	926.694	0.281
		SM	**295.761**	883.909	0.260	656.487	1381.085	0.317	790.229	2099.433	0.198	314.217	903.158	0.275
		U_T	297.022	882.880	0.260	746.641	1379.522	0.326	793.190	2097.146	0.200	318.796	897.967	0.284
		O_T	297.037	890.457	0.262	594.711	1376.383	0.320	794.851	2118.304	0.194	366.836	938.667	0.269
		SM_T	296.703	883.519	0.260	704.741	1385.280	0.316	808.018	**2071.317**	**0.207**	336.467	930.947	**0.287**
		U_TPhi	300.186	**877.760**	0.255	623.283	1381.933	0.317	784.095	2112.352	0.195	342.312	971.720	0.278
		O_TPhi	298.596	880.237	0.255	482.395	1381.569	0.321	798.056	2116.912	0.194	321.374	918.578	0.286
		SM_TPhi	297.058	880.203	0.254	838.139	1383.064	0.319	798.174	2084.353	**0.207**	337.047	934.479	0.277
	svm	Original	**292.759**	886.163	**0.215**	**424.859**	1381.226	**0.282**	736.497	2370.984	0.034	273.085	901.226	0.227
		U	293.282	884.299	0.213	494.675	1385.285	0.240	795.796	2300.716	0.090	273.194	890.162	0.261
		O	301.134	**879.078**	0.194	428.361	**1375.895**	0.264	901.124	**2262.489**	**0.126**	274.328	878.977	0.258
		SM	296.792	879.806	0.201	425.183	1379.886	0.279	751.811	2320.722	0.070	274.146	881.960	0.267
		U_T	293.126	884.385	0.212	425.236	1379.077	0.270	856.955	2289.757	0.113	**272.178**	883.054	**0.281**
		O_T	301.459	881.129	0.199	427.077	1378.833	0.258	740.088	2374.103	0.037	272.933	893.189	0.251
		SM_T	298.159	879.288	0.199	425.000	1378.584	0.276	851.738	2293.877	0.116	273.970	**869.210**	0.279
		U_TPhi	293.551	884.217	0.210	424.922	1380.174	0.277	841.244	2292.430	0.105	276.294	880.143	0.258
		O_TPhi	295.331	887.036	0.214	426.588	1378.077	0.262	**734.323**	2371.684	0.054	274.261	889.533	0.265
		SM_TPhi	298.032	880.809	0.203	424.866	1378.010	0.274	843.426	2294.991	0.121	273.275	881.136	0.265
	rf	Original	297.991	879.563	0.245	440.429	1391.415	**0.332**	717.424	2197.356	0.170	292.294	903.161	0.253
		U	297.550	879.224	0.244	439.916	1390.991	0.328	717.571	2197.205	**0.174**	292.523	903.430	0.253
		O	297.258	879.210	0.243	439.537	1390.731	0.331	717.487	2197.453	0.170	292.752	903.209	0.252
		SM	297.458	879.221	0.243	439.557	**1390.599**	0.327	717.463	2197.333	0.171	292.852	904.081	0.252
		U_T	297.650	879.234	0.245	**439.271**	1391.085	0.329	717.464	2197.333	0.172	292.430	903.358	**0.259**
		O_T	297.349	**878.923**	0.243	440.171	1390.696	0.327	**717.396**	**2197.168**	0.173	292.855	903.448	0.253
		SM_T	297.514	879.393	**0.246**	439.915	1391.140	0.326	717.553	2197.584	0.169	292.866	903.926	0.255
		U_TPhi	297.837	879.569	0.244	440.668	1391.886	0.327	717.465	2197.189	0.172	292.568	903.380	0.256
		O_TPhi	297.517	879.055	**0.246**	439.821	1391.041	0.326	717.524	2197.465	0.173	**292.267**	**903.039**	0.253
		SM_TPhi	**297.119**	879.184	**0.246**	439.795	1390.907	0.329	717.459	2197.455	0.171	292.691	903.769	0.252
		Bandari	**293.287**	892.522	0.211	**425.485**	**1385.502**	**0.309**	**741.610**	2371.126	0.053	**279.986**	907.269	0.194

Continued on the next page

Table 4 – continued from the previous page

			Economy			Microsoft			Obama			Palestine		
		Approach	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u
Google+	lm	Original	**10.722**	36.443	0.183	**29.645**	114.651	**0.206**	**22.172**	79.331	0.133	**15.410**	43.765	**0.133**
		U	10.890	36.194	**0.187**	30.022	114.579	0.199	22.731	78.636	0.138	16.346	43.638	0.110
		O	11.371	**35.713**	0.177	32.470	**114.182**	0.168	24.302	**77.027**	0.137	17.540	43.189	0.099
		SM	10.975	36.128	0.183	30.304	114.499	0.195	22.656	78.682	0.141	18.816	43.298	0.085
		U_T	10.896	36.162	0.185	30.010	114.570	0.198	22.711	78.676	**0.144**	18.115	43.197	0.095
		O_T	11.499	35.892	0.170	31.819	115.336	0.174	23.698	78.274	0.130	17.825	44.522	0.105
		SM_T	10.945	36.060	0.185	30.275	114.456	0.195	22.627	78.756	0.139	21.008	**43.160**	0.082
		U_TPhi	10.891	36.218	0.184	29.991	114.583	0.198	23.243	77.972	0.141	17.132	43.243	0.101
		O_TPhi	11.073	36.043	0.171	31.762	115.556	0.178	23.750	78.264	0.128	17.596	44.369	0.105
		SM_TPhi	11.015	35.946	0.186	30.310	114.551	0.195	22.723	78.709	0.137	18.811	43.497	0.089
	mars	Original	**9.769**	35.773	0.224	**30.065**	114.475	0.234	**21.842**	77.815	0.113	**14.094**	44.200	0.102
		U	10.014	35.071	0.234	38.017	116.545	**0.235**	28.607	**71.225**	0.143	18.639	38.756	0.125
		O	10.437	**34.732**	0.231	36.371	120.066	0.216	25.233	72.399	0.149	16.300	42.132	**0.141**
		SM	10.104	34.876	0.232	31.617	**114.071**	0.230	25.421	72.112	0.148	20.331	**36.178**	0.115
		U_T	10.190	34.755	**0.235**	35.931	116.216	0.232	24.046	74.332	0.149	17.896	40.933	0.122
		O_T	12.301	44.819	0.232	32.801	116.105	0.233	23.598	75.007	0.140	16.219	59.834	0.000
		SM_T	10.629	35.001	0.231	37.307	117.069	0.228	23.427	75.021	0.149	19.747	39.196	0.102
		U_TPhi	9.948	35.067	0.233	43.052	130.622	0.229	24.145	73.965	0.150	16.265	42.682	0.000
		O_TPhi	11.043	38.769	0.217	34.455	116.582	0.233	23.585	75.019	0.143	15.943	43.205	0.130
		SM_TPhi	9.985	34.939	0.231	30.500	114.130	0.231	23.470	74.991	**0.151**	19.329	36.573	0.116
	svm	Original	10.241	35.832	0.144	**27.205**	118.059	0.217	23.964	80.081	0.093	**10.680**	49.294	0.092
		U	11.369	33.449	0.172	38.261	114.618	0.173	26.546	72.610	**0.131**	16.076	42.675	0.137
		O	10.075	35.168	0.167	28.917	115.783	0.211	**22.335**	79.383	0.110	15.478	44.773	0.140
		SM	10.085	35.156	**0.181**	28.170	117.531	0.223	27.353	72.130	0.127	14.631	44.622	**0.147**
		U_T	16.006	34.856	0.133	35.720	**114.079**	0.188	28.754	72.354	0.121	18.717	39.504	0.106
		O_T	**9.681**	35.658	0.168	27.719	117.682	0.221	24.719	80.031	0.097	11.590	45.222	0.110
		SM_T	10.218	35.097	**0.181**	28.050	117.230	0.223	29.698	72.092	0.118	17.049	41.543	0.113
		U_TPhi	16.273	**32.757**	0.154	27.298	117.704	0.215	24.877	74.520	**0.131**	19.171	39.133	0.108
		O_TPhi	9.749	35.551	0.165	28.265	117.001	0.213	23.913	80.825	0.094	11.435	45.983	0.102
		SM_TPhi	10.001	35.348	0.174	27.901	117.614	**0.225**	31.696	**70.424**	0.118	15.138	**38.237**	0.118
	rf	Original	9.838	36.119	0.187	29.436	117.754	0.209	22.071	79.156	**0.138**	11.864	46.373	0.085
		U	9.832	36.137	**0.188**	29.431	117.706	0.209	22.064	79.125	**0.138**	**11.834**	46.402	0.084
		O	9.825	36.141	0.186	**29.418**	117.769	**0.210**	22.069	79.161	0.137	11.836	46.416	0.084
		SM	9.827	36.129	0.186	29.421	**117.686**	0.209	22.071	79.143	0.136	11.840	46.414	0.085
		U_T	9.827	**36.100**	0.184	29.440	117.706	0.208	22.068	**79.118**	0.137	11.871	46.370	0.084
		O_T	9.825	36.124	**0.188**	29.448	117.759	**0.210**	**22.061**	79.142	**0.138**	11.837	46.380	0.088
		SM_T	9.829	36.137	0.186	29.452	117.829	**0.210**	22.085	79.244	**0.138**	**11.834**	46.390	0.085
		U_TPhi	9.833	36.117	0.185	29.456	117.729	0.208	22.078	79.197	**0.138**	11.851	46.392	0.087
		O_TPhi	9.846	36.160	0.186	29.431	117.756	0.208	22.066	79.152	0.135	11.853	**46.363**	**0.089**
		SM_TPhi	**9.824**	36.127	**0.188**	29.423	117.773	0.208	22.082	79.177	0.137	11.866	46.391	0.083
		Bandari	**9.397**	37.347	0.133	**26.666**	119.002	0.147	**20.567**	83.065	0.066	**10.779**	48.583	**0.097**

Continued on the next page

Table 4 – continued from the previous page

			Economy			Microsoft			Obama			Palestine		
		Approach	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u	$rmse$	$rmse_{phi}$	F_1^u
LinkedIn	lm	Original	**108.919**	326.032	0.194	**339.394**	1062.832	0.251	**80.434**	222.722	0.197	**50.538**	**97.143**	0.131
		U	110.782	317.456	0.195	340.666	1056.487	0.258	81.346	219.570	0.202	52.403	98.013	0.130
		O	111.148	317.923	0.195	341.413	1055.550	0.254	81.751	218.521	0.197	59.453	97.458	0.123
		SM	109.229	321.876	**0.200**	341.223	**1053.917**	0.258	80.983	219.793	0.203	61.318	98.297	0.127
		U_T	109.443	321.044	0.198	339.656	1062.083	0.251	80.757	220.265	**0.208**	52.836	98.936	0.127
		O_T	109.969	325.570	0.195	341.750	1057.873	0.250	82.130	221.322	0.188	55.849	102.804	0.126
		SM_T	109.333	321.785	0.198	339.860	1060.896	0.251	81.215	218.763	0.203	53.733	97.186	0.118
		U_TPhi	109.652	321.705	0.199	341.177	1058.810	0.253	80.724	221.381	0.202	55.764	97.622	**0.135**
		O_TPhi	109.639	324.707	0.193	341.632	1057.572	0.253	82.186	**218.494**	0.192	53.297	102.566	0.122
		SM_TPhi	110.195	**316.965**	**0.200**	340.991	1056.286	**0.259**	80.672	221.460	0.202	54.727	97.685	0.128
	mars	Original	**106.702**	315.037	0.202	**338.601**	1060.762	0.254	79.809	222.980	0.205	**26.346**	**89.450**	0.206
		U	110.820	315.076	0.217	339.189	1055.439	0.269	81.622	220.683	0.206	38.185	89.526	0.195
		O	111.855	311.067	0.213	339.057	1050.537	0.270	82.284	217.022	0.209	32.919	99.919	**0.207**
		SM	110.440	**299.066**	0.218	339.644	1051.803	**0.272**	81.740	**214.429**	0.209	39.430	94.185	0.184
		U_T	109.768	314.186	0.218	339.721	1049.999	0.267	**79.670**	219.637	**0.214**	33.293	91.210	0.197
		O_T	109.426	313.445	0.216	340.790	1049.686	0.267	81.799	218.238	0.211	42.257	106.822	0.184
		SM_T	111.888	311.916	0.216	339.960	1048.241	0.271	80.627	217.148	0.209	27.243	90.764	0.200
		U_TPhi	107.877	301.963	0.216	338.852	1053.437	0.263	81.782	220.471	0.212	28.470	96.007	0.194
		O_TPhi	109.420	313.621	**0.219**	340.982	1053.023	0.269	81.552	217.930	0.210	36.308	103.805	0.192
		SM_TPhi	108.774	303.550	0.215	339.639	**1047.305**	0.268	81.239	221.069	0.209	33.382	90.969	0.192
	svm	Original	113.124	351.150	0.153	344.041	1071.672	0.200	83.230	227.696	0.117	26.159	90.040	0.108
		U	115.780	**325.770**	0.169	339.868	1065.746	0.240	**80.167**	219.569	**0.206**	26.252	88.046	**0.232**
		O	112.573	337.499	0.171	340.523	1065.176	0.240	83.832	225.449	0.143	26.185	89.276	0.205
		SM	**112.016**	335.892	**0.180**	341.478	1067.270	**0.245**	85.200	218.685	0.196	26.177	89.339	0.211
		U_T	116.856	326.059	0.175	**339.818**	**1064.983**	**0.245**	87.046	218.733	0.198	26.309	88.285	0.224
		O_T	112.185	341.884	0.165	340.207	1069.847	0.226	83.249	227.656	0.127	26.357	90.465	0.112
		SM_T	112.187	340.209	0.178	340.412	1065.562	0.244	87.225	217.837	0.201	26.444	88.349	0.186
		U_TPhi	113.025	339.399	0.163	340.241	1066.707	0.233	90.898	217.912	0.190	**26.131**	**87.723**	0.224
		O_TPhi	112.401	342.698	0.165	340.337	1070.334	0.227	83.092	228.094	0.127	26.294	90.628	0.118
		SM_TPhi	112.852	336.958	0.179	341.048	1068.278	0.241	83.409	**212.469**	0.181	26.207	88.926	0.230
	rf	Original	112.453	337.039	0.183	**339.141**	1066.389	0.231	80.663	224.595	0.150	28.125	90.394	0.089
		U	112.440	337.007	0.184	339.143	1066.438	0.229	80.690	**224.502**	0.149	28.094	90.396	0.090
		O	112.446	336.998	0.183	339.185	1066.433	**0.232**	**80.660**	224.564	0.149	28.186	**90.386**	0.091
		SM	112.511	337.104	0.183	339.173	1066.353	0.230	80.716	224.645	0.148	28.051	90.401	0.089
		U_T	112.472	337.143	0.183	339.202	1066.490	0.231	80.685	224.640	0.149	28.075	90.397	0.096
		O_T	112.483	336.988	**0.185**	339.168	1066.449	0.230	80.665	224.594	0.147	28.002	90.447	0.088
		SM_T	**112.435**	**336.961**	**0.185**	339.162	**1066.350**	0.228	80.700	224.599	0.147	28.343	90.414	0.084
		U_TPhi	112.463	337.112	0.183	339.179	1066.485	0.230	80.702	224.714	0.147	28.377	90.390	**0.098**
		O_TPhi	112.456	337.096	0.184	339.156	1066.402	0.228	80.702	224.671	**0.151**	28.391	90.412	0.092
		SM_TPhi	112.490	337.138	0.183	339.154	1066.439	0.228	80.713	224.698	0.146	**27.926**	90.417	0.085
		Bandari	115.654	362.753	0.146	340.941	1081.835	0.112	80.729	232.827	0.065	**26.305**	**90.216**	**0.110**

E Evaluation Results of A Posteriori Prediction

Table 5: Evaluation results concerning the evaluation metrics $rmse$, $rmse_\phi$ and F_1^u, for all combinations of social media sources and news topics in *a posteriori* prediction tasks, regarding all baselines and proposed methods.

			Economy			**Microsoft**			**Obama**			**Palestine**		
		Approach	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u
t=1	**Twitter**	ConstScale	99.134	266.217	0.494	**98.130**	**274.511**	0.634	266.891	828.772	0.311	119.927	396.836	0.380
		Linear-log	325.390	851.459	0.329	369.399	1008.403	0.362	697.339	2061.846	0.277	144.111	454.944	0.426
		LM	116.922	326.539	0.339	140.225	403.677	0.447	289.605	866.863	0.244	130.304	420.209	0.332
		ML	116.818	326.300	0.336	140.013	403.207	0.445	289.384	866.375	0.245	129.832	418.902	0.332
		MRBF	116.715	325.643	0.347	140.006	403.114	0.448	289.448	866.038	0.245	131.486	423.734	0.366
		kernel	**98.621**	**263.350**	**0.537**	105.368	295.774	**0.650**	**259.557**	**796.544**	**0.428**	119.986	394.819	**0.519**
		knn	99.388	266.978	**0.537**	105.360	296.501	0.638	266.947	830.101	0.356	**117.911**	**389.919**	0.506
	Facebook	ConstScale	**141.619**	**501.463**	0.230	**135.398**	**514.137**	0.296	**661.749**	2270.259	0.099	**139.277**	**536.081**	0.054
		Linear-log	879.653	3076.914	0.238	306.445	1136.190	0.262	2492.805	8141.008	0.266	216.871	751.448	**0.360**
		LM	259.712	917.235	0.093	363.947	1342.361	0.105	686.703	2375.135	0.085	227.009	858.999	0.226
		ML	262.645	928.057	0.088	366.125	1350.881	0.109	686.622	2376.208	0.085	234.505	884.036	0.225
		MRBF	682.659	2051.217	0.302	367.468	1356.701	0.124	686.673	2375.740	0.087	231.667	870.353	0.220
		kernel	221.419	591.907	0.284	201.353	684.440	0.265	673.381	**2052.759**	0.299	253.435	639.643	0.229
		knn	214.941	596.315	**0.329**	194.181	675.035	**0.335**	664.676	2153.376	**0.310**	245.185	650.518	0.277
	Google+	ConstScale	5.637	39.155	0.111	19.567	**105.016**	0.098	14.373	74.411	0.029	4.623	44.435	0.000
		Linear-log	9.161	46.002	**0.239**	28.746	138.748	0.268	21.091	85.746	**0.180**	9.647	**32.621**	**0.284**
		LM	5.763	40.420	0.102	**18.587**	110.859	0.265	14.867	83.456	0.050	5.843	54.858	0.000
		ML	5.631	39.983	0.215	18.618	114.587	0.287	14.673	83.504	0.060	5.680	53.464	0.072
		MRBF	5.636	40.017	0.215	19.678	122.086	0.253	14.723	82.903	0.142	5.771	54.670	0.051
		kernel	5.666	**37.284**	0.192	23.596	118.781	0.305	15.024	**64.783**	0.150	4.719	39.256	0.206
		knn	**5.586**	38.391	0.209	23.742	127.069	**0.318**	**14.012**	69.848	0.135	**4.563**	40.123	0.234
	LinkedIn	ConstScale	108.099	492.970	0.107	**248.536**	**918.966**	0.070	62.010	242.137	0.062	5.474	43.603	0.102
		Linear-log	187.911	829.120	**0.259**	2149.601	7952.599	0.281	77.504	286.629	**0.283**	8.516	50.843	0.094
		LM	89.141	424.400	0.185	395.821	1672.898	0.269	49.493	222.853	0.158	10.822	106.848	0.000
		ML	89.012	424.950	0.186	419.268	1773.329	0.264	49.379	223.185	0.165	10.539	105.939	0.091
		MRBF	**89.010**	**423.574**	0.187	399.767	1676.686	0.259	**49.318**	**222.739**	0.155	10.529	105.851	0.072
		kernel	111.649	487.449	0.235	315.184	1048.999	**0.290**	59.176	223.936	0.232	**5.426**	**42.536**	0.146
		knn	110.992	495.947	0.205	273.675	925.223	0.239	59.078	230.161	0.193	5.466	43.434	**0.186**

Continued on the next page

Table 5 – continued from the previous page

			Economy			Microsoft			Obama			Palestine		
		Approach	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u
t=2	**Twitter**	ConstScale	**95.238**	**256.055**	0.524	**89.326**	**248.912**	0.673	**253.234**	788.231	0.359	**115.821**	**382.838**	0.399
		Linear-Log	297.727	779.820	0.339	346.090	944.431	0.368	648.412	1926.452	0.281	133.503	419.827	0.431
		LM	115.791	323.421	0.350	137.441	395.399	0.469	284.668	851.454	0.268	128.259	413.680	0.356
		ML	115.391	322.461	0.355	140.460	403.635	0.463	287.535	857.618	0.277	127.714	412.247	0.370
		MRBF	115.320	321.836	0.363	140.441	403.610	0.462	287.485	857.215	0.278	129.352	417.155	0.388
		kernel	98.345	263.332	**0.560**	96.495	269.944	**0.675**	255.459	**787.453**	**0.448**	118.649	390.515	**0.530**
		knn	97.755	263.040	0.541	95.475	267.711	0.669	258.408	804.926	0.377	116.536	385.022	0.516
	Facebook	ConstScale	162.778	573.566	0.246	**320.121**	**1177.864**	0.364	638.624	2179.478	0.141	**134.785**	**527.558**	0.045
		Linear-Log	705.676	2475.415	0.263	355.227	1298.585	0.317	2093.867	6760.458	0.280	216.276	758.529	**0.351**
		LM	257.076	908.490	0.108	368.675	1362.026	0.113	677.546	2341.414	0.157	226.981	862.994	0.226
		ML	257.106	909.189	0.104	357.400	1321.924	0.147	675.975	2343.379	0.157	214.366	831.672	0.217
		MRBF	657.498	2009.394	**0.362**	358.324	1326.854	0.134	676.066	2343.680	0.157	216.173	835.673	0.218
		kernel	176.537	**562.328**	0.285	460.209	1714.828	0.369	655.811	**2026.965**	0.358	189.534	583.768	0.331
		knn	**161.554**	565.706	0.358	455.411	1703.614	**0.463**	**633.345**	2111.491	**0.363**	179.309	604.490	0.287
	Google+	ConstScale	5.472	38.127	0.129	18.801	**99.890**	0.147	**14.136**	73.081	0.037	4.595	42.777	0.000
		Linear-Log	9.001	46.910	0.247	29.237	142.086	0.292	23.567	94.036	0.185	10.269	**33.349**	**0.242**
		LM	5.666	39.980	0.146	17.975	108.152	0.298	14.717	81.639	0.132	5.815	54.039	0.066
		ML	5.538	39.606	0.232	**17.274**	107.946	**0.320**	14.401	82.684	0.110	5.595	53.019	0.134
		MRBF	5.545	39.640	0.229	17.715	111.726	0.312	14.395	82.563	0.122	5.666	53.914	0.121
		kernel	5.347	**35.004**	0.278	22.761	113.566	0.296	15.335	**67.196**	**0.198**	4.702	37.415	0.235
		knn	**5.257**	36.220	**0.281**	20.548	108.562	0.309	14.206	70.980	0.168	**4.483**	39.786	0.176
	LinkedIn	ConstScale	105.357	480.803	0.134	**233.988**	**861.858**	0.077	61.195	239.015	0.082	**5.417**	**43.329**	0.111
		Linear-Log	229.445	1024.893	0.251	1981.599	7316.892	0.308	81.907	306.074	0.281	9.774	57.257	**0.216**
		LM	**88.590**	**422.256**	0.210	430.211	1820.690	0.288	49.011	**220.150**	0.210	9.361	88.180	0.095
		ML	90.040	430.835	0.210	361.154	1528.014	0.291	49.030	221.914	0.206	10.514	105.801	0.092
		MRBF	89.206	426.642	0.210	375.897	1591.305	0.290	**48.787**	220.659	0.209	10.496	105.605	0.078
		kernel	107.652	470.248	**0.282**	306.985	1066.012	**0.325**	59.141	223.030	**0.293**	5.788	47.075	0.119
		knn	106.088	478.720	0.245	265.636	940.580	0.318	59.490	231.372	0.238	5.751	47.728	0.120

Continued on the next page

Table 5 – continued from the previous page

			Economy			Microsoft			Obama			Palestine		
		Approach	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u
t=3	Twitter	ConstScale	93.111	250.463	0.545	**83.283**	**231.634**	0.691	247.141	769.641	0.389	110.468	364.063	0.418
		Linear-Log	282.116	738.852	0.337	342.146	935.754	0.360	604.941	1795.512	0.283	127.947	401.626	0.427
		LM	114.619	320.262	0.363	136.111	391.484	0.480	281.995	843.225	0.283	125.320	403.974	0.372
		ML	115.457	322.707	0.349	139.377	401.437	0.471	276.329	824.258	0.313	119.783	385.478	0.375
		MRBF	115.428	322.260	0.357	139.168	401.042	0.472	275.922	823.142	0.313	120.789	388.431	0.383
		kernel	**90.782**	**242.639**	0.565	99.036	279.588	0.687	**240.140**	**741.669**	**0.475**	113.058	370.975	**0.554**
		knn	91.582	246.313	**0.576**	98.548	278.851	**0.692**	248.452	775.252	0.409	**110.126**	**362.581**	0.551
	Facebook	ConstScale	240.155	851.213	0.269	**299.022**	**1069.749**	0.392	620.848	2109.259	0.171	**132.312**	516.352	0.051
		Linear-Log	2387.559	8404.559	0.281	302.453	1075.164	0.350	1667.431	5372.287	0.302	188.445	672.489	**0.374**
		LM	239.641	846.706	0.084	367.727	1359.644	0.126	668.014	2309.325	0.212	216.254	835.456	0.235
		ML	262.537	931.591	0.114	352.765	1312.942	0.143	666.646	2325.394	0.228	212.974	829.248	0.226
		MRBF	613.605	1833.593	**0.408**	356.898	1327.258	0.125	666.669	2325.509	0.228	213.638	830.420	0.225
		kernel	206.948	**713.287**	0.305	323.287	1154.831	0.433	616.487	**1866.541**	0.403	144.849	**485.243**	0.283
		knn	**202.360**	718.429	0.380	320.128	1152.118	**0.494**	**590.077**	1948.483	**0.416**	136.568	501.453	0.294
	Google+	ConstScale	5.346	37.416	0.148	18.078	**96.376**	0.168	13.897	72.369	0.046	4.553	42.623	0.144
		Linear-Log	9.172	49.637	0.242	28.093	138.295	0.308	25.310	102.482	0.199	9.068	**32.264**	0.263
		LM	5.594	39.742	0.162	17.207	105.163	0.327	14.590	80.835	0.158	5.814	53.693	0.064
		ML	5.466	39.293	0.217	**16.741**	105.296	0.336	14.109	81.724	0.121	5.591	53.267	0.155
		MRBF	5.473	39.303	0.227	17.145	108.471	0.347	14.141	81.174	0.143	5.583	55.788	0.197
		kernel	5.170	**34.625**	**0.349**	21.469	109.917	**0.402**	14.725	**63.324**	**0.227**	4.734	34.794	**0.297**
		knn	**5.155**	35.977	0.288	19.646	103.939	0.394	**13.493**	68.769	0.198	**4.350**	37.316	0.224
	LinkedIn	ConstScale	103.792	475.230	0.161	**221.429**	**815.070**	0.089	60.850	238.052	0.088	**5.310**	**42.591**	0.122
		Linear-Log	233.073	1048.986	0.276	1513.289	5597.678	0.335	82.484	311.311	0.282	9.177	53.617	**0.217**
		LM	**87.781**	**419.429**	0.213	453.970	1929.567	0.304	49.020	219.970	0.245	8.450	78.617	0.098
		ML	89.405	428.076	0.217	364.189	1546.896	0.304	48.154	218.216	0.218	10.321	104.171	0.108
		MRBF	88.829	423.721	0.240	359.636	1524.828	0.307	**47.959**	**217.006**	0.236	10.383	104.685	0.101
		kernel	102.432	451.154	**0.291**	266.896	940.494	0.339	59.266	225.946	**0.287**	5.764	44.182	0.207
		knn	100.853	459.805	0.263	226.114	818.565	**0.366**	59.723	232.910	0.247	5.874	45.263	0.177

F Evaluation Results of Hybrid Methods

Table 6: Evaluation results concerning the evaluation metrics $rmse$, $rmse_\phi$ and F_1^u, for all combinations of social media sources and news topics in *a posteriori* prediction tasks, regarding the proposed hybrid methods (time-based ensembles.

			Economy			**Microsoft**			**Obama**			**Palestine**		
		Approach	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u
t=1	**Twitter**	kernel	124.899	331.663	0.317	154.985	432.001	0.435	**259.557**	796.544	**0.428**	152.941	454.817	0.347
		ENSt_kernel	119.393	**304.849**	**0.397**	**125.056**	**345.495**	**0.519**	285.985	**751.647**	0.312	144.381	421.558	**0.437**
		ENSphi_kernel	119.476	314.492	0.322	137.602	386.283	0.437	291.075	764.366	0.265	148.205	437.015	0.353
		knn	127.567	338.029	0.308	155.455	433.184	0.410	266.947	830.101	0.356	152.558	454.066	0.314
		ENSt_knn	120.268	308.718	0.379	125.245	346.769	**0.519**	287.423	759.377	0.283	**143.613**	**420.667**	0.435
		ENSphi_knn	**119.289**	314.905	0.312	137.336	386.255	0.423	292.926	773.654	0.198	146.647	433.454	0.316
	Facebook	kernel	310.428	926.061	0.101	427.280	1393.641	0.110	758.956	2530.381	0.159	329.333	961.093	0.145
		ENSt_kernel	307.272	**910.962**	**0.268**	499.114	1490.621	0.335	761.863	**2378.565**	0.095	315.738	**900.821**	0.288
		ENSphi_kernel	306.840	911.154	0.112	425.661	**1387.258**	0.118	741.909	2384.592	**0.164**	325.650	937.408	0.165
		knn	311.453	933.720	0.084	427.820	1397.538	0.115	763.224	2588.006	0.133	329.281	980.386	0.152
		ENSt_knn	307.230	911.683	**0.268**	498.942	1490.691	**0.336**	761.728	2387.506	0.090	**315.039**	902.385	**0.294**
		ENSphi_knn	**305.745**	912.242	0.096	**424.913**	1387.272	0.125	**734.794**	2404.295	0.136	321.307	941.384	0.173
	Google+	kernel	9.615	38.873	0.143	26.858	114.541	0.253	20.163	79.045	0.102	11.114	50.455	0.189
		ENSt_kernel	10.211	36.699	**0.233**	27.469	107.490	0.254	23.062	73.092	**0.158**	14.105	46.290	0.134
		ENSphi_kernel	9.714	**36.301**	0.165	27.044	113.189	0.254	22.703	**72.144**	0.106	12.402	**44.450**	0.109
		knn	9.687	40.447	0.129	26.544	121.102	0.245	**19.530**	85.561	0.082	**10.570**	54.929	**0.217**
		ENSt_knn	10.147	36.829	0.232	26.607	**105.822**	0.259	22.507	74.017	0.151	13.807	46.938	0.144
		ENSphi_knn	**9.475**	36.579	0.156	**26.304**	117.109	**0.272**	21.581	73.518	0.085	11.572	44.944	0.119
	LinkedIn	kernel	136.633	425.636	0.164	346.116	1028.013	0.229	77.540	220.524	0.165	28.075	99.929	0.074
		ENSt_kernel	112.823	**328.512**	**0.234**	350.713	1088.490	**0.281**	79.475	218.376	**0.219**	30.789	98.836	0.197
		ENSphi_kernel	134.593	409.120	**0.234**	343.669	1010.278	0.271	76.972	**215.741**	0.202	28.018	**97.010**	0.067
		knn	137.758	437.584	0.134	324.600	982.058	0.159	78.358	227.544	0.112	28.491	102.049	0.088
		ENSt_knn	**112.691**	329.685	**0.234**	347.107	1078.883	**0.281**	79.428	219.004	0.217	30.616	99.041	**0.200**
		ENSphi_knn	133.164	412.057	0.211	**319.401**	**953.548**	0.223	**76.375**	218.036	0.136	**27.973**	97.187	0.079

Continued on the next page

Table 6 – continued from the previous page

			Economy			Microsoft			Obama			Palestine		
		Approach	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u	$rmse$	$rmse_\phi$	F_1^u
t=2	Twitter	kernel	126.679	335.816	0.341	150.660	419.418	0.460	**255.459**	787.453	**0.448**	152.017	452.369	0.356
		ENSt_kernel	**118.733**	**305.486**	**0.390**	**123.885**	**342.838**	0.528	282.607	**743.668**	0.317	144.179	422.452	**0.434**
		ENSphi_kernel	119.599	315.225	0.346	133.138	373.378	0.462	289.508	760.608	0.271	147.230	434.698	0.366
		knn	126.994	336.556	0.326	151.009	420.128	0.444	258.408	804.926	0.377	151.773	451.811	0.333
		ENSt_knn	119.194	307.620	0.375	124.116	344.238	**0.529**	283.946	750.922	0.286	**143.632**	**422.107**	0.430
		ENSphi_knn	118.829	313.862	0.330	132.656	372.662	0.447	289.747	764.679	0.219	145.985	431.805	0.333
	Facebook	kernel	297.612	910.922	0.121	450.788	1484.096	0.142	744.766	2489.714	0.204	297.546	948.712	0.217
		ENSt_kernel	304.123	906.682	**0.279**	485.962	1469.367	0.337	757.418	2371.312	0.114	311.308	**898.681**	**0.300**
		ENSphi_kernel	293.730	**895.943**	0.124	448.810	1477.012	0.143	728.451	**2349.996**	**0.205**	292.123	921.230	0.211
		knn	296.890	916.006	0.106	450.426	1483.846	0.139	746.830	2545.475	0.176	295.337	964.614	0.159
		ENSt_knn	304.044	907.730	0.274	485.764	**1469.116**	**0.341**	757.564	2383.083	0.104	310.582	900.556	0.297
		ENSphi_knn	**291.315**	896.167	0.112	**448.023**	1475.544	0.144	**720.456**	2371.189	0.177	**286.626**	927.459	0.187
	Google+	kernel	9.383	37.758	0.229	25.825	112.099	0.254	20.327	81.444	0.157	10.951	50.566	**0.184**
		ENSt_kernel	10.069	36.483	0.237	26.474	103.610	0.271	22.913	**72.719**	**0.169**	13.828	46.563	0.139
		ENSphi_kernel	9.491	**35.433**	**0.247**	26.031	110.309	0.254	22.981	74.911	0.162	12.221	**44.935**	0.166
		knn	9.519	39.732	0.152	24.212	110.233	0.223	**19.856**	86.735	0.100	**10.480**	55.762	0.129
		ENSt_knn	9.991	36.756	0.238	25.705	**102.330**	0.271	22.398	73.735	0.161	13.493	47.494	0.149
		ENSphi_knn	**9.256**	35.934	0.190	**24.038**	106.479	**0.290**	21.976	75.457	0.111	11.407	46.370	0.081
	LinkedIn	kernel	135.169	418.787	0.206	342.727	1047.740	0.270	77.534	219.194	0.214	28.321	100.852	0.059
		ENSt_kernel	112.457	**328.086**	**0.240**	348.354	1082.116	0.288	79.120	217.499	0.227	30.311	98.495	**0.217**
		ENSphi_kernel	133.662	404.828	0.225	340.481	1033.749	0.277	77.230	**215.782**	**0.231**	28.298	**98.266**	0.053
		knn	135.408	431.904	0.163	320.601	997.277	0.233	78.699	228.352	0.153	28.819	103.015	0.060
		ENSt_knn	**112.237**	329.607	0.239	343.811	1070.342	**0.290**	79.119	218.555	0.223	30.071	98.832	**0.217**
		ENSphi_knn	131.108	407.856	0.181	**316.496**	**976.009**	0.263	**76.834**	219.374	0.166	**28.210**	98.298	0.054
t=3	Twitter	kernel	123.656	327.587	0.349	152.119	424.522	0.469	**240.140**	741.669	**0.475**	149.456	444.766	0.369
		ENSt_kernel	117.865	**304.389**	**0.388**	**124.639**	**345.973**	**0.528**	280.004	**737.404**	0.322	142.573	418.805	0.428
		ENSphi_kernel	**116.390**	306.613	0.351	134.964	379.688	0.471	283.670	746.115	0.285	144.597	427.105	0.369
		knn	124.863	330.669	0.335	152.350	424.913	0.463	248.452	775.252	0.409	148.768	442.811	0.352
		ENSt_knn	118.431	306.779	0.375	124.809	347.093	0.525	282.203	746.775	0.292	**141.671**	**417.351**	**0.431**
		ENSphi_knn	116.483	307.589	0.339	134.501	378.922	0.466	286.609	757.295	0.239	142.882	422.744	0.358
	Facebook	kernel	257.654	786.498	0.132	420.497	1377.988	0.175	728.019	2430.893	0.244	285.294	926.682	0.218
		ENSt_kernel	287.874	858.356	**0.281**	477.788	1454.137	0.345	748.988	2355.826	0.130	306.528	**891.486**	**0.308**
		ENSphi_kernel	253.506	**771.039**	0.136	418.429	1370.586	0.176	710.826	**2289.155**	**0.245**	279.750	899.297	0.220
		knn	257.919	790.755	0.121	420.662	1379.040	0.161	728.398	2485.028	0.218	285.923	941.892	0.161
		ENSt_knn	287.638	858.812	0.279	477.641	1454.049	**0.346**	749.600	2369.697	0.115	306.036	894.320	0.306
		ENSphi_knn	**252.478**	771.496	0.124	**418.116**	**1370.277**	0.162	**702.453**	2315.111	0.222	**277.141**	905.369	0.176
	Google+	kernel	9.268	37.738	0.244	24.838	109.841	**0.331**	19.693	78.713	0.167	10.887	49.792	0.212
		ENSt_kernel	9.908	36.336	0.242	25.974	102.777	0.276	22.602	72.552	0.170	13.596	46.517	0.141
		ENSphi_kernel	9.350	**35.408**	**0.247**	25.045	107.966	0.325	22.330	**72.079**	**0.179**	12.154	**44.491**	**0.282**
		knn	9.620	40.008	0.166	23.613	107.463	0.293	**19.176**	84.424	0.121	**10.406**	55.293	0.150
		ENSt_knn	9.825	36.731	0.226	25.236	**101.296**	0.277	22.056	73.786	0.157	13.243	47.591	0.148
		ENSphi_knn	**9.209**	36.016	0.190	**23.477**	103.983	0.324	21.233	73.178	0.125	11.283	46.087	0.076
	LinkedIn	kernel	130.978	408.598	0.220	315.271	976.993	0.279	77.238	219.753	0.212	**27.801**	99.122	0.117
		ENSt_kernel	112.276	**329.256**	**0.244**	344.178	1071.240	**0.298**	78.219	**215.547**	**0.244**	29.869	97.880	0.209
		ENSphi_kernel	129.098	393.868	0.234	312.573	960.577	0.289	76.972	216.593	0.234	28.041	**97.354**	0.140
		knn	131.463	420.529	0.175	294.332	929.742	0.280	78.450	227.623	0.170	28.454	101.761	0.094
		ENSt_knn	**112.035**	331.117	0.238	338.990	1057.371	0.296	78.119	216.532	0.232	29.648	98.443	**0.216**
		ENSphi_knn	127.071	396.893	0.193	**289.837**	**907.312**	0.294	**76.686**	219.225	0.176	28.055	97.795	0.105

G Evaluation Results of Single-Source Ranking Tasks (Google News)

Table 7: Evaluation results concerning the evaluation metrics MAP, MRR and $NDCG@10$, for all combinations of social media sources and news topics in single-source ranking tasks using Google News rankings' data, regarding all *a posteriori* prediction approaches.

		Economy			**Microsoft**			**Obama**			**Palestine**		
	Approach	MAP	MRR	$NDCG@10$	MAP	MRR	$NDCG@10$	MAP	MRR	$NDCG@10$	MAP	MRR	$NDCG@10$
Twitter	Time	0.119	0.271	0.222	0.348	0.555	0.439	0.154	0.353	0.239	0.511	0.761	0.691
	Live	0.932	0.993	0.915	0.941	0.999	0.945	0.847	0.985	0.870	0.951	0.997	0.954
	Source	0.737	0.965	0.753	0.929	0.999	0.915	0.602	0.893	0.631	0.852	0.902	0.843
	ConstScale	0.937	0.993	0.922	0.950	**1.000**	0.952	0.864	0.987	0.884	0.960	0.997	0.962
	Linear-log	0.945	0.993	**0.927**	0.952	**1.000**	0.952	0.896	**0.992**	0.906	0.960	0.997	0.956
	LM	0.937	0.993	0.918	0.947	**1.000**	0.942	0.894	**0.992**	**0.909**	0.909	0.997	0.893
	ML	0.931	0.989	0.912	0.950	**1.000**	0.943	0.892	0.991	0.899	0.889	0.985	0.870
	MRBF	0.931	0.989	0.910	0.953	**1.000**	0.950	0.893	**0.992**	0.900	0.878	0.988	0.857
	kernel	0.943	**0.997**	0.921	**0.962**	**1.000**	0.952	0.893	**0.992**	0.902	0.961	0.997	0.944
	ENSt_kernel	**0.946**	**0.997**	0.921	0.960	**1.000**	0.948	0.885	**0.992**	0.889	0.942	**0.999**	0.916
	ENSphi_kernel	0.940	**0.997**	0.921	0.961	**1.000**	**0.958**	0.888	**0.992**	0.900	0.961	**0.999**	0.945
	knn	0.936	**0.997**	0.912	**0.962**	**1.000**	0.953	0.874	**0.992**	0.879	0.916	**0.999**	0.900
	ENSt_knn	0.763	0.952	0.768	0.914	0.999	0.903	0.677	0.924	0.702	0.849	0.932	0.841
	ENSphi_knn	0.789	0.987	0.783	0.922	0.999	0.909	0.624	0.896	0.692	0.822	0.906	0.835
	BestAPriori	0.284	0.600	0.378	0.642	0.807	0.685	0.454	0.748	0.524	0.533	0.739	0.639
	Bandari	0.945	0.993	**0.927**	0.952	**1.000**	0.949	**0.899**	**0.992**	**0.909**	**0.962**	0.997	0.946
	Official	0.939	0.993	0.923	0.948	**1.000**	0.951	0.871	0.987	0.889	0.961	0.997	**0.963**
Facebook	Time	0.178	0.371	0.274	0.459	0.713	0.561	0.326	0.603	0.457	0.548	0.688	0.636
	Live	0.706	0.911	0.715	0.773	0.929	0.783	0.806	0.908	0.826	0.927	0.989	0.926
	Source	0.532	0.748	0.572	0.737	0.894	0.780	0.507	0.746	0.617	0.867	0.932	0.846
	ConstScale	0.724	0.911	0.728	0.788	0.930	0.795	0.839	0.923	0.853	0.936	0.989	0.934
	Linear-log	0.748	0.911	0.747	0.787	0.922	0.797	0.875	0.940	0.882	0.948	0.996	0.919
	LM	0.743	**0.940**	0.758	0.771	0.930	0.797	0.887	0.949	**0.906**	0.930	0.994	0.896
	ML	0.696	0.912	0.721	0.652	0.908	0.714	0.826	0.907	0.852	0.884	0.959	0.862
	MRBF	0.724	0.920	0.733	0.718	0.947	0.745	0.825	0.903	0.853	0.876	0.963	0.856
	kernel	0.765	0.939	0.797	**0.891**	0.990	**0.890**	0.877	0.938	0.896	0.947	0.999	0.916
	ENSt_kernel	**0.782**	**0.940**	**0.809**	0.881	0.990	0.883	**0.893**	**0.951**	0.902	0.927	**1.000**	0.893
	ENSphi_kernel	0.762	0.939	0.795	0.886	**0.992**	0.886	0.864	0.925	0.890	0.948	**1.000**	0.916
	knn	0.770	0.939	0.803	0.879	**0.992**	0.879	0.881	0.946	0.890	0.920	**1.000**	0.887
	ENSt_knn	0.549	0.770	0.591	0.708	0.891	0.740	0.546	0.722	0.634	0.811	0.853	0.817
	ENSphi_knn	0.527	0.746	0.584	0.712	0.848	0.778	0.588	0.723	0.694	0.788	0.936	0.798
	BestAPriori	0.253	0.397	0.354	0.611	0.852	0.687	0.476	0.769	0.544	0.676	0.887	0.668
	Bandari	0.736	0.912	0.757	0.809	0.933	0.825	0.884	0.944	0.890	**0.952**	0.999	0.921
	Official	0.730	0.911	0.740	0.791	0.933	0.818	0.865	0.932	0.872	0.944	0.989	**0.949**

Continued on the next page

Table 7 – continued from the previous page

		Economy			Microsoft			Obama			Palestine		
	Approach	MAP	MRR	$NDCG$@10	MAP	MRR	$NDCG$@10	MAP	MRR	$NDCG$@10	MAP	MRR	$NDCG$@10
Google+	Time	0.165	0.382	0.270	0.382	0.618	0.514	0.368	0.597	0.477	0.570	0.818	0.592
	Live	0.752	0.880	0.773	0.906	0.971	0.893	0.806	0.891	0.833	0.836	0.939	0.846
	Source	0.494	0.685	0.574	0.685	0.741	0.768	0.548	0.748	0.647	0.902	0.992	0.894
	ConstScale	0.757	0.881	0.778	0.920	0.971	0.903	0.820	0.897	0.849	0.836	0.939	0.846
	Linear-log	0.746	0.880	0.772	**0.934**	0.972	**0.924**	0.833	0.899	0.869	0.762	0.993	0.757
	LM	**0.781**	0.897	0.796	0.911	0.981	0.896	0.866	0.934	0.888	0.830	**1.000**	0.818
	ML	0.730	0.889	0.747	0.893	0.982	0.881	0.840	0.924	0.872	0.683	0.819	0.714
	MRBF	0.729	0.889	0.745	0.872	0.981	0.865	0.833	0.914	0.868	0.680	0.814	0.713
	kernel	0.777	0.895	**0.806**	0.919	0.985	0.901	**0.876**	0.950	**0.895**	0.882	0.999	0.860
	ENSt_kernel	0.763	**0.900**	0.786	0.903	0.975	0.888	0.870	**0.962**	0.885	0.676	0.932	0.719
	ENSphi_kernel	0.771	**0.900**	0.794	0.929	0.974	0.917	0.850	0.933	0.874	**0.917**	0.999	**0.906**
	knn	0.769	0.894	0.779	0.905	0.979	0.883	0.850	0.944	0.876	0.651	0.902	0.707
	ENSt_knn	0.549	0.701	0.600	0.781	0.960	0.800	0.612	0.798	0.683	0.777	0.997	0.771
	ENSphi_knn	0.566	0.701	0.625	0.858	**0.989**	0.833	0.627	0.791	0.701	0.798	0.999	0.778
	BestAPriori	0.345	0.541	0.422	0.595	0.792	0.683	0.509	0.785	0.573	0.735	0.975	0.664
	Bandari	0.770	0.890	0.800	0.921	0.982	0.904	0.866	0.935	0.890	0.873	0.992	0.861
	Official	0.754	0.879	0.776	0.922	0.971	0.912	0.829	0.900	0.852	0.835	0.939	0.845
LinkedIn	Time	0.195	0.437	0.295	0.378	0.585	0.513	0.289	0.557	0.418	0.305	0.549	0.289
	Live	0.738	0.874	0.770	0.853	0.941	0.872	0.787	0.920	0.822	**0.927**	0.954	**0.940**
	Source	0.597	0.848	0.624	0.762	0.925	0.794	0.575	0.720	0.652	0.871	0.890	0.864
	ConstScale	0.760	0.879	0.789	0.880	0.951	0.894	0.790	0.922	0.824	**0.927**	0.954	**0.940**
	Linear-log	0.790	0.892	0.819	0.925	0.959	**0.926**	0.839	0.938	0.881	0.570	0.985	0.608
	LM	0.760	0.871	0.810	0.919	0.975	0.909	**0.856**	0.957	**0.886**	0.757	0.999	0.770
	ML	0.712	0.830	0.739	0.900	0.983	0.894	0.771	0.901	0.787	0.635	0.975	0.688
	MRBF	0.728	0.833	0.756	0.897	0.980	0.894	0.762	0.900	0.784	0.643	0.975	0.696
	kernel	0.798	0.905	0.830	0.928	**1.000**	0.923	0.854	0.960	0.883	0.675	0.999	0.716
	ENSt_kernel	**0.804**	0.914	0.821	**0.929**	**1.000**	0.912	0.850	**0.966**	0.874	0.544	0.958	0.594
	ENSphi_kernel	0.802	0.908	**0.832**	0.915	0.987	0.919	0.843	0.961	0.874	0.788	**1.000**	0.799
	knn	0.793	**0.915**	0.813	0.913	**1.000**	0.899	0.846	0.963	0.866	0.548	0.989	0.596
	ENSt_knn	0.569	0.810	0.616	0.872	0.997	0.864	0.566	0.753	0.654	0.566	0.945	0.596
	ENSphi_knn	0.531	0.733	0.623	0.729	0.911	0.793	0.594	0.755	0.693	0.320	0.519	0.449
	BestAPriori	0.357	0.596	0.443	0.519	0.689	0.652	0.413	0.600	0.504	0.331	0.548	0.323
	Bandari	0.783	0.878	0.817	0.925	0.964	0.920	0.839	0.946	0.878	0.697	0.999	0.732
	Official	0.790	0.894	0.811	0.906	0.956	0.912	0.809	0.933	0.833	**0.927**	0.954	0.939

H Evaluation Results of Single-Source Ranking Tasks (Yahoo! News)

Table 8: Evaluation results concerning the evaluation metrics MAP, MRR and $NDCG$@10, for all combinations of social media sources and news topics in single-source ranking tasks using Yahoo! News rankings' data, regarding all *a posteriori* prediction approaches.

		Economy			**Microsoft**			**Obama**			**Palestine**		
	Approach	MAP	MRR	$NDCG$@10	MAP	MRR	$NDCG$@10	MAP	MRR	$NDCG$@10	MAP	MRR	$NDCG$@10
Facebook	Time	0.539	0.826	0.550	0.611	0.843	0.700	0.632	0.910	0.656	0.578	0.982	0.527
	Live	0.885	0.958	0.883	0.872	0.908	0.873	0.946	0.983	0.950	0.929	0.988	0.922
	Source	0.845	0.871	0.833	0.921	**1.000**	0.895	0.841	0.956	0.852	0.875	0.866	0.878
	ConstScale	0.889	0.958	0.885	0.872	0.908	0.873	0.960	0.986	0.959	0.929	0.988	0.922
	Linear-log	0.864	0.975	0.849	0.862	0.905	0.854	0.959	**1.000**	0.942	0.885	0.999	0.873
	LM	0.885	**1.000**	0.868	0.936	0.985	0.918	0.925	**1.000**	0.900	0.890	0.911	0.872
	ML	0.740	0.890	0.745	0.755	0.838	0.781	0.896	0.939	0.886	0.736	0.795	0.758
	MRBF	0.724	0.880	0.735	0.790	0.879	0.803	0.894	0.938	0.886	0.748	0.815	0.770
	kernel	**0.921**	0.988	0.916	**0.978**	0.999	0.963	0.908	**1.000**	0.894	0.952	**1.000**	0.919
	ENSt_kernel	0.849	0.997	0.844	0.917	0.993	0.884	0.850	**1.000**	0.846	0.754	0.873	0.773
	ENSphi_kernel	0.918	0.994	0.915	0.970	**1.000**	**0.968**	0.917	**1.000**	0.900	**0.970**	0.999	**0.948**
	knn	0.844	0.997	0.839	0.901	0.992	0.870	0.850	**1.000**	0.846	0.729	0.859	0.755
	ENSt_knn	0.732	0.858	0.755	0.951	**1.000**	0.934	0.766	0.993	0.786	0.841	0.926	0.851
	ENSphi_knn	0.789	0.942	0.799	0.887	0.997	0.877	0.811	**1.000**	0.819	0.824	**1.000**	0.849
	BestAPriori	0.395	0.669	0.400	0.200	0.328	0.290	0.524	0.809	0.569	0.254	0.533	0.333
	Bandari	0.910	0.964	**0.920**	0.933	0.974	0.943	**0.976**	0.997	0.955	0.951	0.993	0.927
	Official	0.894	0.964	0.891	0.878	0.918	0.880	0.971	0.992	**0.969**	0.942	0.989	0.932
Google+	Time	0.147	0.501	0.152	0.440	0.825	0.475	0.532	0.880	0.483	0.194	0.559	0.196
	Live	0.883	0.824	0.898	0.902	0.951	**0.918**	**0.899**	0.981	**0.904**	0.927	0.918	**0.924**
	Source	0.825	0.865	0.818	**0.926**	0.981	0.894	0.753	0.896	0.769	0.919	0.933	0.903
	ConstScale	0.883	0.824	0.898	0.902	0.951	**0.918**	**0.899**	0.981	**0.904**	0.927	0.918	0.923
	Linear-log	0.274	0.732	0.407	0.713	0.958	0.703	0.618	0.963	0.642	0.593	0.933	0.672
	LM	0.531	0.940	0.584	0.813	0.995	0.780	0.770	**1.000**	0.748	0.763	**0.954**	0.770
	ML	0.590	0.923	0.626	0.748	0.936	0.736	0.753	0.981	0.729	0.807	0.863	0.821
	MRBF	0.589	0.924	0.625	0.724	0.921	0.715	0.674	0.885	0.674	0.812	0.869	0.824
	kernel	0.618	0.964	0.633	0.852	0.995	0.815	0.783	**1.000**	0.757	0.810	0.919	0.815
	ENSt_kernel	0.375	0.819	0.411	0.580	0.768	0.639	0.599	0.973	0.628	0.582	0.937	0.600
	ENSphi_kernel	0.824	**0.965**	0.784	0.918	**0.997**	0.883	0.884	**1.000**	0.858	0.853	0.919	0.848
	knn	0.365	0.802	0.404	0.567	0.760	0.634	0.614	0.958	0.637	0.582	0.937	0.600
	ENSt_knn	0.222	0.640	0.325	0.703	0.984	0.668	0.586	0.901	0.580	0.542	0.843	0.560
	ENSphi_knn	0.198	0.570	0.309	0.633	0.995	0.648	0.603	0.957	0.604	0.536	0.846	0.599
	BestAPriori	0.191	0.572	0.242	0.180	0.393	0.251	0.310	0.625	0.344	0.155	0.391	0.213
	Bandari	0.686	0.938	0.680	0.841	0.995	0.820	0.795	**1.000**	0.767	0.833	0.919	0.835
	Official	**0.887**	0.831	**0.901**	0.904	0.951	**0.918**	0.898	0.981	0.903	**0.928**	0.918	0.923

Continued on the next page

Table 8 – continued from the previous page

		Economy			Microsoft			Obama			Palestine		
	Approach	*MAP*	*MRR*	*NDCG*@10	*MAP*	*MRR*	*NDCG*@10	*MAP*	*MRR*	*NDCG*@10	*MAP*	*MRR*	*NDCG*@10
LinkedIn	Time	0.434	0.843	0.411	0.446	0.797	0.491	0.464	0.786	0.431	0.255	0.628	0.276
	Live	0.959	0.989	0.961	0.945	0.946	0.954	**0.958**	0.990	**0.959**	**0.752**	0.980	**0.674**
	Source	0.928	0.978	0.917	0.920	0.968	0.898	0.849	0.988	0.838	0.694	0.918	0.634
	ConstScale	0.960	0.989	0.961	0.945	0.946	0.954	**0.958**	0.990	**0.959**	0.741	0.961	0.664
	Linear-log	0.836	**1.000**	0.811	0.835	**1.000**	0.812	0.639	0.992	0.642	0.545	0.891	0.532
	LM	0.757	0.957	0.762	0.859	**1.000**	0.841	0.862	**1.000**	0.831	0.636	0.982	0.595
	ML	0.609	0.721	0.629	0.715	0.896	0.723	0.794	0.993	0.779	0.724	**0.998**	0.642
	MRBF	0.616	0.744	0.634	0.743	0.925	0.737	0.754	0.959	0.752	0.717	0.997	0.636
	kernel	0.940	0.994	0.905	0.951	**1.000**	0.929	0.814	**1.000**	0.797	0.691	0.976	0.616
	ENSt_kernel	0.729	0.993	0.732	0.722	0.994	0.722	0.533	0.979	0.581	0.433	0.805	0.409
	ENSphi_kernel	0.949	0.992	0.931	**0.975**	**1.000**	**0.964**	0.884	**1.000**	0.860	0.720	0.976	0.635
	knn	0.680	0.982	0.696	0.686	0.978	0.705	0.547	0.976	0.591	0.421	0.805	0.402
	ENSt_knn	0.731	0.983	0.709	0.860	**1.000**	0.835	0.659	**1.000**	0.586	0.409	0.759	0.411
	ENSphi_knn	0.649	0.960	0.681	0.749	0.996	0.745	0.575	0.951	0.591	0.308	0.554	0.329
	BestAPriori	0.366	0.653	0.406	0.215	0.395	0.285	0.500	0.856	0.451	0.135	0.302	0.220
	Bandari	0.957	0.996	0.927	0.952	**1.000**	0.934	0.862	**1.000**	0.834	0.707	0.974	0.630
	Official	**0.965**	0.992	**0.963**	0.950	0.946	0.959	**0.958**	0.990	**0.959**	0.744	0.974	0.668

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, June 2005.

[2] Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 21–30, New York, NY, USA, 2009. ACM.

[3] Charu C. Aggarwal. *Outlier Analysis*. Springer Publishing Company, Incorporated, 2013.

[4] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering – a decade review. *Information Systems*, 53:16 – 38, 2015.

[5] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, January 1991.

[6] Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 607–616, New York, NY, USA, 2013. ACM.

[7] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz. Applying support vector machines to imbalanced datasets. In *Proceedings of the 15th European Conference on Machine Learning*, ECML 2004, pages 39–50, 2004.

[8] Oguz Akbilgic, Hamparsum Bozdogan, and M. Erdal Balaban. A novel hybrid RBF neural networks model as a forecaster. *Statistics and Computing*, 24(3):365–375, 2014.

[9] Hiroshi Akima. Algorithm 761: Scattered-data surface fitting that has the accuracy of a cubic polynomial. *ACM Trans. Math. Softw.*, 22(3):362–371, September 1996.

[10] Hiroshi Akima and Albrecht Gebhardt. *akima: Interpolation of Irregularly and Regularly Spaced Data*, 2015. R package version 0.5-12.

[11] Roberto Alejo, Vicente García, José Martínez Sotoca, Ramón Alberto Mollineda, and José Salvador Sánchez. Improving the performance of the RBF neural networks trained with imbalanced samples. In *Proceedings of the 9th International Work-Conference on Artificial Neural Networks*, IWANN 2007, pages 162–169, 2007.

[12] D.L. Altheide. *Qualitative Media Analysis*. Qualitative research methods. Sage Publications, 1996.

[13] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[14] Gianni Amati, Edgardo Ambrosi, Marco Bianchi, Carlo Gaibisso, and Giorgio Gambosi. Fub, IASI-CNR and university of tor vergata at TREC 2007 blog track. In *Proceedings of The Sixteenth Text REtrieval Conference*, TREC 2007, 2007.

[15] Ioannis Arapakis, B. Barla Cambazoglu, and Mounia Lalmas. On the feasibility of predicting news popularity at cold start. In Luca Maria Aiello and Daniel McFarland, editors, *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, pages 290–299. Springer International Publishing, 2014.

[16] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[17] Kenneth J. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346, 1950.

[18] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

[19] Josh Attenberg and Şeyda Ertekin. *Class Imbalance and Active Learning*, pages 101–149. John Wiley & Sons, Inc., 2013.

[20] A. Azaria, A. Richardson, S. Kraus, and V. S. Subrahmanian. Behavioral analysis of insider threat: A survey and bootstrapped prediction in imbalanced data. *IEEE Transactions on Computational Social Systems*, 1(2):135–155, 2014.

[21] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association, 2010.

[22] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. *CoRR*, abs/1111.4570, 2011.

[23] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[24] Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. The pulse of news in social media: Forecasting popularity. In *Proceedings of the Seventh International Conference on Weblogs and Social Media*. The AAAI Press, 2012.

[25] Albert-László Barabási. *Linked: The New Science of Networks*. Perseus Pub., 2002.

[26] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207, 2005.

[27] Ricardo Barandela, José Salvador Sánchez, Vicente García, and E. Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851, 2003.

[28] Luís Baía. Actionable Forecasting and Activity Monitoring: applications to financial trading. Master's thesis, University of Porto, College of Sciences, Portugal, 2015.

[29] Jonah Berger and Katherine L Milkman. What makes online content viral? *Journal of Marketing Research*, 49(2):192–205, 2012.

[30] Jerzy Blaszczynski and Jerzy Stefanowski. Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing*, 150:529–542, 2015.

[31] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447. Association for Computational Linguistics, June 2007.

[32] Léon Bottou. Stochastic learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, pages 146–168, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[33] Paula Branco. *Re-sampling Approaches for Regression Tasks under Imbalanced Domains*. PhD thesis, Universidade do Porto, 2014.

[34] Paula Branco, Rita P Ribeiro, and Luis Torgo. Ubl: an r package for utility-based learning. *arXiv preprint arXiv:1604.08079*, 2016.

[35] Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, 49(2):31:1–31:50, August 2016.

[36] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[37] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. {*FEBS*} *Letters*, 573(1–3):83 – 92, 2004.

[38] Peter J. Brockwell and Richard A. Davis. *Time series: theory and methods*. Springer Series in Statistics. Springer, New York (N.Y.), 1991.

[39] D.S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[40] Tine Buch-Larsen, Jens Perch Nielsen, Montserrat Guillén, and Catalina Bolancé. Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics*, 39(6):503–516, 2005.

[41] Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 211–223, New York, NY, USA, 2014. ACM.

[42] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the 10th International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[43] F. K. P. Chan, A. W. C. Fu, and C. Yu. Haar wavelets for efficient similarity search of time-series: with and without time warping. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):686–705, May 2003.

[44] Yi Chang, Xuanhui Wang, Qiaozhu Mei, and Yan Liu. Towards twitter context summarization with user influence models. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 527–536, New York, NY, USA, 2013. ACM.

[45] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, 2002.

[46] Chao Chen, Andy Liaw, and Leo Breiman. Using Random Forest to Learn Imbalanced Data. Technical report, Department of Statistics, University of Berkeley, 2004.

[47] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936, New York, NY, USA, 2014. ACM.

[48] Marc Cheong and Vincent Lee. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In *Proceedings of the 2nd ACM workshop on Social Web Search and Mining*, SWSM '09, pages 1–8, New York, NY, USA, 2009. ACM.

[49] Peter F. Christoffersen and Francis X. Diebold. Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics*, 11(5):561–571, 1996.

[50] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[51] William S. Cleveland. *Visualizing Data*. Hobart Press, 1993.

[52] D.R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[53] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585, December 2006.

[54] R Crane and D Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[55] Sven F. Crone, Stefan Lessmann, and Robert Stahlbock. Utility based data mining for time series analysis: Cost-sensitive learning for neural network predictors. In *Proceedings of the 1st International Workshop on Utility-based Data Mining*, UBDM '05, pages 59–68, New York, NY, USA, 2005. ACM.

[56] Easley David and Kleinberg Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[57] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proc. of 23rd ICML*, pages 233–240, New York, NY, USA, 2006.

[58] Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

[59] Munmun De Choudhury, Scott Counts, and Mary Czerwinski. Identifying relevant social media content: leveraging information diversity and user cognition. In

Proceedings of the 22nd ACM conference on Hypertext and hypermedia, HT '11, pages 161–170, New York, NY, USA, 2011. ACM.

[60] Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. From chatter to headlines: Harnessing the real-time web for personalized news recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 153–162. ACM, 2012.

[61] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[62] Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 155–164, New York, NY, USA, 1999. ACM.

[63] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the Essence: Improving Recency Ranking Using Twitter Data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 331–340, New York, NY, USA, 2010. ACM.

[64] Randall L. Dougherty, Alan Edelman, and James M. Hyman. Nonnegativity-, monotonicity-, or convexity-preserving cubic and quintic Hermite interpolation. *Mathematics of Computation*, 52(186):471–494, 1989.

[65] Harris Drucker, Chris, Burges* L. Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, volume 9, pages 155–161, 1997.

[66] Chris Drummond and Robert C. Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Workshop on Learning from Imbalanced Datasets II*, volume 11, pages 1–8. Citeseer, 2003.

[67] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. A Wiley Interscience Publication. Wiley, 1973.

[68] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 613–622. ACM, 2001.

[69] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation revisited, 2001.

[70] Miles Efron. Information search and retrieval in microblogs. *J. Am. Soc. Inf. Sci. Technol.*, 62(6):996–1008, June 2011.

[71] J. P. Egan. *Signal detection theory and ROC analysis*. Series in Cognition and Perception. Academic Press, New York, NY, 1975.

[72] Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, pages 973–978, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[73] Seyda Ertekin. Adaptive oversampling for imbalanced data classification. In *Proceedings of the 28th International Symposium on Computer and Information Sciences*, ISCIS 2013, pages 261–269, 2013.

[74] Seyda Ertekin, Jian Huang, Léon Bottou, and C. Lee Giles. Learning on the border: active learning in imbalanced data classification. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, CIKM 2007, pages 127–136, 2007.

[75] Andrew Estabrooks and Nathalie Japkowicz. A mixture-of-experts framework for learning from imbalanced data sets. In Frank Hoffmann, David J. Hand, Niall Adams, Douglas Fisher, and Gabriela Guimaraes, editors, *Advances in Intelligent Data Analysis: 4th International Conference, IDA 2001 Cascais, Portugal, September 13–15, 2001 Proceedings*, pages 34–43, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[76] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Computational Intelligence*, 20(1):18–36, 2004.

[77] Facebook. Facebook - statistics, 2016. [Online; accessed 23-January-2017].

[78] Tom Fawcett and Foster Provost. Activity monitoring: Noticing interesting changes in behavior. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 53–62, New York, NY, USA, 1999. ACM.

[79] I. Feinerer, K. Hornik, and D. Meyer. Text mining infrastructure in r. *Journal of Statistical Software*, 5(25):1–54, 2008.

[80] Alberto Fernández, Salvador García, María José del Jesus, and Francisco Herrera. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets Syst.*, 159(18):2378–2398, 2008.

[81] Flavio Figueiredo, Jussara M. Almeida, Marcos André Gonçalves, and Fabrício Benevenuto. On the dynamics of social media popularity: A youtube case study. *ACM Trans. Internet Technol.*, 14(4):24:1–24:23, December 2014.

[82] Freemeteo. http://freemeteo.com.pt/. Accessed: 2017-03-30.

[83] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997.

[84] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315:2007, 2007.

[85] C. Friedman and S. Sandow. *Utility-Based Learning from Data*. Chapman & Hall/CRC machine learning & pattern recognition series. CRC Press, 2011.

[86] Michael Gamon. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[87] Shuai Gao, Jun Ma, and Zhumin Chen. Modeling and predicting retweeting dynamics on microblogging platforms. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 107–116, New York, NY, USA, 2015. ACM.

[88] Xingyu Gao, Zhenyu Chen, Sheng Tang, Yongdong Zhang, and Jintao Li. Adaptive weighted imbalance learning with application to abnormal activity recognition. *Neurocomputing*, 173(P3):1927–1935, January 2016.

[89] Zan Gao, Long-fei Zhang, Ming-yu Chen, Alexander Hauptmann, Hua Zhang, and An-Ni Cai. Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset. *Multimedia Tools and Applications*, 68(3):641–657, 2014.

[90] V. García, J. S. Sánchez, and R. A. Mollineda. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21, February 2012.

[91] Florent Garcin, Christos Dimitrakakis, and Boi Faltings. Personalized news recommendation with context trees. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys'13, pages 105–112, New York, NY, USA, 2013. ACM.

[92] Lorenzo Gatti and Marco Guerini. Assessing sentiment strength in words prior polarities. *CoRR*, abs/1212.4315, 2012.

[93] S. Geisser. *Predictive Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1993.

[94] Colin S. Gillespie. Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*, 64(2):1–16, 2015.

[95] M. Gupta, J. Gao, C. Zhai, and J. Han. Predicting future popularity trend of events in microblogging platforms. In Andrew Grove, editor, *ASIS&T 75th Annual Meeting*, 2012.

[96] Gonca Gürsun, Mark Crovella, and Ibrahim Matta. Describing and forecasting video access patterns. In *INFOCOM*, pages 16–20. IEEE, 2011.

[97] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[98] Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[99] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

[100] D. M. Hawkins. *Identification of outliers*. Monographs on applied probability and statistics. Chapman and Hall, 1980.

[101] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284, September 2009.

[102] Urban Hjorth. Model selection and forward validation. *Scandinavian Journal of Statistics*, 9:95–105, 1982.

[103] Tad Hogg and Kristina Lerman. Stochastic models of user-contributory web sites. In *Proceedings of 3rd International Conference on Weblogs and Social Media (ICWSM)*, May 2009.

[104] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 57–58, New York, NY, USA, 2011. ACM.

[105] Kurt Hornik. *openNLP: Apache OpenNLP Tools Interface*, 2016. R package version 0.2-6.

[106] Chiao-Fang Hsu, E. Khabiri, and J. Caverlee. Ranking comments on the social web. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, volume 4 of *CSE '09*, pages 90–97, Washington, DC, USA, 2009. IEEE.

[107] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177. ACM, 2004.

[108] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 173–178, New York, NY, USA, 2010. ACM.

[109] Salman Jamali and Huzefa Rangwala. Digging digg: Comment mining, popularity prediction, and social network analysis. In *Proceedings of the 2009 International Conference on Web Information Systems and Mining*, WISM '09, pages 32–38, Washington, DC, USA, 2009. IEEE Computer Society.

[110] Nathalie Japkowicz. Learning from Imbalanced Data Sets: A Comparison of Various Strategies. In *AAAI Workshop on Learning from Imbalanced Data Sets*, volume 68, 2000.

[111] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proc. of the 23rd ACM SIGIR*, pages 41–48, 2000.

[112] Borda JC. Memoire sur les elections au scrutin, 1781.

[113] Taeho Jo and Nathalie Japkowicz. Class imbalances versus small disjuncts. *SIGKDD Explor. Newsl.*, 6(1):40–49, 2004.

[114] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings*, pages 137–142, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[115] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.

[116] Mahesh V. Joshi, Vipin Kumar, and Ramesh C. Agarwal. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM 2001, pages 257–264, 2001.

[117] Charles M. Judd and Gary H. McClelland. *Data analysis: A model-comparison approach.* Harcourt Brace Jovanovich Inc, New York, 1989.

[118] Andreas Kaltenbrunner, Vicenc Gomez, and Vicente Lopez. Description and Prediction of Slashdot Activity. In *LA-WEB '07: Proceedings of the 2007 Latin American Web Conference*, pages 57–66, Washington, DC, USA, 2007. IEEE Computer Society.

[119] Andreas Kaltenbrunner, Vicenç Gómez, Ayman Moghnieh, Rodrigo Meza, Josep Blat, and Vicente López. Homogeneous temporal activity patterns in a large online communication space. *CoRR*, abs/0708.1579, 2007.

[120] Frigyes Karinthy. *Chains*. unknown, 1929.

[121] Y. Keneshloo, S. Wang, E. H. S. Han, and N. Ramakrishnan. Predicting the shape and peak time of news article views. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2400–2409, Dec 2016.

[122] Elham Khabiri, Chiao-Fang Hsu, and James Caverlee. Analyzing and predicting community preference of socially generated metadata: A case study on comments in the digg community. In *Proceedings of the Third International Conference on Weblogs and Social Media*, ICWSM 2009, 2009.

[123] Su-Do Kim, Sung-Hwan Kim, and Hwan-Gue Cho. Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity. In *Computer and Information Technology (CIT), 2011 IEEE 11th International Conference on*, pages 449–454. IEEE, 2011.

[124] Ryota Kobayashi and Renaud Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, pages 191–200, 2016.

[125] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573, 2012.

[126] Shoubin Kong, Qiaozhu Mei, Ling Feng, Fei Ye, and Zhe Zhao. Predicting bursts and popularity of hashtags in real-time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 927–930, New York, NY, USA, 2014. ACM.

[127] I. Koprinska, M. Rana, and V.G. Agelidis. Yearly and seasonal models for electricity load forecasting. In *Proc. of 2011 IJCNN*, pages 1474–1481, July 2011.

[128] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[129] Bartosz Krawczyk, Mikel Galar, Lukasz Jeleń, and Francisco Herrera. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing*, 38(C):714–726, January 2016.

[130] Bartosz Krawczyk, Leandro L. Minku, Joo Gama, Jerzy Stefanowski, and Micha Woniak. Ensemble learning for data stream analysis. *Inf. Fusion*, 37(C):132–156, September 2017.

[131] Bartosz Krawczyk, Michal Wozniak, and Gerald Schaefer. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Appl. Soft Comput.*, 14:554–562, 2014.

[132] D.G. Krige. A Statistical Approach to Some Mine Valuations and Allied Problems at the Witwatersrand. Master's thesis, University of Witwatersrand, South Africa, 1951.

[133] Andrey Kupavskii, Alexey Umnov, Gleb Gusev, and Pavel Serdyukov. Predicting the audience size of a tweet. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *ICWSM*. The AAAI Press, 2013.

[134] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[135] Himabindu Lakkaraju and Jitendra Ajmera. Attention prediction on social media brand pages. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 2157–2160, 2011.

[136] Chei Sian Lee and Long Ma. News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behaviour*, 28(2):331–339, 2012.

[137] Jong G. Lee, Sue Moon, and Kave Salamatian. An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors. In *IEEE Conference on Web Intelligence*, 2010.

[138] Seungyong Lee, George Wolberg, and Sung Yong Shin. Scattered data interpolation with multilevel b-splines. *IEEE Transactions on Visualization and Computer Graphics*, 3(3):228–244, July 1997.

[139] Kristina Lerman. Social information processing in news aggregation. *IEEE Internet Computing*, 11(6):16–28, 2007.

[140] Kristina Lerman and Aram Galstyan. Analysis of social voting patterns on digg. In *Proceedings of the First Workshop on Online Social Networks*, WOSN '08, pages 7–12, New York, NY, USA, 2008. ACM.

[141] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In William W. Cohen and Samuel Gosling, editors, *ICWSM*. The AAAI Press, 2010.

[142] Kristina Lerman and Tad Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 621–630, New York, NY, USA, 2010. ACM.

[143] Ho Leung Li and Vincent T. Y. Ng. Discovering associations between news and contents in social network sites with the d-miner service framework. *J. Netw. Comput. Appl.*, 36(6):1651–1659, November 2013.

[144] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[145] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):1019–1031, May 2007.

[146] E Limpert, WA Stahel, and M Abby. Log-normal distributions across the sciences: Keys and clues. *Bioscience*, 51(5):341–352, 2001.

[147] Shili Lin. Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):555–570, 2010.

[148] Shili Lin. Space oriented rank-based data integration. *Statistical Applications in Genetics and Molecular Biology*, 9(20), 2010.

[149] Shili Lin and Jie Ding. Integration of ranked lists via cross entropy monte carlo with applications to mrna and microrna studies. *Biometrics*, 65 1:9–18, 2009.

[150] Charles X. Ling and Victor S. Sheng. Class imbalance problem. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, page 171. Springer, 2010.

[151] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463. Springer, 2012.

[152] Huan Liu, Fred Morstatter, Jiliang Tang, and Reza Zafarani. The good, the bad, and the ugly: uncovering novel research opportunities in social media mining. *International Journal of Data Science and Analytics*, 1(3):137–143, 2016.

[153] Xiaoming Liu, Johan Bollen, Michael L. Nelson, and Herbert Van de Sompel. Co-authorship networks in the digital library research community. *Inf. Process. Manage.*, 41(6):1462–1480, December 2005.

[154] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Systems, Man, and Cybernetics, Part B*, 39(2):539–550, 2009.

[155] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013.

[156] Zhunchen Luo, Miles Osborne, Sasa Petrovic, and Ting Wang. Improving twitter retrieval by exploiting structural information. In Jörg Hoffmann and Bart Selman,

editors, *Proceedings of the 26th AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.* AAAI Press, 2012.

[157] Ilias N. Lymperopoulos. Predicting the popularity growth of online content: Model and algorithm. *Information Sciences*, 369:585 – 613, 2016.

[158] Craig Macdonald, Iadh Ounis, and Ian Soboroff. Overview of the TREC 2009 blog track. In *Proceedings of The Eighteenth Text REtrieval Conference*, TREC 2009, 2009.

[159] M. A. Maloof. Learning when data sets are Imbalanced and when costs are unequal and unknown. In *ICML-2003 Workshop on Learning from Imbalanced Datasets II*, volume 2, 2003.

[160] Agostino Manduchi and Robert Picard. Circulations, revenues, and profits in a newspaper market with fixed advertising costs. *Journal of Media Economics*, 22(4):211–238, 2009.

[161] Travis Martin, Jake M. Hofman, Amit Sharma, Ashton Anderson, and Duncan J. Watts. Exploring limits to prediction in complex social systems. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 683–694, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[162] Kamran Massoudi, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 362–367, Berlin, Heidelberg, 2011. Springer-Verlag.

[163] Yasuko Matsubara, Yasushi Sakurai, B. Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: Model and implications. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 6–14, New York, NY, USA, 2012. ACM.

[164] Andrew McCallum and Kamal Nigam. A comparison of event models for naive Bayes text classification. In *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, pages 41–48, 1998.

[165] Richard M. C. McCreadie, Craig Macdonald, and Iadh Ounis. News article ranking: Leveraging the wisdom of bloggers. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 40–48, Paris, France, France, 2010. Le Centre de Hautes Etudes Internationales D'Informatique Documentaire.

[166] Charles E. Metz. Ce: Basic principles of roc analysis. In *Seminars in Nuclear Medicine*, volume 8, pages 283–298, 1978.

[167] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2012. R package version 1.6-1.

[168] S. Milborrow. *earth: Multivariate Adaptive Regression Spline Models*, 2013. R package version 3.2-6.

[169] Stanley Milgram. The Small World Problem. *Psychology Today*, 2:60–67, 1967.

[170] Gilad Mishne and Maarten de Rijke. A study of blog search. In *Proceedings of the 28th European Conference on Advances in Information Retrieval*, ECIR'06, pages 289–301, Berlin, Heidelberg, 2006. Springer-Verlag.

[171] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

[172] Tsendsuren Munkhdalai, Oyun-Erdene Namsrai, and Keun Ho Ryu. Self-training in significance space of support vectors for imbalanced biomedical event data. *BMC Bioinformatics*, 16(7):S6, 2015.

[173] E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[174] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 8:1–8:7, New York, NY, USA, 2011. ACM.

[175] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[176] Nic Newman, Richard Fletcher, David A. L. Levy, and Rasmus Kleis Nielsen. Reuters institute digital news report 2016. Technical report, Reuters Institute for the Study of Journal (University of Oxford), 2016.

[177] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, Heraklion, Crete, Greece, May 30, 2011*, pages 93–98, 2011.

[178] Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In William W. Cohen and Samuel Gosling, editors, *Proceedings of the 2010 ICWSM*. The AAAI Press, 2010.

[179] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke. Predicting imdb movie ratings using social media. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, pages 503–507, Berlin, Heidelberg, 2012. Springer-Verlag.

[180] Albert Orriols-Puig and Ester Bernadó-Mansilla. Evolutionary rule-based systems for imbalanced data sets. *Soft Computing*, 13(3):213, 2008.

[181] M. Osborne and M. Dredze. Facebook, twitter and google plus for breaking news: Is there a winner? In *Proc. of the 8th ICWSM 2014*, 2014.

[182] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[183] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895.

[184] Robin Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4:1–79, 2007.

[185] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.

[186] Sasa Petrovic, Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimpton. Can twitter replace newswire for breaking news? In *Proc. of 7th ICWSM*. The AAAI Press, 2013.

[187] Clifton Phua, Damminda Alahakoon, and Vincent Lee. Minority report in fraud detection: Classification of skewed data. *SIGKDD Explor. Newsl.*, 6(1):50–59, 2004.

[188] R.G. Picard. Commercialism and Newspaper Quality. *Newspaper Research Journal*, 25(1):54–66, 2004.

[189] James III Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975.

[190] R. Pincus. Barnett, v., and lewis t.: Outliers in statistical data (3rd edition). *Biometrical Journal*, 37(2):256–256, 1995.

[191] Henrique Pinto, Jussara M. Almeida, and Marcos A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 365–374, New York, NY, USA, 2013. ACM.

[192] Foster Provost and Tom Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 43–48. AAAI Press, 1997.

[193] Foster J. Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 445–453, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[194] Hemant Purohit, Yiye Ruan, Amruta Joshi, Srinivasan Parthasarathy, and Amit Sheth. Understanding user-community engagement by multi-faceted features: A case study on twitter. In *Social Media Engagement Workshop, WWW-2011*, 2011.

[195] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[196] R. Ribeiro. *Utility-based Regression*. PhD thesis, Dep. Computer Science, Faculty of Sciences - University of Porto, 2011.

[197] Rita Ribeiro and Luis Torgo. Predicting harmful algae blooms. In *Proceedings of the 2003 Portuguese Conference on Artificial Intelligence*, EPIA 2003, pages 308–312. Springer Berlin Heidelberg, 2003.

[198] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[199] Tyler W. Rinker. *qdap: Quantitative Discourse Analysis Package*. University at Buffalo/SUNY, Buffalo, New York, 2013. 2.2.5.

[200] Suman Deb Roy, Tao Mei, Wenjun Zeng, and Shipeng Li. Towards cross-domain learning for social video popularity prediction. *IEEE Trans. Multimedia*, 15(6):1255–1267, 2013.

[201] Reuven Y. Rubinstein and Dirk P. Kroese. *The Cross Entropy Method: A Unified Approach To Combinatorial Optimization, Monte-carlo Simulation*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2004.

[202] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[203] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *Proc. of the 33rd ACM SIGIR*, pages 555–562, 2010.

[204] Michael G. Schimek and Marcus Bloice. Modelling the rank order of web search engine results. In *Proceedings of the 27th International Workshop on Statistical Modelling*, IWSM'12, pages 303–308. Statistical Modelling Society, 2012.

[205] Chao Shen, Fei Liu, Fuliang Weng, and Tao Li. A participant-based approach for event summarization using twitter streams. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, NAACL-HLT '13, pages 1152–1162. The Association for Computational Linguistics, 2013.

[206] Hua-Wei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 291–297, 2014.

[207] Benjamin Shulman, Amit Sharma, and Dan Cosley. Predictability of popularity: Gaps between prediction and understanding. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, pages 348–357, 2016.

[208] Mikhail V. Simkin and Vwani P. Roychowdhury. Why does attention to web articles fall with time? *Journal of the Association for Information Science & Technology*, 66(9):1847–1856, 2015.

[209] Daniel B. Stouffer, R. Dean Malmgren, and Luis A. N. Amaral. Log-normal statistics in e-mail communication patterns. *CoRR*, 2006.

[210] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proc. of the 2nd IEEE SOCIALCOM*, pages 177–184, DC, USA, 2010. IEEE.

[211] Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719, 2009.

[212] Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, August 2010.

[213] Leonard J. Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000.

[214] Alexandru Tatar. *Predicting User-Centric Behavior: Mobility and Content Popularity.* PhD thesis, UPMC Sorbonne Universités - Paris VI, 2014.

[215] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida. Ranking news articles based on popularity prediction. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 106–110, Washington, DC, USA, 2012. IEEE Computer Society.

[216] Alexandru Tatar, Marcelo Dias de Amorim, Serge Fdida, and Panayotis Antoniadis. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5(1), 2014.

[217] Alexandru Tatar, Jérémie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo Dias de Amorim, and Serge Fdida. Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 67:1–67:8, New York, NY, USA, 2011. ACM.

[218] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December 2010.

[219] L. Torgo. An infra-structure for performance estimation and experimental comparison of predictive models in r. *CoRR*, abs/1412.0436, 2014.

[220] L. Torgo. An infra-structure for performance estimation and experimental comparison of predictive models in r. *CoRR*, abs/1412.0436, 2014.

[221] Luis Torgo and Rita Ribeiro. Predicting outliers. In *Proceedings of the 2003 European Conference on Principles of Data Mining and Knowledge Discovery*, ECML 2003, pages 447–458. Springer Berlin Heidelberg, 2003.

[222] Luís Torgo, Paula Branco, Rita P. Ribeiro, and Bernhard Pfahringer. Resampling strategies for regression. *Expert Systems*, 32(3):465–476, 2015.

[223] Manos Tsagkias. *Mining Social Media: Tracking Content and Predicting Behavior*. PhD thesis, Faculteit der Natuurwetenschappen, Wiskunde en Informatica, 2012.

[224] Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1765–1768, New York, NY, USA, 2009. ACM.

[225] Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. News comments:exploring, modeling, and online prediction. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan Rüger, and Keith van Rijsbergen, editors, *Advances in Information Retrieval: 32nd European Conference*

on IR Research, ECIR'2010, pages 191–203, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[226] Oren Tsur and Ari Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 643–652, New York, NY, USA, 2012. ACM.

[227] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[228] Twitter. Twitter - statistics, 2016. [Online; accessed 23-January-2017].

[229] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011.

[230] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[231] Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–33, 1989.

[232] Benjamin X. Wang and Nathalie Japkowicz. Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, 25(1):1–20, 2010.

[233] Geoffrey S. Watson. Smooth regression analysis. *The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

[234] D.J. Watts. *Six Degrees: The Science of a Connected Age*. W. W. Norton, 2004.

[235] Geoffrey I. Webb. Decision tree grafting. In *Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence - Volume 2*, IJCAI'97, pages 846–851, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[236] Gary Weiss. Foundations of imbalanced learning. In Haibo He and Yunqian Ma, editors, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 1st edition, 2013.

[237] Gary M. Weiss, Maytal Saar-Tsechansky, and Bianca Zadrozny. Report on UBDM-05: workshop on utility-based data mining. *SIGKDD Explorations*, 7(2):145–147, 2005.

[238] Gary M. Weiss, Bianca Zadrozny, and Maytal Saar-Tsechansky. Guest editorial: special issue on utility-based data mining. *Data Mining and Knowledge Discovery*, 17(2):129–135, 2008.

[239] L. Weng, F. Menczer, and Y.-Y. Ahn. Virality prediction and community structure in social networks. *Scientific Reports*, 3(2522), 2013.

[240] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[241] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 272–278. AAAI Press, 2016.

[242] F. Wu and B. A. Huberman. Novelty and collective attention. *Proc. Natl. Acad. Sci. USA*, 104(45):17599–17601, 2007.

[243] Gang Wu and Edward Y. Chang. Kba: Kernel boundary alignment considering imbalanced data distribution. *IEEE Transactions on Knowledge & Data Engineering*, 17:786–795, 2005.

[244] Ruifeng Xu, Tao Chen, Yunqing Xia, Qin Lu, Bin Liu, and Xuan Wang. Word embedding composition for data imbalances in sentiment and emotion classification. *Cognitive Computation*, 7(2):226–240, 2015.

[245] Wei Xu, Ralph Grishman, Adam Meyers, and Alan Ritter. A preliminary study of tweet summarization using information extraction. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 20–29, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[246] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 177–186, New York, NY, USA, 2011. ACM.

[247] Qiang Yang and Xindong Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making (IJITDM)*, 05(04):597–604, 2006.

[248] Tae Yano, William W Cohen, and Noah A Smith. Predicting response to political blog posts with topic models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 477–485. Association for Computational Linguistics, 2009.

[249] Peifeng Yin, Ping Luo, Min Wang, and Wang-Chien Lee. A straw shows which way the wind blows: Ranking potentially popular items from early votes. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 623–632, New York, NY, USA, 2012. ACM.

[250] Youtube. Youtube - statistics, 2016. [Online; accessed 23-January-2017].

[251] Linyun Yu, Peng Cui, Fei Wang, Chaoming Song, and Shiqiang Yang. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics. In Charu Aggarwal, Zhi-Hua Zhou, Alexander Tuzhilin, Hui Xiong, and Xindong Wu, editors, *ICDM*, pages 559–568. IEEE Computer Society, 2015.

[252] Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, ICDM 2003, page 435. IEEE Computer Society, 2003.

[253] Bianca Zadrozny, Gary M. Weiss, and Maytal Saar-Tsechansky. UBDM 2006: Utility-based data mining 2006 workshop report. *SIGKDD Explorations*, 8(2):98–101, 2006.

[254] Tauhid Zaman, Emily B. Fox, and Eric T. Bradlow. A bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics*, 8(3):1583–1611, 2014.

[255] Wenbin Zhang and Steven Skiena. Improving movie gross prediction through news analysis. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, pages 301–304, Washington, DC, USA, 2009. IEEE Computer Society.

[256] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams, editors, *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522. ACM, 2015.

[257] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.

[258] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition, 2012.

[259] Jingbo Zhu. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of 45th 2007 Annual Meeting of the Association for Computational Linguistics*, ACL 2007, pages 783–790, 2007.

[260] X. Zhu and I. Davidson. *Knowledge discovery and data mining: challenges and realities*. Premier reference source. Information Science Reference, 2007.

[261] Águas do Douro e Paiva. `http://addp.pt/`. Accessed: 2017-03-30.

[262] O Özgöbek, Jon Atle Gulla, and R Cenk Erdur. A survey on challenges and methods in news recommendation. *Proc. of 10th WEBIST*, 2014.