Graph-based Topic Extraction Using Centroid Distance of Phrase Embeddings on Healthy Aging Open-ended Survey Questions

Dijana Kosmajac*, Kirstie Smith†, Vlado Keselj‡ and Susan Kirkland§
*‡Faculty of Computer Science, †§Department of Community Health and Epidemiology, Dalhousie University
Halifax, NS, Canada
Email: *dijana.kosmajac@dal.ca, †kirstie.smith@dal.ca, ‡vlado@dnlp.ca, §susan.kirkland@dal.ca

***Abstract*—Open-ended questions are a very important part of research surveys. However, they can pose a challenge when it comes to processing since manual processing requires a labour-intensive human effort. Automation of the task requires application of NLP methods since free text does not ensure standardized structure. To tackle this problem, we present a solution for topic discovery and analysis of open-ended survey items. We use graph-based representation of the text that adds structure and enables easier manipulation and keyphrase retrieval. Additionally, we use pre-trained *fastText* aligned word vectors to cluster similar phrases even if they are written in different languages. The goal is to produce topic word and phrase representatives that are easy to interpret by a domain expert. We compare the method with traditional LDA and two state-of-the-art algorithms: BTM and WNTM. The resulting keyphrases representing topics are more intuitive to the domain experts than the ones obtained by reference topic models in similar experimental settings.**

***Index Terms*—Topic Modelling, LDA, healthy aging, short texts, open-ended survey responses**

I. Introduction

Survey research is a very common approach when it comes to gaining insights into a research subject. For example, it is used in different domains, such as health and health services [1], marketing and consumer analysis [2], [3], but it originated in social sciences. Although the survey data is collected using a standardized form, open-ended questions (OE) can be part of it. Its primary role is to to clarify ambiguities and provide explanations and potentially identify opinions that researchers did not include in the standardized form [4], [5]. Another important point to mention is that OE questions expand the capability of the survey to capture spontaneous thoughts, sentiments and attitudes. This is useful in marketing research where companies can measure consumers' attitude towards their products.

Nonetheless, processing such questions requires great human effort. Because of the nature of OE questions, the standard approach in identifying the topics requires researchers to go through all the answers and label them manually. This may not be a challenge for smaller studies, but in the case of tens of thousands of samples the task can take a lot of resources to accomplish. If the data is labeled by multiple researchers, the process is prone to errors, which is usually measured with between-rater variance [6], [7]. An important challenge in automated processing of the OE answers is that the texts are relatively short. Extracting topics from short texts is difficult because most of the traditional methods rely on word co-occurrence, which assumes that the related words occur together relatively frequently, and this is not a reasonable assumption in the sparse data collections such as survey answers [8].

In this study our focus is on a survey from Canadian Longitudinal Study on Aging (CLSA) conducted on over 50,000 older adults living in the 10 Canadian provinces. The survey responses were collected in two official languages of Canada: French and English. Demographic forecasts indicate that Canada's population is aging and the demographic structure will change dramatically over the next two decades. The numbers show that 25% of the population will be over 65 by 2036, almost double compared to 2009 [9]. The consequences of the demographic shift are among Canada's most pressing health and social policy issues. To put it into perspective, the total health and social care expenditures in Canada now exceed $300 billion with healthcare alone at approximately $211 billion, the largest expenditure item in provincial budgets [10]. Optimizing population health and wellness over the trajectory of aging — i.e. optimizing "healthy aging" — is therefore a major research and policy goal in Canada [11]. Therefore, we are analyzing the answers on the following OE question: "What do you think makes people live long and keep well?"

The aim of this study is to analyze open-ended survey responses by applying a combination of Information Retrieval (IR) and unsupervised Machine Learning (ML) techniques to discover the potential differences among certain subgroups, including gender, age, and presence of health conditions. We describe an interesting solution in a form of framework for group profiling based on difference in opinions (that is, topics) and compare it with state-of-the-art probabilistic topic modeling approaches. Our goal is to extract the topic-representative keyphrases that are more intuitive for topic labeling by the domain expert by introducing part-of-speech information, as well as semantic relatedness in a form of word embeddings.

II. Related Work

Domain of the topic identification refers to tasks of finding semantically meaningful topics from a document corpus. The base assumption says that there are hidden variables (topics) which describe the similarities between observable variables (that is, documents). Some of the most influential representatives of topic modeling methods are probabilistic Latent Semantic Indexing (pLSA) [12] and its generalization – Latent Dirichlet Allocation (LDA) [13]. LDA has been around for awhile, and has been applied to different domains, such as short [8], [14] and long texts [5], [15], genetic data [16], and images [17]. However, an LDA model in its original setup has a few shortcomings, especially when the target documents are short, or there are too many topics. This paper focuses on the former.

In the literature there are several studies on topic extraction from survey OE responses. The main characteristics of these texts are that they are short (usually between one and a few tens of words), not complete sentences, may or may not have punctuation, and prone to a degree of grammatical mistakes. In [4] authors propose a Structural Topic Model (STM) for topic discovery in OE responses. The main difference between traditional LDA and STM is that they include covariates of interest into the prior distributions for document-topic proportions and topic-word distributions. With this setup the result is a model where each OE response is a mixture of topics with incorporated prior knowledge about topical variance. Thorough experiments on topic modeling on OE responses were performed by [5] using two state-of-the-art algorithms (BTM, WNTM) [18], [19] and LDA as a baseline. They examine suitability of the automated algorithms to replace manual analysis and give some general recommendations for researchers and practitioners how to choose the right method for a given research task. They particularly chose the algorithms which are designed to address the issue of short documents. We conduct our comparative analysis with the same set of algorithms, hence the next few paragraphs are dedicated to description of same.

Biterm Topic Model (BTM) [18] is a model that does not use an external knowledge source to deal with the short documents or missing context as some other methods (LF-LDA). The main difference between BTM and LDA is that the input for it is not a set of documents D, but set of biterms B calculated on the corpus level. A biterm b represents a word pair that co-occurred in a specified short context window. Additionally, LDA uses the word co-occurrence pattern per document to generate words while BTM generates biterms.

Word Network Topic Model (WNTM) [19] is a recent model that infers topic distributions for words instead of documents to avoid the disadvantage of LDA with short texts. The core of the algorithm is word co-occurrence network which is created by moving a sliding window of length S through each document. The network nodes are the vocabulary of the corpus and the edges represent the co-occurrences of each word pair weighted by the number of co-occurrences in the corpus. In another words, for each word w_v a pseudo-document d_p is created that consists of all words that co-occur with w_v, i.e. all words that are direct neighbours of w_v in the word network. The generated pseudo-documents are used as input in WNTM.

An extension of traditional LDA is Latent Feature LDA (LF-LDA). It adresses the sparsity of short texts by using pre-trained word vector representations (Word2Vec [20] and GloVe [21]). In LF-LDA the generative process is similar to original LDA but differs in the way how words are generated from topics. In LDA, a word can only be drawn from the Dirichlet multinomial distribution ϕ that is trained on the target corpus while LF-LDA additionally allows draw from the multinomial distribution based on word vector representation of words and topics. This means that LF-LDA incorporates semantic knowledge from external corpora. They also introduce additional hyperparameter λ which determines the probability of word sampling from external latent feature component.

There are other variants of pseudo-document generation to improve short text topic modeling. Authors of the paper [22] use IR technique to cluster similar tweets in larger pseudo-documents. The process consists of three steps. The first step is preliminary set generation (set length n), where they cluster tweets based on cosine similarity. Second step is aggregation of similar preliminary sets into pooled set representation (of length m, where $m < n$), and this set is basically a set of pseudo-documents used for the next step. The final step is traditional LDA. They compare the results of different variants of the methodology with LDA trained on one document (whole dataset merged into one) and BTM. Another interesting work [23] proposes a general framework for addressing the issues with short text topic modeling. The build the model on BTM, WNTM and LF-LDA by applying an expansion procedure on each document. They describe two variants: co-frequency expansion (CoFE) and distributed representation-based expansion (DREx).

A significant amount of literature has studied transferring the probabilistic topic modeling concept from monolingual to multilingual settings [24]. In one of the early works [25] authors proposed an extension of standard pLSA to extract topics from cross-lingual datasets. They bridge the gap between different languages $(L_1, L_2, ...L_n)$ by introducing aligned dictionary. In this setting they define word distribution of a cross-lingual topic θ for language L_i as $p_i(w_i|\theta) = \frac{p(w_i|\theta)}{\sum_{w \in V_i} p(w|\theta)}$, where V_i is vocabulary of language L_i. These formulations are extension of the traditional maximum likelihood estimator to estimate parameters and discover cross-lingual topics. A more recent papers by authors [26] consider the topic modeling for multilingual datasets by training bilingual word embeddings. It is important to note that our approach is different in a sense that we are using pre-trained aligned word vectors due to the fact that the dataset presented in this study has a couple of limitations, such as size, length of documents and imbalance between the languages.

III. Methodology

Our approach consists of a number of steps towards building a set of phrase groups that represent meaningful topics. Unlike probabilistic topic modeling methods, the method relies on IR techniques and a ML unsupervised method – clustering. The intuition behind our approach is that IR methods can facilitate and speed up researcher's learning about the data by introducing the structure to the unstructured text documents. With the right data representation model, one can exploit the full power of other variables in the survey and get insights into possible correlations. We refer to this method as Graph-based Topic Clustering (GTC).

A. *Dataset*

CLSA is a study and national platform of adult development and aging individuals, each with unique experiences of their environments, communities, and health and social systems. The CLSA follows 50,000 Canadians between the ages of 45 and 85 years over a 20-year period. However, the data utilized in this paper come from the study baseline, collected between 2010-2015. CLSA is designed as a research platform with the aim to accelerate understanding of the complex interplay among the vast array of determinants of health, from gene-environment interactions, to lifestyles, social networks and transitions in retirement and wealth.

After applying the pre-processing tasks described further in the text the number of responses in English is 41,496 and 9,296 in French. To get a better understanding of the data, we conduct a simple statistical analysis. In English subset 24.60% of responses are in length range 1–3, 33.41% in 4-6, 20.44% in 7-9, 9.95% in 10-12 and 11.60% longer than 12 relative to total responses in English. In French subset 39.37% of responses are in length range 1–3, 38.02% in 4-6, 14.16% in 7–9 and 8.44% longer than 9 words. That means that more than a half of responses are shorter than 7 words.

1) Pre-processing: We conducted a couple of pre-processing tasks to decrease the noise in the dataset and to transform the data in such a way that it complies with the requirements of the methods for topic modeling. First, standard pre-processing techniques are performed, such as conversion to lowercase and the removal of numbers, punctuation [27]. Using [28] Stanford tool for French and English we tokenized and tagged the entire corpus. Lemmatization was performed using Spacy, and for the French language we used dictionary-based lemmatizer [29]. Although the dataset consists of English and French responses, we did not perform translation. For unsupervised spelling correction (unsupervised in a sense that we did not know which words are misspelled) contextual grammar correction [30] was used which relies on external Google 1T N-grams corpus. To identify the candidates for spell correction, we scanned through words that have frequency less than 5 and checked if they exist in FastText aligned word vectors [31] used later in the process. If the words do not exist it the FastText word vectors, they are flagged for spell correction. In general, the dataset did not contain many misspelings, and the number of flagged words is less than 200.

B. *Graph Representation of Text*

The dataset is represented by a directed graph $G = (V, E, C)$, where $V = w_1, w_2, ..., w_N$ is the set of nodes (i.e. vertices), each representing a word token. $E \subset \{(w_i, w_j) \mid w_i, w_j \in V\}$ is the set of edges between the vertices and it represents a direct neighbour connection between two word tokens. Each edge $e \in E$ is an ordered pair $e = (w_i, w_j)$ and is associated with a weight $we_{w_i,w_j} > 0$, which indicates the strength of the relation (frequency of the relation between two tokens in the dataset). Fig. 1 illustrates an example of graph representation of two sentences.

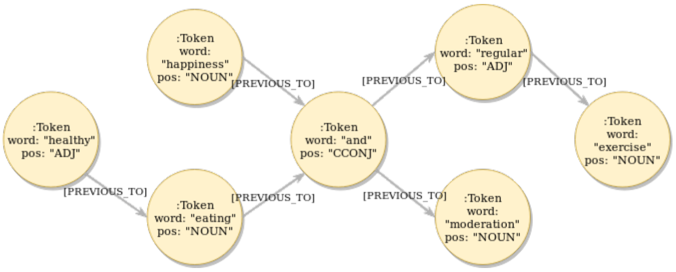

Fig. 1. Example — graph representation of two answers “healthy eating and regular exercise” and “happiness and moderation”

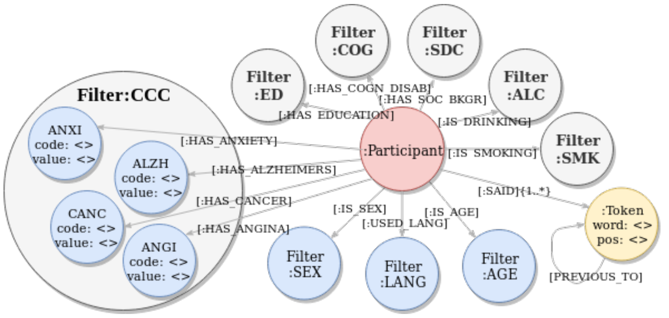

Fig. 2. Conceptual graph model of the survey dataset. Filters: SDC — socio-demographic characteristics; ED — education; COG — cognitive disabilities; CCC — health conditions; ALC — alcohol consumption; SMK — smoking.

In [32] authors use similar concept to represent a set of unstructured and short texts and perform summarization. Our work is different in couple of aspects. First, our goal is to extract characteristic keyphrases of 1–3 words length, while they try to capture longer common sequences of words. Second, each token is enriched with additional information such as lemma form and part-of-speech tag which are used in keyword extraction process. Third, the whole word graph is extended with other fields from the survey, such as participant id and other variables of interest. This makes it possible to reconstruct each participant's response to its original form. We used [33] graph database because of its powerful SQL-like declarative graph query language called Cypher and its accompanying graph-specific features. Fig. 2 shows the conceptual model of a part of CLSA survey (the survey itself is far more complex including over 300 variables) that is relevant to this study.

C. *Centroid of Phrase Word Embeddings*

To extract the word phrases consisting of one, two or three words we used the tag information. The only considered words are verbs, nouns, adjectives and adverbs. The meaningful phrases are constructed by considering neighbouring words with the rules:

- `^(DET)?(-?ADJ)*-?NOUN(-ADJ)?$`

- `^VERB-NOUN$`
- `^ADV-ADJ$`
- `^VERB-ADV$`
- `^ADJ$`

Word2vec [20] is known as a computationally efficient predictive vector space model (VSM) for learning word embeddings from raw text. The FastText implementation is considered to be a state-of-the-art for a couple of reasons. First, the models are trained using subword information, meaning that words are represented as a sequence of character n-grams. Second, the models for different languages can be aligned in a same vector space so the words from different languages with high semantic similarity are close to each other [31]. We opted for using FastText pre-trained aligned word vectors for English and French.

To represent a multi-word phrase, we calculate a centroid of word vectors. The centroid of a finite set of m ($m = 3$ in our case) word vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m \in \mathbb{R}^d$ (n is the vector dimension and in our case $d = 300$) is given as follows:

$$\mathbf{p_C} = \frac{\sum_{i=1..m} \mathbf{w}_i}{m} \tag{1}$$

Note that this is a very simple representation and there is a significant work done in document and sentence vector representations [34]. Our motivation to use centroid stems from the fact that the phrases are very short and the neighbouring words are likely to be semantically close. However, different phrase representations will be investigated in the follow-up work.

D. Spectral Clustering

The extracted phrases represented as the phrase vectors are clustered using spectral clustering algorithm. Given a set of n vectors (phrases) $\mathbf{P} = \{\mathbf{p}_{C1}, \mathbf{p}_{C2}, \ldots, \mathbf{p}_{Cn} \in \mathbb{R}^d\}$, the objective of spectral clustering is to divide these vectors into k clusters. The steps of the algorithm for spectral clustering can are:

- Construct an affinity matrix A, consisting of pairwise similarities a_{ij}. The similarity measure method used to calculate a_{ij} in this paper is Gaussian kernel function for constructing the similarity $a_{ij} = exp(-\gamma \mathbf{p}_{Ci} - \mathbf{p}_{Cj}{}^2)$, where $\gamma (= \sigma^2)$ is a specified scaling parameter used for determining the size of neighborhood.
- Compute normalized Laplacian matrix L based on affinity matrix A as $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where D is an $n \times n$ diagonal matrix with $d_i = \sum_{j=1}^{n} a_{ij}$ on the diagonal.
- Compute the k largest eigenvectors of the normalized Laplacian matrix L, and form the matrix $V = (v_{ij})_{n \times k}$ using these eigenvectors as its columns.
- Form the matrix $U = (u_{ij})_{n \times k}$ by normalizing the rows of V, such that $u_{ij} = v_{ij} / \sqrt{\sum_j v_{ij}^2}$.
- Each row of U represents a new vector for a phrase in $\mathbb{R}^k$ space. Then cluster the vectors using the k-means method.
- Assign each phrase $\mathbf{p}_{Ci}$ to a given cluster c if the corresponding row i in U is assigned to this cluster.

TABLE I
HYPERPARAMETERS FOR THE MODELS USED.

Method	# of topics	# of models	Hyperparameters
LDA	$k \in \{2, 4, .., 50\}$	20	$\alpha = 0.05, \beta = 0.01$
BTM	$k \in \{2, 4, .., 50\}$	20	$\alpha = 0.05, \beta = 0.01$
WNTM	$k \in \{2, 4, .., 50\}$	20	$\alpha = 0.05, \beta = 0.01$
GTC	$k \in \{2, 4, .., 50\}$	2	$\gamma = \{0.1, 1.0\}, \texttt{kernel} = rbf$

E. Hyperparameter settings

For the experiments on BTM, WNTM and LDA we used implementations described in the paper [35]. The reason for choosing hyperparameters values α, β and λ as shown in Table I is simply because they are recommended settings for short texts by the original authors [18], [19], [36] and in the paper [5].

F. Evaluation Method

The state-of-the-art evaluation methods for topic coherence are the intrinsic measure *UMass* [37] and the extrinsic measure *UCI* [38] which depends on external reference corpora.

$$\mathit{score}_{\mathit{UMass}}(w_i, w_j) = \log \frac{D(w_i, w_j) + \epsilon}{D(w_i)} \tag{2}$$

where $D(w_i)$ is the count of documents containing the word w_i, $D(w_i, w_j)$ the count of documents containing both words w_i and w_j, and D the total number or documents in the corpus. This score measures how much, within the words used to describe a topic, a common word is in average a good predictor for a less common word.

$$\mathit{score}_{\mathit{UCI}}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{3}$$

where $p(w_i, w_j) = D_{ref}(w_i, w_j)/D_{ref}$ and $p(w_i) = D_{ref}(w_i)/D_{ref}$, D_{ref} is the total number of documents in the external reference corpus, $D_{ref}(w_i)$ is the count of documents of reference corpus containing the word and $D_{ref}(w_i, w_j)$ the count of documents containing both words.

IV. RESULTS

A. Quantitative evaluation

Fig. 3 gives an overview of the coherence scores (UMass top and UCI bottom row) produced for the different methods. The topic is considered more coherent if the score is higher. UMass coherence score, as mentioned earlier, is calculated on the corpus itself. It indicates that the coherence slowly decreases with the number of topics. It also shows significantly lower value for GTC approach. The reason for this is that the responses (documents) are very short and the number of topically related terms within a response is low (1-3 related terms). Hence, the point-wise mutual information statistic is unable to pick up semantically related terms from different documents because they rarely occur in the same context. The coherence measure performed on the reference external corpus (Wikipedia with longer documents and more samples) demonstrates almost opposite results. GTC shows better coherence scores for French and English ($k > 20$). Entire French

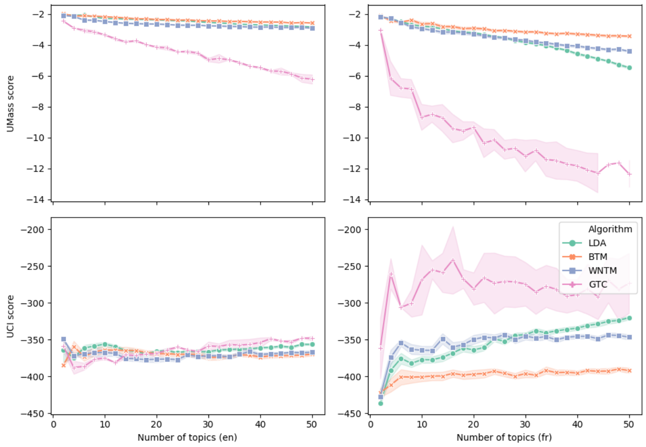

Fig. 3. *UMass* (top row) and *UCI* (bottom row) coherence measures calculated and averaged over different models for top 10 representative terms for English (left column) and French (right column) subsets.

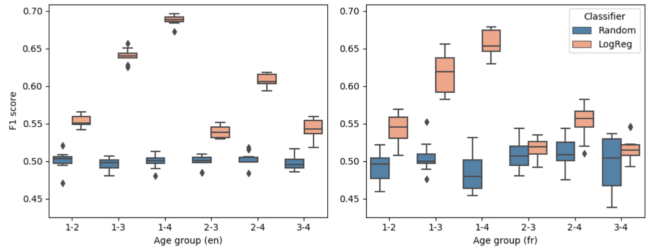

Fig. 4. Pairwise classification with 10-fold validation between age groups for English (left) and French (right) subsets.

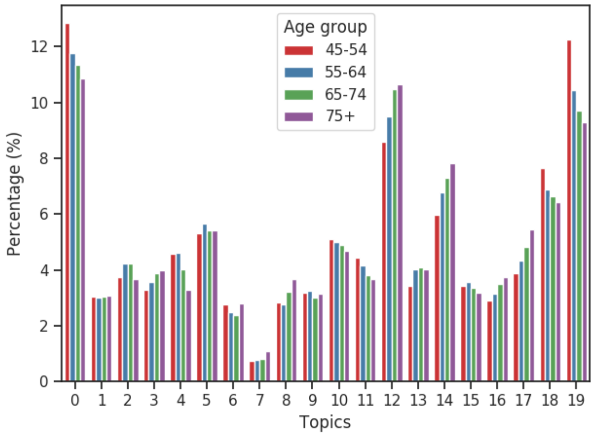

Fig. 5. Difference in topics among age groups.

Wikipedia (around 2.2 million documents) and entire simple English Wikipedia (around 200 thousand documents) were used as reference corpora. Due to the volume of standard edition of English Wikipedia (5.9 million articles at the time of writing) we were unable to use it as a reference, which may have reflected in the results.

B. Qualitative Evaluation

We explore the quality of the topics based on the opinions of three domain experts. We used topics generated for the setting where $k = 16$. The reason for choosing this number is based on empirical assumption about the number of topics derived from prior analysis of the graph representation. The coherence scores did not provide a definitive choice in terms of the number of topics.

1) Age Groups: To examine the differences between age groups in the dataset the experiment is set up as a set of binary classification problems. The classes are: 1 (45-54 age range), 2 (55-64 age range), 3 (65-74 age range) and 4 (75+ age). The classification is applied pairwise with all possible age group combinations. Fig. 4 shows logistic regression results with 10-fold validation on each pair. Interesting observation is that with the bigger age gap the classification accuracy tends to increase and the trends are similar in both languages. Please note that the features for the classification consist of filtered lemmas which are nouns, adjective, adverbs and verbs. The classification results and differences would be likely higher if we included the filtered words which is out of the scope of this paper.

Most notable trend on the Fig. 5 is that most of the topics show ordered gradual increase/decrease in a topic involvement per group. Most notable difference is for the first age group which use phrases from topic 0 cluster and topic 19 cluster more than other groups. Topic 0 cluster contains words about exercise and topic 19 is about healthy eating and diet.

2) Gender: To examine the differences between genders in the dataset the experiment is set up as a binary classification problem. The classes are: F (women) and M (men). Using logistic regression and the same set of feature as for the age groups we show that there is a difference between men's and women's responses. Fig. 6 illustrates the results on 10-fold cross validation.

Fig. 7 shows the differences between genders. Most notable difference is topics 0, 2, 3, 4 and 5. Topic 0 cluster has terms mostly about exercise. Male participants use words from this cluster more than female. Clusters 2, 3 and 4 contain word related to family, children and relationships. Female participants tend to use talk about these topic slightly more than male. The other clusters seem more or less balanced.

3) Pre-existing Conditions: In this section we examine the topical differences in participants that reported health conditions. On the conceptual graph 1 the filter is referred as ”CCC“. Similar classification experiments were conducted on subsets of participants who reported anxiety versus who did not, cancer versus who did not and Alzheimer's disease versus participants who did not. However, there was no significant difference between the groups and logistic regression classifier did not perform better than random. Although the difference was not detected in the classification experiments, the topic modeling methodology can help in discovering the differences on a semantic level. Fig. 8 and Fig. 9 show the topical distribution for three setups: anxiety-no anxiety, cancer-no cancer and Alzheimer's-no Alzheimer's.

TABLE II
Top 10 terms and coherence scores for two example topics per method for English subset, where $k = 16$. (a) best topic according to UMass score, (b) best topic according to UCI score.

Method	UMass	UCI	Terms
LDA (a)	-1.8966	-326.8925	exercise good social family diet friend healthy relationship life activity
LDA (b)	-2.5279	-171.6791	eat exercise properly right healthy active n't eating food drink
BTM (a)	-1.9055	-276.5447	exercise activity social active diet physical mental mind healthy good
BTM (b)	-2.5599	-198.8898	exercise eat food good diet vegetable healthy not n't fruit
WNTM (a)	-2.2245	-182.6595	positive attitude life outlook mental good n't people happy not
WNTM (b)	-2.5611	-234.1146	good healthy active prop regular positive social balanced attitude activity
GTC (a)	-2.8286	-357.9075	active activity important interest physical physically mind mentally interested mental
GTC (b)	-3.9109	-254.3686	positive attitude moderation outlook good humour fun laugh humor mental

TABLE III
Top 10 terms and coherence scores for two example topics per method for French subset, where $k = 16$. (a) best topic according to UMass score, (b) best topic according to UCI score.

Method	UMass	UCI	Terms
LDA (a)	-2.0289	-375.1233	physique vie social activite alimentation exercice bien bon mental travail
LDA (b)	-2.5971	-93.5049	physique activite alimentation bon exercice nutrition mental stress sain activites
BTM (a)	-2.0253	-280.7812	actif pas physiquement bien bon alimentation exercice vie sante stress
BTM (b)	-3.2007	-182.1551	pas problemes regulier vis mental trop difference physique alimentation vie
WNTM (a)	-2.1108	-320.6495	plus possible vie bon alimentation pas medecin moins stress exercice
WNTM (b)	-2.9369	-104.4695	soin sante gens plus bien mental pas exercice personne c'est
GTC (a)	-2.7957	-378.3365	alimentation bon nourriture nutrition sain physique exercice gestion genetique activite
GTC (b)	-8.1423	47.5306	activite physique actif genetique activites activities gene excess genes hygiene

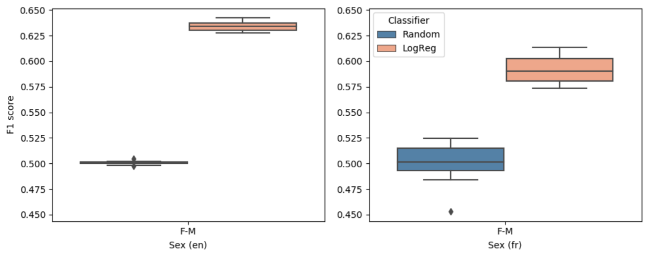

Fig. 6. Classification with 10-fold validation between genders for English (left) and French (right) subsets.

V. Conclusion

In summary, the current work has demonstrated an alternative method for topic extraction from OE responses. We compared the method with probabilistic state-of-the-art approaches for short texts: BTM and WNTM, and LDA as a baseline. The results are compared based on *Umass* and *UCI* coherence measures which are two common unsupervised evaluation approaches. The observation is that these two measures, although based on the same idea (point-wise mutual information) show different results on the dataset. The main difference is that the former is intrinsic (based on the statistics of the dataset) and the latter is extrinsic (based on the statistics of the larger external corpus). We show and discuss why, in this case study, the extrinsic measure is more suitable to measure topic coherence. Additionally, we explore topical distributions with different grouping setups and discover some interesting insights about the data.

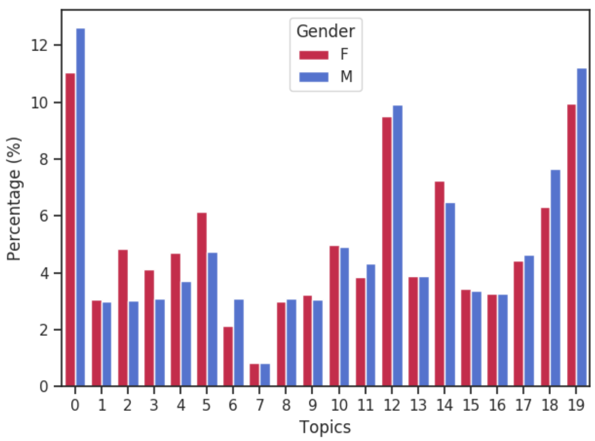

Fig. 7. Difference in topics between genders.

Nevertheless, there are a couple of drawbacks of this approach that are important to mention. First, it is not suitable for online topic modeling as it depends on clustering which is too slow for real-time settings. However, the surveys are closed sets that are primarily focused on exploratory analyses and the prompt performance time is not a requirement. Second, the quality of the results largely depends on the quality of pre-trained word vectors. To put it into perspective, for domain-specific datasets this can pose a challenge in a sense that the word vectors may not have a good coverage for domain-specific terms.

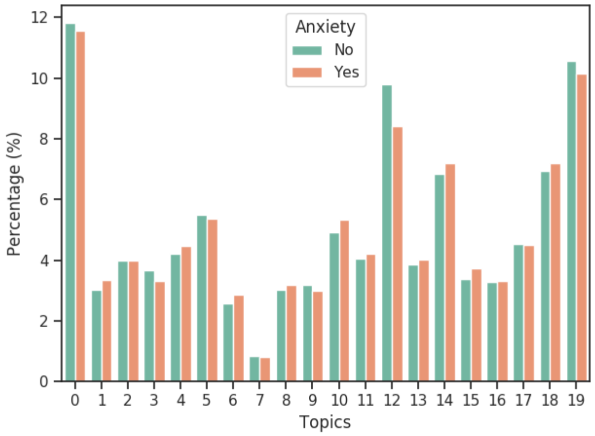

Fig. 8. Difference in topics in setup anxiety-no anxiety.

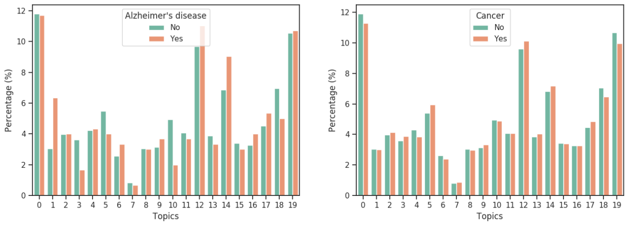

Fig. 9. Difference in topics in setups: Alzheimer's-no Alzheimer's (left), cancer-no cancer (right).

References

[1] G. Brokos, P. Malakasiotis, and I. Androutsopoulos, "Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering," in *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 114–118. [Online]. Available: https://www.aclweb.org/anthology/W16-2915

[2] S. Spinelli, C. Dinnella, C. Masi, G. P. Zoboli, J. Prescott, and E. Monteleone, "Investigating preferred coffee consumption contexts using open-ended questions," *Food Quality and Preference*, vol. 61, pp. 63–73, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950329317301039

[3] F. ten Kleij and P. A. Musters, "Text analysis of open-ended survey responses: A complementary method to preference mapping," *Food Quality and Preference*, vol. 14, no. 1, pp. 43–52, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950329302000113

[4] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand, "Structural topic models for open-ended survey responses," *American Journal of Political Science*, vol. 58, no. 4, pp. 1064–1082, 2014.

[5] A.-S. Pietsch and S. Lessmann, "Topic modeling for analyzing open-ended survey responses," *Journal of Business Analytics*, vol. 1, no. 2, pp. 93–116, 2018. [Online]. Available: https://doi.org/10.1080/2573234X.2019.1590131

[6] H. E. Tinsley and D. J. Weiss, "Interrater reliability and agreement of subjective judgments," *Journal of Counseling Psychology*, vol. 22, no. 4, p. 358, 1975.

[7] J. W. Fleenor, J. B. Fleenor, and W. F. Grossnickle, "Interrater reliability and agreement of performance ratings: A methodological comparison," *Journal of Business and Psychology*, vol. 10, no. 3, pp. 367–380, Mar 1996. [Online]. Available: https://doi.org/10.1007/BF02249609

[8] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA '10. New York, NY, USA: ACM, 2010, pp. 80–88. [Online]. Available: http://doi.acm.org/10.1145/1964858.1964870

[9] N. Bohnert, J. Chagnon, and P. Dion, "Population projections for Canada (2013 to 2063), Provinces and Territories (2013 to 2038)," Statistics Canada, Tech. Rep., 2015. [Online]. Available: https://www150.statcan.gc.ca/n1/en/pub/91-520-x/91-520-x2014001-eng.pdf

[10] D. J. Sheets and E. M. Gallagher, "Aging in Canada: State of the Art and Science," *The Gerontologist*, vol. 53, no. 1, pp. 1–8, 11 2012. [Online]. Available: https://doi.org/10.1093/geront/gns150

[11] "CIHR Institute of Aging. IA strategic plan 2013-2018: Living longer, living better," http://www.cihr-irsc.gc.ca/e/47179.html, 2013, accessed: 2019-05-30.

[12] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 289–296. [Online]. Available: http://dl.acm.org/citation.cfm?id=2073796.2073829

[13] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012. [Online]. Available: http://doi.acm.org/10.1145/2133806.2133826

[14] K. W. Lim, C. Chen, and W. L. Buntine, "Twitter-network topic model: A full bayesian treatment for social network and text modeling," *CoRR*, vol. abs/1609.06791, 2016. [Online]. Available: http://arxiv.org/abs/1609.06791

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[16] P. Pinoli, D. Chicco, and M. Masseroli, "Latent Dirichlet allocation based on Gibbs sampling for gene function prediction," in *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, May 2014, pp. 1–8.

[17] E. Hörster, R. Lienhart, and M. Slaney, "Image retrieval on large-scale image databases," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, ser. CIVR '07. New York, NY, USA: ACM, 2007, pp. 17–24. [Online]. Available: http://doi.acm.org/10.1145/1282280.1282283

[18] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13. New York, NY, USA: ACM, 2013, pp. 1445–1456. [Online]. Available: http://doi.acm.org/10.1145/2488388.2488514

[19] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: A simple but general solution for short and imbalanced texts," *Knowledge and Information Systems*, vol. 48, no. 2, pp. 379–398, Aug. 2016. [Online]. Available: http://dx.doi.org/10.1007/s10115-015-0882-z

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[21] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[22] M. Hajjem and C. Latiri, "Combining ir and lda topic modeling for filtering microblogs," *Procedia Computer Science*, vol. 112, pp. 761–770, 2017, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917315235

[23] P. Bicalho, M. Pita, G. Pedrosa, A. Lacerda, and G. L. Pappa, "A general framework to expand short text for topic modeling," *Information Sciences*, vol. 393, no. C, pp. 66–81, Jul. 2017. [Online]. Available: https://doi.org/10.1016/j.ins.2017.02.007

[24] I. Vulić, W. D. Smet, J. Tang, and M.-F. Moens, "Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications," *Information Processing & Management*, vol. 51, no. 1, pp. 111 – 147, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457314000739

[25] D. Zhang, Q. Mei, and C. Zhai, "Cross-lingual latent topic extraction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1128–1137. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858681.1858796

[26] I. Vulić and M.-F. Moens, "Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '15. New York, NY, USA: ACM, 2015, pp. 363–372. [Online]. Available: http://doi.acm.org/10.1145/2766462.2767752

[27] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Natural Language Engineering*, vol. 16, no. 1, pp. 100–103, 2010.

[28] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, pp. 252–259. [Online]. Available: https://www.aclweb.org/anthology/N03-1033

[29] B. Sagot, "The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French," in *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, May 2010. [Online]. Available: https://hal.inria.fr/inria-00521242

[30] A. Islam and D. Inkpen, "Real-word spelling correction using Google Web 1T 3-grams," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, ser. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 1241–1249. [Online]. Available: http://dl.acm.org/citation.cfm?id=1699648.1699670

[31] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2979–2984. [Online]. Available: https://www.aclweb.org/anthology/D18-1330

[32] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions," in *Proceedings of the 23rd International Conference on Computational Linguistics*, ser. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 340–348. [Online]. Available: http://dl.acm.org/citation.cfm?id=1873781.1873820

[33] "Neo4j graph platform," https://neo4j.com/, 2019, accessed: 2019-05-30.

[34] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *ICLR*, 2017.

[35] J. Qiang, Y. Li, Y. Yuan, W. Liu, and X. Wu, "STTM: A tool for short text topic modeling," *arXiv preprint arXiv:1808.02215*, 2018.

[36] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 299–313, 2015. [Online]. Available: https://www.aclweb.org/anthology/Q15-1022

[37] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 262–272. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145462

[38] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 100–108. [Online]. Available: http://dl.acm.org/citation.cfm?id=1857999.1858011