Comparison of Methods for Topical Clustering of Online Multi-speaker Discourses

Vid Stropnik
University of Ljubljana,
Faculty of Computer and
Information Science,
Velenje, Slovenia
vs6309@student.uni-lj.si

Zoran Bosnić
University of Ljubljana,
Faculty of Computer and
Information Science,
Ljubljana, Slovenia
zoran.bosnic@fri.uni-lj.si

Evgeny Osipov
Luleå University of Technology,
Department of Computer Science,
Electrical and Space Engineering,
Luleå, Sweden
evgeny.osipov@ltu.se

ABSTRACT

Discussions held on online forums differ from traditional text documents in several ways. In addition to individual text-bodies (submission comments, forum posts etc.) being very short, they also have multiple messengers, each of whom may exhibit unique patterns of speech. Consequently, state of the art methods for text summarization are often rendered inapplicable for these sorts of corpora. This paper evaluates the topic-clustering algorithm used in the state-of-the-art online comment clustering techniques, as parts of commonly used summarizer models. It proposes two alternative, vector-based approaches and presents results of a comparative external analysis, concluding in the three methods being comparable.

KEYWORDS

latent Dirichlet allocation, word embeddings, GloVe, hyperdimensional computing, self-organized maps, topical clustering, clustering evaluation, discussion summarization

1 Introduction

User generated comments carry a great amount of useful information. Big data researchers have successfully used them to predict stock market volatility [1] and predict the characteristics of such comments that perform the best on a given online platform [2]. User comments can also offer vast amounts of complementary information, as well as being forms of information surveillance, entertainment or social utility [3]. Existing mechanisms for displaying comments on websites do not scale well and often lead to *cyberpolarization* [4]. Furthermore, they are platform-specific and often fail to offer an overall image of the topics discussed in a given comments section.

A comprehensive, easily understandable automatic summary of the online discourse at hand can be instinctively understood as a solution to this problem. This, however, is no easy task, seeing as these corpora are often very short and come from multiple speakers. Consequently, traditional summarization methods do not translate well to these sorts of text bodies.

In Section 2 of this paper, the related work establishes the general framework that other authors generally use for the task at hand. It establishes the Latent Dirichlet Allocation (LDA) topic modeling algorithm as the current leading method for topical grouping of individual comments. These topical groups play a pivotal role in later summarization steps, also presented in Section 2.

In this paper, we externally evaluate and compare LDA versus two frameworks, using word representations in semantic vector space. We describe the analyzed methods in Sections 3 and 4. In Section 5, we describe the comparative evaluation methodology used to determine the applicability of each modeling technique and present our results. We follow it up by discussing further work in the conclusion of this paper.

2 Related work

Online discussion summarization is a field that has not been addressed directly by many authors. One group of works [5-7] have roughly described a three-step process, commonly presented as the state of the art. The approaches includes a topical clustering of all the observed comments, establishing a ranking method for determining the most salient ones in each cluster, and later summarizing this selection. Between them, the authors confidently establish Latent Dirichlet Allocation (LDA) topic modeling as the most human like grouping algorithm. Further work also proposes a novel graph-based linear regression model based on the Markov Cluster Algorithm (MCL), [8] which outperforms LDA, but uses the knowledge of multidomain knowledge bases for implementation. While we argue that extractive summarization is not an ideal method for the analysis of multi-speaker corpora, the first step of identifying and topically clustering individual comments in each comment section is assumed as a required step towards successful summarization of the topics discussed therein.

To the best of our knowledge, popular NLP word embedding algorithms (i.e. *word2vec, GloVe*) have not been used directly for comment summarization applications up until now. Similarly. neither have hyperdimensional representations, another topic of interest.

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

3 NLP Methods

In this work, we examine three distinct topical clustering models, the output of which is always a set of comment clusters, given a multi-comment input.

The first is an LDA model, using a Term frequency – inverse document frequency (TFIDF) word representation as an input. In this representation, the comments were hard-clustered into the groups, determined by the degree of membership of which had been the highest in a soft-clustering approach, provided by the LDA model.

The second examined model uses *GloVe* word embeddings clustered with the k-means clustering algorithm, thus portraying words in semantic vector space using information of contexts in which words often appear.

The third model creates Hyperdimensional representations of words, mapped them into a two-dimensional topology using the self-organized maps algorithm and then clustered it like the preceding model. This approach is the least explored for this use-case and is inspired by the observed differences between the functionality of the human brain and the traditional von Neumann architecture for modern computing.

We performed the comparative evaluation of the models on the *Reddit Corpus (by subreddit)* dataset, provided by the Cornell Conversational Analysis Toolkit (Convokit)[1]. Five *Conversations,* corresponding to as many treads on the website Reddit were extracted from the corpus. We selected threads, discussing topics from different subject domains, where each contained at least 50 non-removed comment text bodies. Two human annotators were then asked to manually identify topical clusters in the selected *Conversations*. The comment texts were provided to them in the form of a set of numbered text files, containing only the text data in chronological order of submission. Reddit post titles or other metadata were not available to the annotators and no guidance was given as to the number of topics required. The clusterings were examined as-is, with no singleton removal performed.

We describe the NLP techniques used to create the three clustering models in the following subsections, with external evaluation results being presented in Section 5.

3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a topic modeling technique initially proposed in the context of population genetics, but later applied in machine learning in the early 21st century. It assumes a generative process of documents as random mixtures over a collection of latent topics. Each of these topics, in turn, is characterized by a certain distribution over words. A topic model can be created by estimating the per document distribution of topics θ and the per topic distribution over words φ. [9] Many methods, such as variational inference, Bayesian parameter estimation [9] and Collapsed Gibbs sampling [10], have been used to approximate these values. In the end, they all boil down to maximizing the model's probability of creating the exact documents, provided to it in the input, assuming the knowledge of the number of topic distributions.

3.2 Word Embeddings

Word Embeddings is a collective name for a set of language modeling and feature learning techniques, yielding word representations using vectors, the relative similarities of which correlate with the semantic similarity of the represented words. These meanings are extracted from the contexts – fixed-size windows of preceding and succeeding words, in which individual words appear in in the training corpus. The generation of these vectors is achieved by Context counting [11] or Context prediction [12]. While there have been several claims of one of the methods for synthetizing word embeddings being superior over another, recent work implies the correspondence between these model-types [13]. Whichever way these word-vectors are created, they represent semantic meaning in vector space. Using algebraic similarity measures (in our case, cosine distance) on comment-word averages, the relative likeness of the examined comments' meanings is calculated. Comment clusters can then be created by clustering the semantic-space points into groups with high intra-cluster and low inter-cluster similarity. These groups represent topical clusters, used in our examination.

3.3 Hyperdimensional Computing

Hyperdimensional computing is a family of biologically inspired methods for representing and manipulating concepts and their meanings in high-dimensional space. Random Bipolar vectors of high, but fixed dimensionality (≥ 1000) are initialized as individual word representations and are then transformed in ways that represent semantically similar comments closer in the high-dimensional vector space, while the similarity of dissimilar comments is likely close to zero due to their inherent orthogonality. The methods used to transform these vectors are binding, bundling and permuting [14]. By using these methods, individual hyperdimensional vectors are created for each comment, encoding the used words and their position in the comment in the vector.

Similar to the clustering of word embeddings, semantically similar comment groups can be found by clustering, thus determining the outputs of the third model. However, the performance of this method did not yield comparative results at first. We hypothesised that this might be due to the high component count of the used vectors (more than double the dimensions of the Word Embedding approach), so a method of dimensionality reduction was examined, aiming to improve its results. It is described in the next sub-section.

3.4 Self-Organized Maps

Self-organizing maps (SOM), also known as Kohonen networks are computational methods for the visualization and analysis of high-dimensional data. The output of the algorithm is a set of nodes, arranged in a certain topology that represents the nodes' mutual relation, with each node being represented with a weight vector of *t* dimensional components, with *t* corresponding to the uniform dimensionality of data being reduced [15]. As data representations in high-dimensional vector spaces are inherently vulnerable to sparseness, clustering outputs can differ in cases where the clustered data is first dimensionally reduced. Thus, we used the SOM algorithm to examine if the results (of the examination in Section 5) of any of the proposed frameworks can

[1] https://convokit.cornell.edu/documentation/index.html/

be improved by dimensionally reducing the vector representations prior to clustering.

SOM proved to drastically improve the performance of the Hyperdimensional computing model, while making the Word Embeddings-based model perform worse. Consequently, we only use SOM prior to clustering the HD-based approach in the evaluation, presented in Section 5.

4 Implementation

All implementational work was done with the *Python* programming language. All text corpora were pre-processed using the WordNetLemmatizer and PorterStemmer from NLTK.[2] Stop word removal was done in the pre-processing step using the topic modeling package Gensim[3], which also provided the submodules for TFIDF and LdaModel, used for the implementation of Latent Dirichlet Allocation. GloVe word embeddings were provided as part of the NLP open-source library SpaCy[4] as part of the "*en_core_web_md"* pretrained statistical model for the English Language. The SOM algorithm was implemented using the SimpSOM package[5], with k-means clustering being provided by Scikit-Learn.[6]

5 Evaluation

To analyze the applicability of LDA, Word Embeddings and dimensionally reduced Hyperdimensional computing for the discussed use-case, topical clustering outputs were created for 5 Reddit *Conversations*. Two human annotators also manually created topical groups for these conversations. The goal of our evaluation was to see which model created the most *human-like* clusters; consequently having the highest average agreement measure with the clustering samples, provided by the two annotators.

Topical clusters, created by the three models, were externally evaluated using four symmetric agreement measures: The V-Measure [16], The Fowlkes-Mallows Index [17], the Rand Index [18] and the Mutual information score [19]. The latter two were also adjusted for variance. For each examined model, the best performing number of topic clusters was selected. The agreement of the clustering output of each model was measured against both of the manual clusterings, with the *per annotator* average of each metric being the final output.

Figure 1 shows the result scores of all four metrics for each analyzed method. In the top row, the average agreement between the two annotators is also shown. This is, expectedly, higher than the average agreement between any examined model and the human outputs. A few takeaways can be addressed, examining the figure. Firstly, the different methods were successful to a varying degree, depending on the used metric, with each performing the best according to at least one. Secondly, when comparing their average relative success in relation to the agreement scores between Annotator A and Annotator B, we can see that their performances are very similar. This can be seen even clearer in Figure 2, which shows each model's performance with respect to the agreement score between the two human annotators. The percentage is calculated as an averaged sum of all four metric scores, weighted by the sum of these scores, achieved by the human versus human evaluation. In the figure, Word Embeddings can be seen as the best-performing approach, reaching 54.18 % of the Human agreement. The performance of LDA presented in Figure 2 is also comparable to that found in [5].

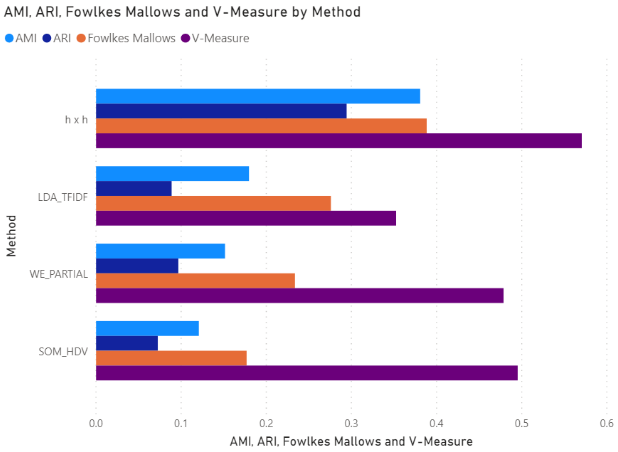

Figure 1: Visualization of agreement metric results between the human annotators (top) and the average annotator vs. model agreement (bottom three)

However, the difference in results between the best and the worst performing models being less than 7% of the total human agreement score, this metric is not enough to establish Word Embeddings as superior to LDA or indeed, dimensionally reduced High-dimensional computing. We can conclude that both Hyperdimensional computing and Word Embeddings can produce topical clusters, comparable to the current state of the art LDA method.

Semantic document representations performing as well as the state-of-the-art topic modeling framework using LDA opens up plentiful possibilities in the field of multi-speaker conversation analysis. Whereas topic modeling's more direct approach of inferring latent conversation topics might be useful in their discovery, the possibility of applying algebraic functions to individual comment vectors might enable further topic mining and experimentation. While the k-means clustering algorithm requires a desired number of clusters at input, similar to LDA, its job is not to encode semantics in the Word Embedding or SOM-HDC framework. This means that an alternative clustering algorithm – one without the need for an input number of medoids - could be used for the task of grouping comments. This, in turn, would result in a truly unsupervised topical clustering framework. A comparative evaluation of these approaches is a field of interest in the future, as our non-conclusive experiments have

[2] https://www.nltk.org/
[3] https://radimrehurek.com/gensim/
[4] https://spacy.io/
[5] https://github.com/fcomitani/SimpSOM/
[6] https://scikit-learn.org/

already shown a vast variance in results when using different clustering approaches.

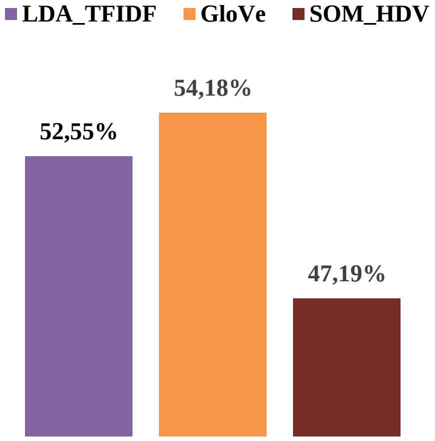

Figure 2: Percentage of the Human versus human agreement score achieved by each model (averaged between 4 agreement metrics)

6 Conclusion

In this article, we work from our hypothesis that popular semantics-laden vector representations of text data can be applicable in the established framework for extractive online discussion summarization. We present two models using different vector-based representation techniques and conclude that they are both comparable to the Latent Dirichlet Allocation topic modelling technique, used in most literature, with the Word Embeddings-based framework outperforming it in our external evaluations.

As mentioned in Section 2, the authors of this article argue that extractive summarizations are intrinsically less suitable when working with multi-speaker corpora. Our future work in this field includes the modeling of an abstractive summarizer framework, using the findings presented in this paper. Our intent is to use them in conjunction with graph-based approaches that take advantage of multidomain knowledge bases like DBPedia for both clustering and topic-labelling [8, 20].

Whether used in extractive or abstractive applications, we presume that the field will greatly benefit from our findings, seeing that the two vector-based representation frameworks open a plethora of new possibilities for other researchers. These include the detailed data manipulation using algebraic operations on individual comment vectors, as well as said vectors being suitable inputs for deep learning models using neural networks.

References

[1] W. Antweiler and M. Z. Frank, ‘Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards’, *J. Finance*, vol. 59, no. 3, pp. 1259–1294, Jun. 2004, doi: 10.1111/j.1540-6261.2004.00662.x.

[2] T. Weninger, ‘An exploration of submissions and discussions in social news: mining collective intelligence of Reddit’, *Soc. Netw. Anal. Min.*, vol. 4, no. 1, p. 173, Dec. 2014, doi: 10.1007/s13278-014-0173-9.

[3] E. Go, K. H. You, E. Jung, and H. Shim, ‘Why do we use different types of websites and assign them different levels of credibility? Structural relations among users’ motives, types of websites, information credibility, and trust in the press’, *Comput. Hum. Behav.*, vol. 54, pp. 231–239, Jan. 2016, doi: 10.1016/j.chb.2015.07.046.

[4] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg, ‘Opinion space: a scalable tool for browsing online comments’, in *Proceedings of the 28th international conference on Human factors in computing systems - CHI ’10*, Atlanta, Georgia, USA, 2010, p. 1175, doi: 10.1145/1753326.1753502.

[5] C. Llewellyn, C. Grover, and J. Oberlander, ‘Summarizing Newspaper Comments’, *Proc. Eighth Int. AAAI Conf. Weblogs Soc. Media*, pp. 599–602, Jun. 2014.

[6] Z. Ma, A. Sun, Q. Yuan, and G. Cong, ‘Topic-driven reader comments summarization’, in *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM ’12*, Maui, Hawaii, USA, 2012, p. 265, doi: 10.1145/2396761.2396798.

[7] E. Khabiri, J. Caverlee, and C.-F. Hsu, ‘Summarizing User-Contributed Comments’, presented at the International AAAI Conference on Weblogs and Social Media, pp. 534–537, Barcelona, Spain, Jul. 2011.

[8] A. Aker *et al.*, ‘A Graph-Based Approach to Topic Clustering for Online Comments to News’, in *Advances in Information Retrieval*, vol. 9626, N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello, Eds. Cham: Springer International Publishing, 2016, pp. 15–29.

[9] D. Blei, A. Y. Ng, and M. I. Jordan, ‘Latent Dirichlet Allocation’, *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2000, doi: 10.1162/jmlr.2003.3.4-5.993.

[10] W. M. Darling, ‘A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling’, *Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.*, pp. 642–647, Dec. 2011.

[11] 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, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ‘Efficient Estimation of Word Representations in Vector Space’, *ArXiv13013781 Cs*, Sep. 2013, Accessed: Aug. 19, 2020. [Online]. Available: http://arxiv.org/abs/1301.3781.

[13] A. Österlund, D. Ödling, and M. Sahlgren, ‘Factorization of Latent Variables in Distributional Semantic Models’, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 227–231, doi: 10.18653/v1/D15-1024.

[14] D. Kleyko, E. Osipov, D. De Silva, and U. Wiklund, ‘Distributed Representation of n-gram Statistics for Boosting Self-organizing Maps with Hyperdimensional Computing’, in *Perspectives of System Informatics, 12th International Andrei P. Ershov Informatics Conference, Revised Selected Papers*, pp. 64-79, Novosibirsk, Russia, 2019.

[15] T. Kohonen, T. S. Huang, and M. R. Schroeder, *Self-Organizing Maps*. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2012.

[16] A. Rosenberg and J. Hirschberg, ‘V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure’, in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, Jun. 2007, pp. 410–420, Accessed: Aug. 20, 2020. [Online]. Available: https://www.aclweb.org/anthology/D07-1043.

[17] E. B. Fowlkes and C. L. Mallows, ‘A Method for Comparing Two Hierarchical Clusterings’, *J. Am. Stat. Assoc.*, vol. 78, no. 383, pp. 553–569, Sep. 1983, doi: 10.1080/01621459.1983.10478008.

[18] W. M. Rand, ‘Objective Criteria for the Evaluation of Clustering Methods’, *J. Am. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, Dec. 1971, doi: 10.1080/01621459.1971.10482356.

[19] N. X. Vinh, J. Epps, and J. Bailey, ‘Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance’, *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Oct. 2010.

[20] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, ‘Unsupervised graph-based topic labelling using dbpedia’, in *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM ’13*, Rome, Italy, 2013, p. 465, doi: 10.1145/2433396.2433454.