International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 2, Number: 1, Pages: 46- 51, Year: 2023

Enhancement of Embedded Topic Model

Asma Gul, S. Zafar Ali Shah and Sadaqat Jan

Department of Computer Software Engineering, University of Engineering & Technology Mardan, Pakistan
Corresponding author: First Author (E-mail: asmagul1882@gmail.com)

Received: 12/03/2023, Revised: 22/05/2023, Accepted: 10/06/2023

Abstract— This **comparative study aims to examine the performance of the Embedded Topic Model (ETM) in producing coherent topics when applied to noun-restricted corpora. As nouns in any dataset are the most informative features involving them will improve the topic quality. To evaluate this hypothesis, we compare the performance of two topic models: ETM (Embedded Topic Model) and LDA (Latent Dirichlet allocation), on three dataset variations. The first dataset is the original pre-processed dataset, while the second version consists of the dataset reduced to noun phrases only; the third version represents the dataset reduced to nouns only. To assess the performance of both models, we employ two widely used measures: Topic Coherence (TC) and Topic Diversity (TD). The experimental results revealed that the embedded topic model outperforms LDA across all the variations of the datasets. Remarkably it exhibits exceptional performance for the dataset having only nouns. In addition, the time to train the model is also reduced when the vocabulary is reduced to nouns only. This paper evaluates how the Embedded Topic Model significantly improves topic quality, especially in noun-restricted contexts. These findings provide insightful information for researchers and practitioners regarding the possible advantages of using noun-based corpus reduction strategies in topic modelling tasks.**

Index Terms— **Text mining, topic modelling, text summarization, topic diversity, embedded topic modelling.**

I. Introduction

Due to the rapid growth of online textual data such as social media platforms, discussion forums, surveys, and personal blogs, Topic modelling is a technique to find patterns in a set of documents; these patterns are then considered as hidden "topics" present in the dataset. Topic modelling is quite challenging when the text is shorter, for example, when the dataset consists of tweets or product reviews instead of articles. Such short text doesn't have a context where LDA and its variants usually work based on word co-occurrences; therefore, these models don't perform satisfactorily on short text. similarly, BTM (Bit Term topic model) overcomes the data sparsity problem to some extent, but these probabilistic models require high run time and computational complexity. The output of these topic modelling techniques is difficult to interpret [1].

Pre-trained word embedding such as wordtovec are trained on large datasets which can capture word semantics in different contexts. where context is highly important in the disambiguation of meanings and understanding the actual meaning of the words, as each word in a document is related to its context, that is, words with similar contexts target a similar topic. Word embedding is a popular word representation technique useful for finding semantically. meaningful topics. It generates vector representations for words and is a good measure of semantic relatedness[2].

Similar to human beings who understand the text based on their background knowledge, machine learning algorithms also require (background knowledge) semantic relationships between words which is difficult to find in short text. Still, this problem can be solved by using word embedding, as it can capture general word co-occurrence patterns. This study is a continuation of [3] in which the author has suggested an embedded topic model (ETM). In this study, our main focus is on the nouns in the dataset. As nouns in the text extract specific types of entities, such as people, organizations or place names, in fact, any concrete thing with a name. Named entities have much expressive power, and involving them in a topic will improve topic quality. Short texts such as news mainly discussed named entities/nouns such as people, organizations, places and events. That's why if nouns/named entities are included in the topic modelling process, it will positively affect topic quality. Most corpuses have large vocabularies, and reducing the size of the dataset to nouns only before topic modelling will ultimately reduce the time to train models on such datasets.

In this article, we sought to find that whether there is any improvement in topic quality by limiting the corpus to nouns only or not. To achieve our goal, we have compared the performance of ETM and LDA on three different dataset versions. The results showed that ETM has shown better performance and produces more meaningful topics on the corpus version reduced to nouns only, as measured by topic coherence and topic diversity [3]. The results also showed that more coherent topics can be produced in less time, even with a reduced corpus. Some of the related work is given as follows.

In recent years much research has been carried out that has shown that word embedding is quite useful for topic modeling. ETM model [3] combines LDA with word embedding, as LDA faces problems while dealing with a larger vocabulary, The ETM model, on the other hand, uses an embedding representation of both words and topics. D-ETM [4] (Dynamic Embedded Topic Model) extends both D-LDA and ETM. In

contrast, D-LDA is an extension of LDA that uses probabilistic time series to allow the topics to vary over time. The main difference between ETM and D-ETM is that the word embedding of D-ETM varies over time. EETM [11] is a hybrid model in which the benefits of both LDA and word embedding are combined through an integration framework. The integration framework connects these models by making them share. consistent internal semantic structure of the text content Concept embedded topic model [12] comprises three phases. In the first phase, semantically related words will be generated from the collection of documents through word Net synonyms. These words are clustered together to create a group of semantically related words.

This group of semantically related words is interpreted as a concept. In GLTM(Global and Local word embedding-based Model) [5], word embedding trained on large external datasets and a continuous skip-gram model with negative sampling is used to obtain local word embedding. In [15] and [18], the word embedding sequences are directly modelled by assuming that topics are multivariate Gaussian distributions in the embedding space or von Mises-Fisher models. The focused topic model [16] is a model where a topic focuses on words and is informed by word embedding. WELDA(Word embedding and LDA) [6]is a new model combining the positive points of latent Dirichlet allocation and word embedding to form a new topic model to improve topic quality. It works on the bag of words assumption and generates topics based on global context.

In a clustering-based topic model [7], a word network graph is used, such as the network nodes representing different definitions of words and phrases and edges representing the similarity between words based on word embedding. The cross-contextual word embedding model [8] obtains the first global word embedding for each word. The second part obtains the local word embedding for each polysemous word in different contexts. A word embedding is adaptively adjusted and updated concerning different contexts. R-BERT (Relational Bidirectional Encoder Representations from Transformers) [9] is an extension of BTM (Bit Term Topic Model), which aims to solve the problem of sparsity in the short text. In R-BTM, short texts are linked using word embedding.

Latent Feature Topic Models (LFTM) [10] is a model in which word embedding is used to sample words from the multinomial topic distribution and the embedding space. Distribution over all the words is required for sampling in the embedding space. Gaussian mixture topic model (GMTM) [17] proposes that topics are multivariate Gaussian distributions on the embedding space and then directly models sequences of word embedding. The model [19] jointly learns word embedding and latent topics over the dataset. However, this model performs well for long text, not short text.

A generative topic embedding model [20] and [21] combines the two types of patterns that are topic Word embedding maps words into a low-dimensional continuous embedding space, and topic modelling maps documents onto a low-dimensional topic space. The model [22] directly takes in word semantic relations learned from a large text dataset, which is domain-free and easy to access and uses relatedness knowledge based on word embedding with the GPU model. LCTM (Latent Concept Topic Model) [23] generates topics via the co-occurrence of latent concepts, where concepts are the clusters of conceptually similar words in embedding space. Some researchers propose global topic embedding vectors such as [24] and [25] to get the dataset-level embedding vector; they average the embedding of words in the same topic. The novel correlated topic model [26] exploit the additional word-level correlation information in word embedding and directly model topic correlation in the continuous word embedding space. A new topic model [27] uses an external source to train word embedding upon it. The resulting semantic regularities are then used as supplementary information to overcome the data sparsity problem in short texts. Model [28] uses the von Mises-Fisher distribution to model the density of words over a unit sphere. This model naturally exploits the semantic structures of word embedding while flexibly discovering several topics.

Some models use specific speech parts for topic modelling, such as the composite model [29-35], which can capture the interaction between short- and long-range word dependencies. It can simultaneously learn syntactic classes and semantic topics and identify words' roles in documents. It is competitive in part-of-speech tagging and classification with models specializing in only one dependency form. A hybrid model [30] embeds hidden Markov models (HMMs) within LDA topics to jointly model the topics and the syntactic structures within each topic. Part-of-Speech LDA (POSLDA) [13], is a syntactically and semantically consistent generative probabilistic model. This model discovers POS-specific topics from an unlabeled dataset. In [14], the author suggests that eliminating all words except nouns would provide an alternative to finding the most informative features of a dataset.

II. MATERIALS AND METHODS

The proposed work is a comparative study to compare the performance of ETM against LDA on three versions of the dataset (Original dataset, Noun phrase only dataset, Nouns only Dataset). We measure the performance of both models in terms of topic interpretability and topic diversity.

TABLE I.
TOP 5 WORDS OF THE FIVE TOPICS GENERATED BY ETM AND LDA

(a) LDA				
Topics with Original Dataset				
Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**
show	campaign	percent	house	son
music	state	money	building	family
film	new	bank	water	home
art	party	fund	food	mother
book	election	share	apartment	friend
Topics with Noun phrases only Dataset				
Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**
case	year	time	company	first
Law	city	show	Year	second
People	school	way	Business	team
Year	home	thing	Money	player
Police	work	people	Percent	two
Topics with Nouns only Dataset				
Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**
percent	season	case	campaign	room
company	team	police	state	water
year	game	government	government	art
market	play	court	country	city
business	time	law	election	space
(b) ETM				
Topics with Original Dataset				
Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**
son	American	game	house	health
friend	democratic	season	building	medical
family	clinton	team	room	researchers
house	sander	player	food	drugs
man	london	win	city	doctors
Topics with Noun phrases only Dataset				
Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**
health	state	media	case	first
medical	government	world	office	game
researchers	officials	times	lawyer	team
test	campaign	wrote	killing	second
drugs	political	news	dead	player
Topics with Nouns only Dataset				
Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**
family	company	team	show	trump
son	director	game	music	island
man	business	player	film	sander
mother	executive	league	artist	democratic
wife	firm	victory	movie	obama

A. Text Mining Steps

The following steps are performed before topic modeling.

SELECTION OF DATASET: The NYT (New York Times) dataset is used which consist of 8888 number of articles comprise of different topics.

TEXT PREPROCESSING**:** After selection of data set, the following text preprocessing steps are performed to remove noise, stop words and redundancies from dataset.

TOKENIZATION: It splits the text sentences in to words removing the blank spaces and punctuations.

STOP WORDS REMOVAL: The stop words are removed from the dataset.

B. Proposed Approaches for Topic Extraction

Previous study [14] suggests that nouns can be used to find the most informative feature of dataset. Keeping this in mind we have suggested three approaches for topic extraction. The three approaches are:

APPROACH 1:

In this approach the original preprocessed dataset is used for extracting topics through ETM and LDA topic models.

APPROACH 2:

In this approach before extracting topics through ETM and LDA an NLTK POS tagger is used to tag the tokenized text as NN (nouns), VB (verbs), JJ (adjectives) etc.) In the dataset. After tagging the dataset the nouns are extracted and saved in another data frame.

APPROACH 3:

In this approach before extracting the topics the noun phrases are extracted from the dataset and saved in another data frame.

III RESULTS AND DISCUSSION

The proposed approach is tested on NYT (New York Times) dataset which consist of 8888 number of articles.
The quality of topics produced by both the ETM and LDA on three different versions of the dataset has been tested on the basis of two metrics such as topic coherence and topic diversity. The number of topics is set to K= {40} for each model.

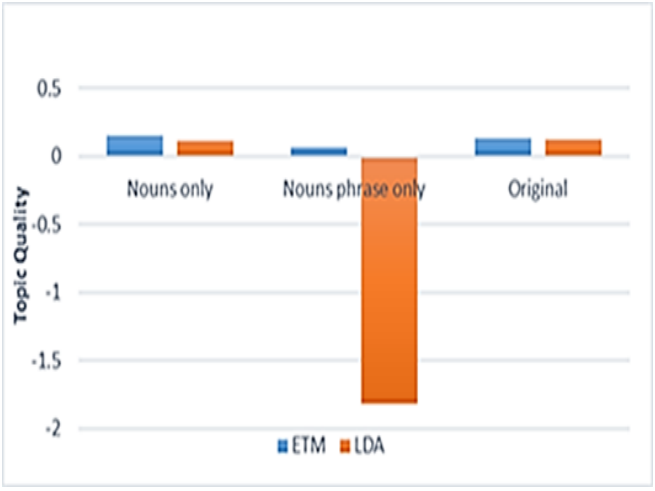

Fig.1 Graphical comparison of Topic quality

TABLE II.
TOPIC QUALITY ON NYT DATASET K=40

Model	Dataset version	TC	TD	TQ
ETM	Nouns only	0.169	0.914	0.155
	Nouns phrase only	0.071	0.98	0.069
	Original	0.158	0.872	0.137
LDA	Nouns only	0.127	0.9125	0.115
	Nouns phrase only	-2.045	0.8875	-1.814
	Original	0.150	0.837	0.125

A. Quantitative Analysis:

We study the two models quantitatively. We measure the quality of the topics on the basis of two metrics used in study [10] known as Topic coherence and topic diversity. Both these metrics are also used in current study. The topic coherence is the measure that shows that how frequently words co –occur in a dataset [32, 36-39]. The higher the topic coherence the more interpretable is the topic
model. According to [32], the coherence for a topic can be calculated as

$$TC = \sum_{M=2}^{m} \sum_{l=1}^{m-1} log \frac{\left|D(w_m^k, w_l^k)\right| + 1}{\left|D(w_l^k)\right|}$$

Where $w^k = [w_l^k, w_l^k, w_l^k \ldots \ldots w_m^k]$ is the top M words under the topic k sorted by probability in descending order. $\left|D(w_m^k, w_l^k)\right|$ Is the number of documents in dataset containing both word w_m^k *and* w_l^k.The main idea behind the diversity is the percentage of unique words in all the topics, the value close to 0 indicate more redundant topics whereas the value close to 1 indicate varied topics.
The topic quality is the product of topic coherence and topic diversity. Table 1 shows that the quality of topics generated by ETM is better than LDA. The results in TABLE II also shows that the topic quality is enhanced while reducing the dataset to nouns only. The coherence score for noun only dataset shows that it will have lower number of junk topics.

B. Qualitative Analysis:

We next study the models qualitatively. TABLE I illustrates the top 5 topics generated from three different approaches by ETM and LDA. The table shows that ETM on noun only dataset gives the highly interpretable topics as compared to LDA and other versions of the dataset. Cosine similarity score in Table III shows that the words in the topics generated by ETM on nouns only dataset are semantically meaningful.

TABLE III.
FIVE WORDS SELECTED FROM THE TOPICS GENERATED BY ETM ON NOUNS ONLY DATASET ALONG WITH THE COSINE SIMILARITY WITH OTHER WORDS IN THE SAME TOPIC

Word	Topic words with cosine similarity
Car	vehicle (0.1098) truck (0.1093) bus (0.093) driver (0.104) tow (0.051) plane (0.075) charges (0.011) freeway (0.04) crashes (0.071)
medicine	Psychiatry (0.095) biology (0.0086) professor (0.087) metabolism (0.025) pathology (0.081) molecular (0.054), preventive (0.075)
film	Movie (0.11) theater (0.081) art (0.070) starred (0.090) novel (0.085) sequel (0.085) concert (0.086) Showrunner (0.0071) television (0.081)
house	Manor (0.065) apartment (0.091) building (0.098) kitchen (0.077) auction (0.059) mansion (0.095) Clapboard (0.045) stucco (0.046)
food	Pantry (0.103) beer (0.078) madeleine (0.0138) necessities (0.086) grocery (0.0885) stalls (0.058) menu (0.073) nutrition (0.024)

TABLE IV.

Model	Dataset version	Time
ETM	Nouns only	2m 21s
	Nouns phrase only	23 m
	Original	34 m
LDA	Nouns only	10m 5 s
	Nouns phrase only	26m
	Original	36m

TIME TO GENERATE 40 TOPICS

C. *Efficiency*

As given in TABLE II that topics generated by ETM on dataset reduced to nouns not only give more interpretable topics but as the vocabulary is reduced than the time to generate topics is also reduced. As there are further steps involved such as parts of speech tagging (POS tagging) etc., where POS tagging takes less than a minute. If the other steps are ignored than comparatively the time taken by nouns only dataset is less than the other versions of the dataset.

TABLE I shows that ETM performs better than LDA on all the three approaches, but ETM itself performs better on approach 2 compare to approach 1 and approach 3. The coherence score in TABLE II prove that nouns only dataset could produce semantically more coherent topics than original dataset. Also topic modeling with original dataset on both ETM and LDA shows that the frequency of nouns is high in topics as compared to other part speech such as verbs etc.

It shows that removing extra vocabulary may not cause any problem and will produce positive results as nouns have a high frequency in the dataset. Along with this noun phrases version of the dataset do not shows any good results both for ETM and LDA, but the coherence score for LDA for nouns phrases dataset is much worse than ETM. The reason for worse result of nouns phrase dataset is given in [33] which states that the semantics of many phrases do not necessarily relate to their component words in natural language. Co-occurrence information of phrases is far lower than single word, which results in a very low probability of many phrases in topics. This lead to the Phrase LDA to be worse. Whereas the ETM shows better results as compare to LDA because it uses word embedding for contextual information. TABLE IV shows that the time to generate topics from the reduced data set is much less than the other versions of the corpus.

IV. CONCLUSION

This is a comparative study to show that Embedded topic model (ETM) performance can be further improved by reducing the dataset to nouns only.

This study found that after limiting the dataset to nouns only Embedded Topic Model (ETM) gives highly coherent and diverse topics, along with this as the vocabulary size is reduced to nouns only it will ultimately result in less training time. But as other steps are involved so we cannot claim it strictly but further investigation will be done in this regard in future work.

REFERENCES

[1] B. M. Schmidt, “Words alone: Dismantling topic models in the humanities,” Journal of Digital Humanities, 2012.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of word representations in vector space,” .2013.

[3] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, “Topic Modeling in Embedding Spaces,” Transactions of the Association for Computational Linguistics, vol. 8, pp. 439–453, Jul. 2020.

[4] A. B. Dieng, Joaquín Rodríguez Vidal, and D. M. Blei, “The Dynamic Embedded Topic Model.,” Jan. 019.

[5] W. Liang, R. Feng, X. Liu, Y. Li, and X. Zhang, “GLTM: A Global and Local Word Embedding-Based Topic Model for Short Texts,” IEEE Access, vol. 6, pp. 43612–43621, 2018.

[6] S. Bunk and Ralf Krestel, “WELDA,” May 2018.

[7] W. Mu, K. H. Lim, J. Liu, S. Karunasekera, L. Falzon, and A. Harwood, “A clustering-based topic model using word networks and word embeddings,” Journal of Big Data, vol. 9, no. 1, Apr. 2022.

[8] S. Li, R. Pan, H. Luo, X. Liu, and G. Zhao, “Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modeling,” Knowledge-Based Systems, vol. 218, p. 106827, Apr. 2021.

[9] X. Li, A. Zhang, C. Li, L. Guo, W. Wang, and J. Ouyang, “Relational Biterm Topic Model: Short-Text Topic Modeling Using Word

Embeddings," The Computer Journal, vol. 62, no. 3, pp. 359–372,May 2018.
[10] 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, Dec. 2015.
[11] P. Zhang, S. Wang, D. Li, X. Li, and Z. Xu, "Combine Topic Modelingwith Semantic Embedding: Embedding Enhanced Topic Model," IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, 2019.
[12] Dakshi Tharanga Kapugama Geeganage, "Concept Embedded Topic Modeling Technique," Jan. 2018.
[13] W. M. Darling, M. J. Paul, and J. W. Wells, "Unsupervised Part-of-Speech Tagging in Noisy and Esoteric Domains With a Syntactic Semantic Bayesian HMM," pp. 1–9, Apr. 2012.
[14] F. Martin and M. H. Johnson, "More Efficient Topic Modelling Through a Noun Only Approach," pp. 111–115, Dec. 2015.
[15] R. Das, M. Zaheer, and C. Dyer, "Gaussian LDA for Topic Models with Word Embeddings," Jul. 2015.
[16] H. Zhao, L. Du, and Wray Buntine, "A Word Embeddings Informed Focused Topic Model," pp. 423–438, Nov. 2017.
[17] V. K. Rangarajan Sridhar, "Unsupervised Topic Modeling for Short Texts Using Distributed Representations of Words," Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015.
[18] Naseer, Fawad, Muhammad Nasir Khan, Akhtar Rasool, and Nafees Ayub. "A novel approach to compensate delay in communication by predicting teleoperator behaviour using deep learning and reinforcement learning to control telepresence robot." Electronics Letters 59, no. 9, e12806, 2023.
[19] Naseer, Fawad, Muhammad Nasir Khan, and Ali Altalbe. "Intelligent Time Delay Control of Telepresence Robots Using Novel Deep Reinforcement Learning Algorithm to Interact with Patients." Applied Sciences 13, no. 4, 2462, 2023.
[20] Naseer, Fawad, Muhammad Nasir Khan, and Ali Altalbe. "Telepresence Robot with DRL Assisted Delay Compensation in IoT-Enabled Sustainable Healthcare Environment." Sustainability 15, no. 4, 3585, 2023.
[21] Altalbe, Ali, Muhammad Nasir Khan, Muhammad Tahir, and Aamir Shahzad. "Orientation Control Design of a Telepresence Robot: An Experimental Verification in Healthcare System." Applied Sciences 13, no. 11, 6827, 2023.
[22] Khan, Muhammad Nasir, Syed K. Hasnain, and Mohsin Jamil. Digital Signal Processing: A Breadth-first Approach. Stylus Publishing, LLC, 2016.
[23] Naseer, Fawad, Muhammad Nasir Khan, Akhtar Rasool, and Nafees Ayub. "A novel approach to compensate delay in communication by predicting teleoperator behaviour using deep learning and reinforcement learning to control telepresence robot." Electronics Letters 59, no. 9, e12806, 2023.
[24] X. Li, J. Chi, C. Li, J. Ouyang, and B. Fu, "Integrating Topic Modeling with Word Embeddings by Mixtures of vMFs," pp. 151–160, Dec.2016.
[25] B. Shi, W. Lam, S. Jameel, S. Schockaert, and Kwun Ping Lai, "Jointly Learning Word Embeddings and Latent Topics," Aug. 2017.
[26] S. Li, T.-S. Chua, J. Zhu, and C. Miao, "Generative Topic Embedding: Continuous Representation of Documents," Jan. 2016,
[27] S. Li, T.-S. Chua, J. Zhu, and C. Miao, "Generative Topic Embedding: a Continuous Representation of Documents (Extended Version with Proofs)," Jun. 2016.
[28] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic Modeling for Short Texts with Auxiliary Word Embeddings," Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16, 2016.
[29] W. Hu and Jun'ichi Tsujii, "A Latent Concept Topic Model for Robust Topic Inference Using Word Embeddings," Jan. 2016.
[30] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical Word Embeddings," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, no. 1, Feb. 2015.
[31] J. Law, Hankz Hankui Zhuo, J.-H. He, and E. Rong, "LTSG: Latent Topical Skip-Gram for Mutually Learning Topic Model and Vector Representations," Feb. 2017.
[32] Guangxu Xun, Y. Li, Wayne Xin Zhao, J. Gao, and A. Zhang, "A Correlated Topic Model Using Word Embeddings," Aug. 2017.
[33] G. Xun, V. Gopalakrishnan, F. Ma, Y. Li, J. Gao, and A. Zhang, "Topic Discovery for Short Texts Using Word Embeddings," IEEE Xplore,Dec. 01, 2016.
[34] K. Batmanghelich, A. Saeedi, K. Narasimhan, and S. Gershman, "Nonparametric spherical topic modeling with word embeddings", Proc. Conf. Assoc. Comput. Linguist. Meet., vol.2016, pp. 537–542, 2016.
[35] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, "Integrating Topics and Syntax," vol. 17, pp. 537–544, Dec. 2004.
[36] J. Jiang, "Modeling Syntactic Structures of Topics with a Nested HMM-LDA," Dec. 2009.
[37] Blei, D. M., Ng, A. Y., and Jordan, M. I,"Latent Dirichlet allocation". Journal of Machine Learning Research, 3:993 1022.2003.
[38] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum." Optimizing semantic coherence in topic models",proceedings of the conference on empirical methods in natural language processing,pages 262-272,2011.
[39] J. Ma, J. Cheng, L. Zhang, L. Zhou, and B. Chen, "A Phrase Topic Model Based on Distributed Representation," Computers, Materials