A DistilBERTopic Model for Short Text Documents

Junaid Rashid[1], Jungeun Kim[2], Usman Naseem[3], Amir Hussain[4]
[1]Department of Computer Science and Engineering, Kongju National University, South Korea
[2]Department of Software, Kongju National University, South Korea
[3]School of Computer Science, University of Sydney, Australia
[4]School of Computing, Edinburgh Napier University, UK
[1]junaidrashid062@gmail.com,[2]jekim@kongju.ac.kr
[3]usman.naseem@sydney.edu.au,[4]a.hussain@napier.ac.uk

Abstract

The analysis of short text documents has become a vital and challenging task. Topic models are utilized to extract topics from a large amount of text data. However, these topic models typically suffer from data sparsity problems when applied to short texts because of relatively lower word co-occurrence patterns. As a result, they tend to provide repetitive or trivial topics of poor quality. Therefore, we presented a DistilBERTopic model to remove the sparsity problem and discover quality topics more accurately from short texts. DistilBERTopic model utilized the pre-trained transformer-based language models, reduced the dimensionality effect on embedding, clustered these embeddings, and discovered the topics from short text documents. Experimental results demonstrate that the DistilBERTopic model achieves better classification and topic coherence than other state-of-the-art topic models for real-world datasets.

1 Introduction

Numerous Web applications, including online social networks, recommendation systems, and question and answer systems, have recently grown in popularity. User-generated content has proliferated, particularly the massive increase in short text in various contexts like blogging, text messages, or customer reviews. It has become a crucial and difficult challenge in many applications to automatically discover latent semantic topics from huge amounts of short texts.

Considerable effort has been devoted to tackling the issue of data sparsity in topic modeling for short text documents. In prior work, for instance, a method is developed for aggregation of a few specific sentences that recreate a lengthier pseudo document by employing appropriate strategies like as combining all text messages originating from a single author (Hong and Davison, 2010a) or establishing relation information between hashtags (Wang, Liu, Qu, Huang, Chen, and Feng, 2014). In addition, some brief messages can contain contextual information such as URL, location, or timestamp. A large amount of the world's textual data comes from news sources and web portals, and all these sources often include various descriptions (Ramage, Hall, Nallapati, and Manning, 2009). However, these strategies may fail in the absence of contextual information (Naseem, Razzak, Khan, and Prasad, 2021). Conventional topic modeling techniques like Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, and Harshman, 1990), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), and Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2001) is extensively employed for discovering the topics from documents.

The formation of co-occurring word pairs across the same document is explicitly modeled in the biterm topic model (BTM) (Cheng, Yan, Lan, and Guo, 2014). The technique avoids the sparsity issue at the document level by aggregating the corpora biterms into a big pseudo document from which the topic distribution is inferred. However, the method does not consider word order. LF-DMM (Nguyen, Billingsley, Du, and Johnson, 2015) enriches Dirichlet Multinomial Mix with latent feature word representations by substituting the topics term with a combination of a Dirichlet multinomial and word embedding. In particular, (Weng, Lim, Jiang, and He, 2010) combines all of the shorter texts produced by that particular individual into a trained model before applying a standard LDA model. The two different aggregation strategies for short texts include the authors of the text and each word in the corpus(Hong and Davison, 2010b). The data preprocessing step for LDA (Mehrotra, Sanner, Buntine, and Xie, 2013) presents alternative tweet clustering strategies to build pseudo documents.

A word network topic model is presented that generates pseudo documents based on the network of words used together in the network (Zuo, Zhao, and Xu, 2016c). In (Zuo, Wu, Zhang, Lin, Wang, Xu, and Xiong, 2016a), inferring topic distribution from the high number of hidden pseudo documents with a drastically reduced amount of these documents. WNTM (Zuo, Zhao, and Xu, 2016d) constructs a network based on how often certain words appear, find word groups and distribute across word topics, as opposed to documents. The model makes the conceptual density in a dataset and creates topic inference slightly sensitive to differences in text length and how topics are spread out. Previous studies showed that using fuzzy clustering for extracting topics from documents also improved the performance for classification and clustering tasks (Rashid et al., 2022).

An efficient topic model derived from the Dirichlet Multinomial Mixture (DMM) model is called GPU-DMM (Li, Duan, Wang, Zhang, Sun, and Ma, 2017). They use the extended Polya urn (GPU) model for short texts, which uses auxiliary embeddings to get generic word semantic information (Mahmoud, 2008). PTM (Zuo, Wu, Zhang, Lin, Wang, Xu, and Xiong, 2016b) assumes that a substantial majority of text documents are produced from a small number of frequent texts and that the idea of a "pseudo document" is used to affirmatively group shorter text together in the presence of sparse data without the necessity of further context. TRNMF (Yi, Jiang, and Wu, 2020) topic model utilizing regularized nonnegative matrix factorization for short documents. Some methods (Gruber et al., 2007) attempt to reduce the sparsity problem by presuming that terms in all sentences interact with the same topic. In addition, the findings of these topic models are typically attained at the posterior in topics, which makes the topic model susceptible to overfitting (Blei et al., 2001).

Therefore, in this research, we presented a DistilBERTopic model that discovers the semantically relevant quality topics and removes the sparsity issue for short text, where probabilities of documents and topics are defined. DistilBERTopic model used the pre-trained transformer-based language models, and before Density-based Spatial clustering, the detrimental impact of high dimensionality is minimized by singular value decomposition. DistilBERTopic model is compared with the state-of-the-art short text topic models using real-world short text document datasets. Experimental results show that DistilBERTopic performs better than other models in terms of topic coherence and classification.

2 Methodology

Consider a variety of Z short text containing vocabulary of size V, which are denoted by $X = x_1, x_2, x_3, ..., x_V$ and K is the number of topics. Dirichlet parameters are α and β.

2.1 Pre-processing

The text data probably contain a significant amount of noise, including different word forms, stop words, punctuation, and special characters. The text data is converted to lowercase to eliminate any potential confusion caused by word variances. The text is first broken up into phrases, which are subsequently tokenized into individual words. A document is broken down into tokens. Stop words are eliminated. Words are normalized by deploying Porter's stemmer algorithm (Patil and Sandip, 2013; Porter, 1980), which culminates in eliminating inflectional endings for the words.

2.2 DistilBERT

DistilBERT (Sanh, Debut, Chaumond, and Wolf, 2019) is developed from BERT by applying knowledge distillation (Kenton and Toutanova, 2019). DistilBERT is a compact Bidirectional Encoder Representation of BERT that preserves the BERT comprehension capabilities by adopting a knowledge distillation technique. The model is distilled in very large batches through the use of dynamic masking and with the assistance of the next sentence prediction. In this context, masking and next sentence prediction refer to the procedure in which a word that is to be predicted is transformed to the value ["MASK"] in the Masked Language model, and the entire sequence is trained to predict that specific word. The trained model aids in establishing the context of words by attempting to identify the meaning of a document. The implementation comprises a loss function comprised of a distillation loss and a cosine embedding loss. To build a more compact version of BERT, the architects of DistilBERT eliminated token-type embeddings and the pooler from the architecture and decreased the number of layers by a factor of two. DistillBERT is used to turn the documents into embeddings.

2.3 Singular Value Composition

The documents with related topics are clustered together to discover the topics in these clusters. The embedding dimensionality is reduced because many clustering methods poorly handle high dimensionality. To reduce the negative effects of higher dimensionality, we apply singular value decomposition (SVD), a well-known technique for reducing data dimension before clustering (Fodor, 2002).

2.4 Hierarchical Density-based Spatial Clustering

The HDBSCAN (Hierarchical Density-based Spatial Clustering) (Campello, Moulavi, and Sander, 2013) is used. HDBSCAN clustering technique represents clusters and allows noise to be treated as outliers. When dealing with noise and varied cluster densities, HDBSCAN is used to discover the dense regions of document vectors. The utilization of HDBSCAN is motivated by the fact that it produces only significant clusters and does not cluster noise. Thus, compared to other clustering algorithms, the quality of the clusters is high. The interactive use of cluster selection epsilon, which hierarchically mixes and separates clusters, allows us to control the size of the clusters. This allows us to discover more specific topics within a specific cluster.

2.5 Probability of the Documents

The probability of Z documents j is calculated by equation 1. Where n is the amount of data.

$$P(Z_j) = \frac{\sum_{i=1}^{m} (X_i, Z_j)}{\sum_{i=1}^{m} \sum_{j=1}^{n} (X_i, Z_j)} \tag{1}$$

2.6 Probability of the Documents to Topics

Equation 2 calculates the probability for documents j with topics k.

$$P(Z_j, Y_k) = P(Y_k|Z_j) \times P(Z_j) \tag{2}$$

Then, for each topic, the normalization probability of documents in the topic is defined by equation 3.

$$P(Z_j|Y_k) = \frac{P(Z_j|Y_k)}{\sum_{j=1}^{n} P(Z_j|Y_k)} \tag{3}$$

2.7 Probability of the Words in Documents

Equation 4 finds the probability of words in the documents.

$$P(X_i|Z_j) = \frac{P(X_i|Z_j)}{\sum_{j=1}^{m} P(X_i|Z_j)} \tag{4}$$

Table 1: Dataset statistics

Datasets	Labels	Z	X	V
TMNews	7	32503	4.9	6347
Twitter	4	2520	5.0	1390

3 Experiments and Results

In this section, DistilBERTopic model is compared with other state-of-the-art topics models. The classification and topic coherence results are given for two real-world datasets TMNews and Twitter.

3.1 Datasets

TWNews and Twitter datasets are selected for the experiments due to the diversity between datasets. TWNews dataset is English news articles taken from the RSS feeds of three prominent newspaper websites[1]. The dataset comprises business, sports, health, U.S., science technology, world, and entertainment. We keep news descriptions because it is often comprised of brief sentences.

The Twitter corpus contains categorized tweets[2]. These tweets are assigned to one of four categories: Apple, Google, Microsoft, and Twitter. Table 1 shows the statistics of the datasets.

3.2 Baseline Topic Models

We compared the presented DistilBERTopic model with BTM (Cheng, Yan, Lan, and Guo, 2014),LF-DMM (Nguyen, Billingsley, Du, and Johnson, 2015), WNTM (Zuo, Zhao, and Xu, 2016d), GPU-DMM (Li, Duan, Wang, Zhang, Sun, and Ma, 2017), PTM (Zuo, Wu, Zhang, Lin, Wang, Xu, and Xiong, 2016b) and TRNMF (Yi, Jiang, and Wu, 2020) over short text data. The topic models BTM, WNTM, and GPU-DMM all use the same hyperparameter values of $alpha = 50/K$ and $beta = 0.01$. For WNTM, the sliding window length was set at 10. As indicated by the authors, for LF-DMM, we adjusted the parameters $\lambda = 0.6$, $\alpha = 0.1$ and $\beta = 0.01$. We used the values of $\alpha = 0.1$, $\lambda = 0.1$, and $\beta = 0.01$, respectively, for PTM and TRNMF. Therefore, in the evaluation of experiments, $\alpha = 0.1$, $\beta = 0.01$ and $\lambda = 0.1$ values are set. The Gibbs sampling method is applied to each model for a total of 1,000 iterations, with the

[1] (http://acube.di.unipi.it/tmn-dataset/), (nyt.com, usatoday.com, reuters.com),

[2] (http://www.sananalytics.com/lab/index.php)

Table 2: Classification accuracy for TMNews and Twitter datasets with 30, 50 and 90 topics

Dataset	Model	K=30	K=50	K=90
TMNews	BTM	0.626	0.526	0.395
	LF-DMM	0.635	0.592	0.658
	WNTM	0.705	0.691	0.701
	GPU-DMM	0.424	0.364	0.336
	PTM	0.443	0.310	0.296
	TRNMF	0.763	0.735	0.646
	DistilBERTopic	**0.785**	**0.757**	**0.668**
Twitter	BTM	0.586	0.474	0.272
	LF-DMM	0.183	0.234	0.241
	WNTM	0.810	0.807	0.764
	GPU-DMM	0.683	0.568	0.510
	PTM	0.340	0.0.248	0.267
	TRNMF	0.821	0.816	0.771
	DistilBERTopic	**0.842**	**0.837**	**0.792**

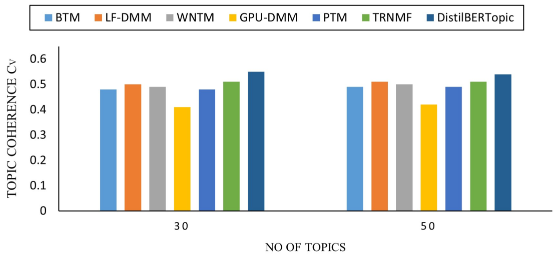

Figure 1: Topic coherence with TMNews dataset with 5 topic words

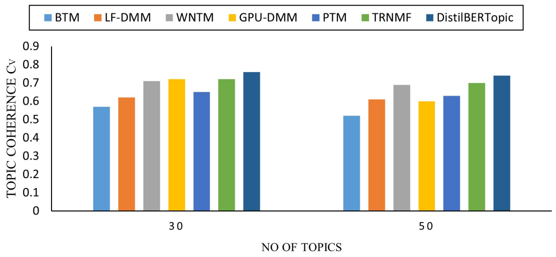

Figure 2: Topic coherence with Twitter dataset with 5 topic words

number of latent topics set to 30, 50, and 90.

3.3 Classification

We represent each document using topic modeling by using P(Y|Z) for topic distribution. P(Y|Z) means the probability of topics with the documents. P(Y|Z) represents the probability of a given topic appearing in a given set of documents. As a result, the topic's quality is efficiently assessed using text classification accuracy. The high classification accuracy shows that the topics are more discriminate and comprehensive. We used Weka for the classification with Naive Bayes. A 5-fold cross-validation method is utilized to assess classification accuracy. The classification accuracy for two datasets with baseline topic models is shown in Table 2. In terms of accuracy of classification across a diverse range of topics, DistilBERTopic model performs significantly better than the other topic models. The classification results showed that DistilBERTopic performs better than several baseline topic models for both datasets with 30, 50, and 90 topics.

3.4 Topic Coherence

Topic coherence is determined by the co-occurrence of words in external corpora. It is revealed that a correlation exists between topic coherence and human judgments and that this correlation has a high degree of generalizability. Topic coherence numerous approaches have been presented for the automatic assessment of individual topics and the automatic evaluation of entire topic models (Newman, Lau, Grieser, and Baldwin, 2010; Lau, Newman, and Baldwin, 2014). We prefer to use the CV approach (Röder, Both, and Hinneburg, 2015). This consistency metric retrieves the co-occurrence value counts of the specified words using a sliding window. The normalized point-wise mutual information calculates co-occurrence counts (NPMI) (Bouma, 2009) between each top word. Equation 5 is used to calculate the NPMI score.Where, $P(w_i)$ probability of encountering the word w_i in any text and $P(w_i, w_j)$ probability of finding the words w_i and w_j together in a randomized documents. The most likely word sequence is $x, 1$,x, 2, $x, 3, ...$x, with N as the total.

$$NPMI\left(x_i, x_j\right) = \sum_{j}^{N-1} \frac{log \frac{P(x_i,x_j)}{P(x_i)P(x_j)}}{-logP(x_i, x_j)} \qquad (5)$$

Figures 1 and 2 demonstrate the TWNews and Twitter topic coherence outcomes of DistilBERTopic model and all comparable topic models. We specifically set T = 5 for the number of top words per topic, K = 30, and 50 as the number of topics.

DistilBERTopic model outperforms and competing others baseline topic models on all two datasets, whereas the WNTM model beats other models on Twitter. Our proposed topic model gives a higher performance in comparison to WNTM and LF-DMM. PTM performs the best among baseline approaches on TWNews datasets, but BTM provides the lowest coherence score. Despite poor prior results, GPU-DMM outperforms LF-DMM in terms of topic coherence. GPU-DMM performs poorly on TMnews, which may indicate that news descriptions frequently encompass multiple topics. On the other hand, GPU-DMM gives a fairly high score of topic coherence for Twitter, which may mean that titles in Twitter data hide rarer topics than news descriptions. Overall, the DistilBERTopic model achieved higher topic coherence results than other baseline topic models.

4 Conclusion

Finding informative content is becoming more challenging as the volume of short texts available increases. In the absence of context information, the short text has sparseness issues. In this paper, we presented the DistilBERTopic model, which extracts semantically coherent topics from short text and ameliorates the sparsity issue. The document embedding is constructed with pre-trained transformer-based language models and clustered using Hierarchical Density-based Spatial Clustering. The singular value composition method reduced the higher dimensionality effect before clustering. We conducted comprehensive experiments on two short corpora of real-world short text data. The experimental outcomes demonstrate that the DistilBERTopic model is more effective and efficient than existing state-of-the-art topic models. DistilBERTopic model achieved better classification and topic coherence results. We will use other word embedding methods with hierarchical and partitioning clustering in the future.

Acknowledgements

This research was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.2020R1I1A3069700) and by the National University Development Project by the Ministry of Education in 2022.

References

David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. *Advances in neural information processing systems*, 14.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Imola K Fodor. 2002. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US).

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *Artificial intelligence and statistics*, pages 163–170. PMLR.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Liangjie Hong and Brian D Davison. 2010a. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88.

Liangjie Hong and Brian D Davison. 2010b. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Chenliang Li, Yu Duan, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2017. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems (TOIS)*, 36(2):1–30.

Hosam Mahmoud. 2008. *Pólya urn models*. Chapman and Hall/CRC.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892.

Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. 2021. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Chaitali G Patil and Patil Sandip. 2013. Use of porter stemming algorithm and svm for emotion extraction from news headlines. *International Journal of Electronics, Communication and Soft Computing Science and Engineering (IJECSCSE)*, 2(7):9.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multilabeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256.

Junaid Rashid, Jungeun Kim, Amir Hussain, Usman Naseem, and Sapna Juneja. 2022. A novel multiple kernel fuzzy topic modeling technique for biomedical data. *BMC bioinformatics*, 23(1):1–19.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yuan Wang, Jie Liu, Jishi Qu, Yalou Huang, Jimeng Chen, and Xia Feng. 2014. Hashtag graph based topic model for tweet mining. In *2014 IEEE International Conference on Data Mining*, pages 1025–1030. IEEE.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270.

Feng Yi, Bo Jiang, and Jianjun Wu. 2020. Topic modeling for short texts via word embedding and document correlation. *IEEE Access*, 8:30692–30705.

Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. 2016a. Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2105–2114.

Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. 2016b. Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2105–2114.

Yuan Zuo, Jichang Zhao, and Ke Xu. 2016c. Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 48(2):379–398.

Yuan Zuo, Jichang Zhao, and Ke Xu. 2016d. Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 48(2):379–398.