Anh Dang[1], Abidalrahman Moh'd[1], Aminul Islam[2], Rosane Minghim[3], Mike Smit[1], and Evangelos Milios[1]

[1]Dalhousie University, Halifax, Nova Scotia, Canada
[2]*University of Louisiana at Lafayette, LA, USA 70503*
[3]*University of Sa˜o Paulo-USP, ICMC, Sao Carlos, Brazil*

5. Evaluation

5.1 Baselines

- **Random**: Each sentence pair is assigned a random real semantic similarity score between [0, 1].
- **Weighted Matrix Factorization (WTMF):** This baseline uses the state-of-the-art unsupervised method [2].
- **Random 5-gram**: we select random 5-grams from the Reddit temporal n-gram corpus with the same size of the 5-grams that contain the topic.
- **Google Tri-gram Method** (GTM): Google Tri-gram Method [3] assigns a semantic similarity score between two sentences using the unigrams and trigrams of the Google Web 1T corpus.

5.2 SEMEVAL 2015 Methods

- **Columbia**: This method used Orthogonal Matrix Factorization to compute a representation vector for each sentence..
- **Yamraj**: This method learned sentence vectors from Google News dataset (about 100 billion words) and Wikipedia articles
- **MathLingBp:** This method exploit the use of the align-and-penalize architecture.

5.3 Topic-based LSA Parameter Setting

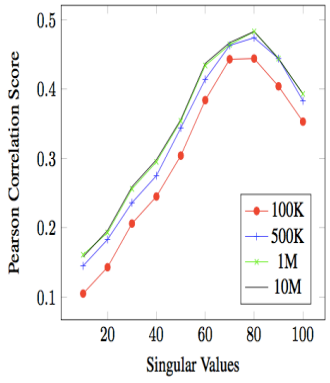

Figure 4: For SS task, Pearson correlation score with an increasing singular values and 5-gram dimensionality. TLSA achieves the best Pearson correlation score for $k = 80$ and the dimensionality of 5-grams = 1M.

5.4 Topic-based LSA versus Baselines and other Methods

Table 5: TLSA results with other baselines and compared methods. Combining TLSA with sentiment analysis achieves the best result for both PI and SS tasks.

Methods / *Baselines*	Paraphrase Identification			Semantic Similarity			
	F1	Precision	Recall	Pearson	maxF1	maxPrec	maxRecall
Human Upperbound	0.823	0.752	0.0908	0.735	–	–	–
TLSA & Sentiment	**0.591**	0.764	0.480	**0.483**	0.582	0.761	0.472
COLUMBIA	0.588	0.593	0.583	0.425	0.599	0.623	0.577
TLSA	**0.585**	0.761	0.474	**0.483**	0.585	0.761	0.474
YAMRAJ	0.496	0.725	0.377	0.360	0.542	0.502	0.589
WTMF	0.536	0.450	0.663	0.350	0.587	0.570	0.606
Random 5-gram	0.504	0.716	0.389	0.466	0.564	0.824	0.429
GTM	0.495	0.391	0.674	0.371	0.582	0.761	0.472
Random	0.266	0.192	0.434	0.017	0.350	0.215	0.949

6. Conclusions and Future Work

- We introduced Reddit temporal n-gram corpus, which is designed specifically for social media text.
- This large-scale terabyte corpus includes all the word unigram to 5-gram, and their frequency per month from October, 2007 to August, 2016.
- We propose a novel Topic-based Latent Semantic Analysis approach which exploits the 5-grams of the corpus.
- The proposed TLSA outperforms all the state-of-the-art unsupervised and semi-supervised methods in SEMEVAL 2015 Task 1.
- For future work, we aim to use this corpus to study the linguistic patterns of social media text, for example, finding the meaning of new words in social media. In addition,

References

[1] Xu, W., Callison-Burch, C., & Dolan, W. B. (2015). SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). *Proceedings of SemEval*.

[2] Guo, W., & Diab, M. (2012, July). Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 864-872). Association for Computational Linguistics.

[3] Islam, A., Milios, E., & Kešelj, V. (2012, May). Text similarity using google tri-grams. In *Canadian Conference on Artificial Intelligence* (pp. 312-317). Springer Berlin Heidelberg.