**A SEMINAR REPORT ON**

**Topic Modeling for Short Texts via Word Embedding and Document Correlation**

PRESENTED BY**: -**

**KUKADE SHRAYESHA NILKANTH**

**UNDER GUIDANCE OF**

**Assistant Prof. M. M. Swami**



**DEPARTMENT OF COMPUTER ENGINEERING**

**AISSMS COLLEGE OF ENGINEERING**

**KENNEDY ROAD, PUNE 411001**

**AFFILATED TO**

**UNIVERSITY OF PUNE**

**AISSMS College of Engineering**

**Department of Computer Engineering Kennedy road, near RTO, pune-411001**



**CERTIFICATE**

This is to certify that the

### “Shrayesha N. Kukade“

from Third Year Computer Engineering has successfully completed her seminar work titled ‘**Topic Modeling for Short Texts via Word Embedding and Document Correlation**’ at AISSMS COE, Pune in the particular fulfillment of the Bachelor’s Degree in Computer Engineering. This work is done during year 2019-2020.

#### Date: / /

**(Assistant Prof. M. M. Swami) (Dr. D. P. Gaikwad) (Dr. D. S. Bormane) Seminar Guide HOD. Comp. Dept Principal**

**ACKNOWLEDGEMENT**

We would like to extend my sincere gratitude and thanks to my guide **Assistant Prof. M. M. Swami**, for her valuable guidance and for giving us useful inputs and encouragement time and again, which inspired us to work harder. Due to her forethought, appreciation of the work involved and continuous imparting of useful tips, this report has been successfully completed.

We are extremely grateful to **Dr. D.P. Gaikwad**, Head of the Department of Computer Engineering, for his encouragement during the course of the project work.

We also extend our heartfelt gratitude to the staff of Computer Engineering Department for their cooperation and support.

We also take this opportunity to thank all our classmates, friends and all those who have directly or indirectly provided their overwhelming support during our project work and the development of this report.

KUKADE SHRAYESHA NILKANTH

Roll No. :17CO028

**ABSTRACT**

The algorithm TRNMF (Transition regularized non-negative matrix factorization) topic model for short texts. TRNMF model uses clustering algorithm to organize a large short text into several semantic clusters and helping topic inference efﬁciently. Due to the limited length of short text, data sparsity prevents the inference process of conventional topic models and causes unsatisfactory results on short texts. Conventional topic models based on word co-occurrences infer the hidden semantic structure from a corpus of documents.

The TRNMF model leverages pre-trained distributional vector representation of words to overcome the data sparsity problem of short texts. Meanwhile, the method employs the clustering mechanism under document-to-topic distributions during the topic inference by using Gibbs Sampling Dirichlet Multinomial Mixture model. TRNMF integrates successfully both word co-occurrence regularization and sentence similarity regularization into topic modeling for short texts. Through extensive experiments on constructed real-world short text corpus, experimental results show that TRNMF can achieve better results than the conventional topic model methods in term of topic coherence measure and text classiﬁcation task.

**TABLE OF CONTENTS**

1. **Index terms…………………………………………….06**
2. **Introduction……………………………………………07**
3. **Characteristics of TRNMF model……………………10**
4. **Literature survey………………………………………11**
5. **System Architecture………………………………..….14**
6. **Theoretical study……………………………………....15**
7. **Algorithm ………………………………………………24**
8. **Experiment……………………………………………..26**
9. **Advantages and challenges……………………………31**
10. **Conclusion………………………………………..32**
11. **Future scope……………………………………...32**
12. **Appendix**

References……………………………...…...33

**INDEX TERMS**

1. **Topic model**

Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrases pattern within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.

1. **Short texts**

Short texts are mainly abbreviation. Abbreviation is the shortest form of a word or a phrase used mainly in writing/typing to

represent the complete form.

1. **Word embedding**

a word embedding is a learned representation for text where words with same meaning have similar representation.

1. **Document correlation**

Mutual relationship of two or more documents, words, etc.

1. **Non-negative matrix factorization**

NMF is a group of algorithms in multivariate analysis and linear algebra where a matrix ‘V’ is factorized into(usually) two matrices ‘W’ and ‘H’, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect.

1. **Regularization**

The act of bringing to uniformity, making regular.

**Chapter 1**

**Introduction**

Recent years have witnessed the increased development and popularity of various kinds of Web applications such as online social networks, recommender systems and Q&A systems. These platform services have led to an explosion of user-generated content, especially the explosively growth of short texts in a wide variety of scenarios such as blog posts, instant messaging, or product reviews. From the analytical and decision-making perspective, these data provide an unprecedented opportunity to mine user’s topic interests and understand the mechanisms of individual decisions. Hence, automatically identifying the latent topic semantic information from massive amounts of short texts has become a fundamental and challenging task in many applications, such context analysis, user interest proﬁle, and text classiﬁcation.

In conventional models, topic modeling is formalized as inferring document-to-topic and topic-to-word probabilistic distributions from the co-occurrence of words within documents, which get a good performance in many different document collections like news corpus, image data and biological information. However, the main problem is that data sparsity under short texts scenario seriously hinders the process of ﬁnding document-topic distributions due to the lack of word co-occurrence information. As a result, conventional topic modeling methods cannot be achieved satisfactory performance in short text topic modeling domain.

Considerable work has been conducted to handle the problem of data sparsity for topic modeling over short texts. For example, a simple method in the early work is to aggregate some speciﬁc short texts to reconstruct a longer pseudo-document by utilizing speciﬁed schemes such as merging all messages generated by the same author, or establishing relation information between hashtags and then conventional topic modeling methods are employed in collection of these pseudo-documents. Besides, some short texts can also hold the contextual information like URL, location, or timestamp. Many research efforts have been developed in term of these contextual information before performing topic modeling. However, these methods may be failure when the contextual information is unavailable.

To solve the problem, some studies take self-aggregation into the generative process of pseudo-documents generated. Toward another direction, these methods focus on learning the latent semantic structure information from additional auxiliary resources like long texts or knowledge bases. However, due to the highly dynamic of topics like Twitter, it is difﬁcult to match strongly related long texts via a search engine or a static knowledge base such as Wikipedia. There is also a risk that certain portion of the auxiliary data may bring unrelated noisy auxiliary topics and affect the performance of topic modeling. For large scale of short texts, the better strategy is to devise a new topic modeling method by modifying the conventional topic modeling algorithm, including modeling word co-occurrence pattern or leveraging prior knowledge sources to improve a topic model.

However, these models only explore explicit word co-occurrence information that can be always captured in a corpus of short texts, it does not take full advantage of deep semantic relations and semantic sentence structure between two words in the semantic level. Hence, research on topic modeling for short text has still many unsolved problems in this ﬁeld.

As we know, understanding length-limited short text not only depends on its literal words, but also needs to a great prior background knowledge like knowledge base or similar semantic words. It is a simple and method-independent scheme that leverages external knowledge base to alleviate the data sparseness of short texts and discover the latent semantic information over short texts. Existing works along this line largely depend on either external thesaurus (e.g., WordNet) or lexical knowledge derived from documents in a speciﬁc domain (e.g., Wikipedia). However, such naive techniques do not typically work well for modeling short texts. This is because such knowledge base must be universally applicable and expansible under kinds of scenarios. Actually, such knowledge base is not always available.

In order to enable an effective topic modeling process, TRNMF is designed to leverage the word vector representation learning during the topic inference process to deal with the data sparsity problem. The learned word representation is to make predictions within local context windows in a simple neural network architecture for language modeling [25]. Speciﬁcally, word2vec is a particularly computationally-efﬁcient predictive model for learning word embeddings from raw text. It comes in two ﬂavors, the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model. Algorithmically, these models are similar, except that CBOW predicts target words (e.g. ’mat’) from source context words (’the cat sits on the’), while the skip-gram does the inverse and predicts source context-words from the target words.

Recently, a new unsupervised learning algorithm which is named GloVe by reformulating word2vec optimizations as a special kind of factorization for word cooccurrence matrices is proposed. The GloVe model is trained on the non-zero entries of a global word-word cooccurrence matrix, which tabulate show frequently words cooccur with one another in a given corpus. Similar words being close together allow us to generalize from one sentence to a class of similar sentences. So far, a lot of pre-trained word embeddings learned from resources like Wikipedia, Twitter, and Freebase are publicly available on the Web 2.

TRNMF extends the non-negative matrix factorization model by introducing topic regularization from large text corpus in the term of topic word distribution and document regularization by employing clustering mechanism to cluster short texts in the term of document-topic, respectively. More speciﬁcally, TRNMF exploits GloVe model to get the global semantically relevant words for target words under the entire corpus.

TRNMF links the semantically relevant words together to alleviate the data sparsity problem, even if they share very limited or no co-occurrences in the current collection of short texts being modeled. Because the global word embedding vectors are trained from external documents, so the global word matrix construction is fast and ﬂexible taking in word embeddings learned from any other large text collections. Besides, TRNMF model uses clustering algorithm to organize a large short text into several semantic clusters and helping topic inference efﬁciently. TRNMF shows more prominent topics and achieves better classiﬁcation accuracy than existing state-of-the-art alternatives on constructed real-world datasets.

**Chapter 2**

**CHARSCTERISTICS OF TRNMF MODEL**

The main characteristics of this TRNMF model are summarized as follows:

* TRNMF topic model is based on regularized non-negative matrix factorization to learn the latent topic patterns over short texts. The model not only leverages the global word-word co-occurrence information learned from large text corpus to alleviate the data sparsity problem, but also uses clustering method to improve topic inference quality.
* A new word-weighting schema is used to measure the importance of a word in collection of short texts. The TRNMF method must consider the distribution characteristics of words for choosing more important words, therefore it is less sensitive to document length or document distribution over topics, and better suited for the task of short text topic mining.
* The TRNMF model is evaluated against the conventional algorithms for short texts on constructed real-world short text documents. Experimental results demonstrate TRNMF model’s superiority in term of topic coherence, classiﬁcation accuracy, and topic readability.

**Chapter 3**

**LITERATURE SURVEY**

**RELATED WORK**

In this section, we will review two lines of relevant research work:

1) topic modeling for short text,

2) topic modeling for short text via vector embeddings.

**A. TOPIC MODELING FOR SHORT TEXT**

Several ingenious schemes create larger pseudo documents by merging short text documents, and then perform current topic modeling methods to infer the latent topics. For instance, Weng et al. aggregate all short texts generated by the same user into a training proﬁle and then apply standard LDA model. Hong et al. design two different short text aggregation schemes including the text authors and each word of corpus vocabulary. Mehrotra et al. propose different tweet pooling schemes to generate pseudo-documents in a data preprocessing step for LDA. In contrast, Zuo et al. propose a word network topic model based on the word co-occurrence network to generate pseudo-documents. Zuo et al. also propose a pseudo document-based topic model for short text by leveraging much less pseudo documents to infer the topic distributions of tremendous latent pseudo documents. Kou et al. propose a multi-feature probabilistic graphical model (MFPGM) for social network search tasks by utilizing the concept of the special region to aggregate short text into long text. Besides, Jin et al. use URLs present in short texts that reference longer documents to produce pseudo-documents. Other contextual auxiliary information in short texts has been used for aggregation such as hashtags, locations, and named entities. Kou et al. propose a social network short text semantic modeling method called STTM based on its spatial and temporal characteristics. However, the aggregation strategies mentioned above can be weaken the data sparsity problem in some extent, and may boost the quality of topic inference.

The problem lies in that auxiliary information may not be always available or topical inconsistencies between the target and auxiliary data. Recently, many models that extend word co-occurrence information have been proposed. For example, Yanetal propose a novel bi-term topic model (BTM) by directly modeling the generation of word pair co-occurring in the same document. The model aggregates all corpus bi-terms in a big pseudo-document that is used to infer the topic distribution, overcoming the sparsity problem at a document level, but the method does not consider the order of words. Lin et al. propose a dual-sparse topic model that addresses the sparsity in both the topic mixtures and the word, which apply a Spike and Slab prior to decouple the sparsity and smoothness of the document-topic and topic-word distributions. In light of this line, Quanetal propose a self-aggregation-based topic model (SATM) for short texts. The model assumes that each piece of short text snippet is sampled from a long pseudo-document unobserved in current text collection and shares the same topic proportion each other.

The topic inference and the aggregation process is conducted in a mutual reinforcement manner, such that the aggregation build upon general topical afﬁnity of texts. However, the number of parameters in SATM increases with the size of data, and setting an appropriate number of long pseudo-documents is an intractable problem.

Meanwhile, the inference process involving both text aggregation and topic sampling is time consuming. Bicalho et al. propose a general framework for topic modeling of short text by creating larger pseudo document representations from the original documents. The framework respectively uses word co-occurrence and word vector representations to generate pseudo-documents and then performs topic modeling algorithms.

**B. TOPIC MODELING FOR SHORT TEXT VIA WORD EMBEDDINGS**

Word embeddings are the collective name for a set of language modeling and feature learning techniques in natural language processing. It can simultaneously learn both syntactic and semantic information of words into continuous vectors.

Topic models have also been devised by using word embeddings representations. Most relevant work will be described as follows. Nguyen et al. propose two new latent feature topic models called LF-LDA and LF-DMM, respectively. The former integrated a latent feature model into Latent Dirichlet Allocation model, and the latter rely on a one topic-per-document Dirichlet Multinomial Mixture model. Speciﬁcally, these models replace the topic-to word Dirichlet multinomial component which generates the words from topics in each Dirichlet multinomial topic model by a two components mixture of a Dirichlet multinomial component and a latent feature component. Each word in a short text is generated from either the Dirichlet multinomial distribution or the probability estimated by using word embeddings. Similarly, Das et al. developed a variant of LDA that operates on continuous space embeddings of words by using multivariate Gaussian distributions. Li et al. propose a simple, fast, and effective topic model for short texts, named GPU-DMM, based on the Dirichlet Multinomial Mixture model.

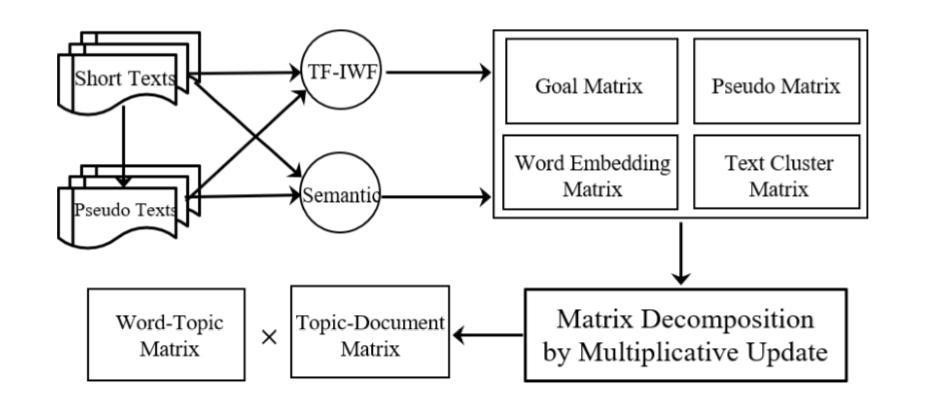
The model exploits the general word semantic relatedness knowledge provided by auxiliary word embeddings by using the generalized Polya urn (GPU) model in the topic inference of short texts. Shi et al. proposed a semantics-assisted non-negative matrix factorization (SeaNMF) model by introducing additional dependence of the keywords on their contexts via neural word embedding to discover topics for the short texts. Wang et al. integrated both word embeddings as supplementary information and an attention mechanism that segments short text documents into fragments of adjacent words receiving similar attention for short texts. Liang et al. proposed a global and local word embedding-based topic model (GLTM) for short texts, where the global word embeddings is learned from large external corpus and the local word embeddings is obtained by employing the continuous skip-gram model with negative sampling. In all, these methods either do not handle word cooccurrence features or do not obtain similar content from short texts, preventing us from further understanding the hidden semantic structure.

Different from existing works, TRNMF algorithm is proposed for topic modeling on short texts through joint word embeddings learned by pre-trained model GloVe and similar semantic information used by short text clustering algorithm GSDMM.

**Chapter 4**

**SYSTEM ARCHITECTURE**

In this section, we ﬁrst review the basic settings of NMF with a discussion of how it can be used for topic modeling. Then, given a set of short text documents, we detail how TRNMF can ﬁt into the task of modeling short texts. Next, we present the details of the TRNMF model, as shown in Figure 1.



**Figure 1: Framework of TRNMF model**

TF-IWF : ratio of Term frequency-inverse word frequency

Pseudo texts : unreal/pretended text

Semantic : meaning in the language/logic

**Chapter 5**

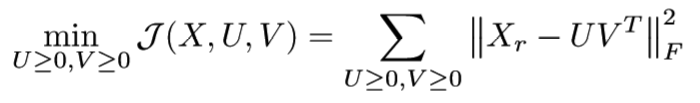
**THEORETICAL STUDY**

1. **BASE NMF FOR TOPIC MODELING**

Given a non-negative matrix Xr ∈ R+MxN , and an integer K << min(M,N), Non-negative Matrix Factorization (NMF) ﬁnds a lower-rank approximation given by

Xr ≈ UVT …………………………………………(1)

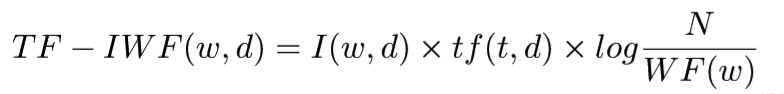
Where U ∈ R+MxN and V ∈ R+NxK are non-negative factors. NMF is typically formulated in terms of the Frobenius norm as

…………………………………………(2)

where ≥ applies to every element of the given matrix in the left-hand side.

In the topic modeling context, xi ∈ R+M×N , the i-th column of X, corresponds to the bag-of-words representation of document i with respect to m words, possibly with some preprocessing, e.g., inverse-document frequency weighting and column-wise l2-norm normalization. In general, each entry of X can be represented by tf-idf which is a weighting scheme that assigns each term in a document a weight based on its term frequency (tf) and inverse document frequency (idf). K corresponds to the number of topics. uk ∈ R+M×K , the k-th non-negative column vector of term-topic matrix U, represents the k-th topic as a weighted combination of m words. A large value indicates a close relationship of the topic to the corresponding word. vj ∈ R+N×K , the j-th column vector of topic-document matrix VT, represents document j as a weighted combination of k topics.

As we know, the purpose of tf-idf is to highlight important words and suppress minor words but the simple structure of idf cannot effectively reﬂect the importance of words and the distribution of feature words, so that they cannot achieve the right weight adjustment function, so the accuracy of tf-idf is not very high. Moreover, (1) tf-idf does not consider the distribution difference of words in short texts, and (2) term frequency contribute to a heavy weight. Therefore, in order to learn better quality of word semantic information, we choose more important words in collection of short texts to learn high quality of topic representation. Speciﬁcally, we propose a variation of the tf-idf weighting scheme term frequency-inverse word frequency namely TFIWF to measure the importance of a term in the given document collection. The actual TF-IWF formula we used is

……………………….(3)

Where,

w represents a word,

d represents a document,

I(w,d) is the indicator function that whether word w appears in document collection d or not,

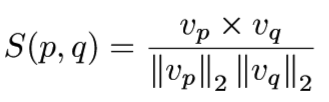
N is the sum of counting the number of times each word occurs in each document in the collection, and

WF(w) is the number of times word w occurs in the collection.

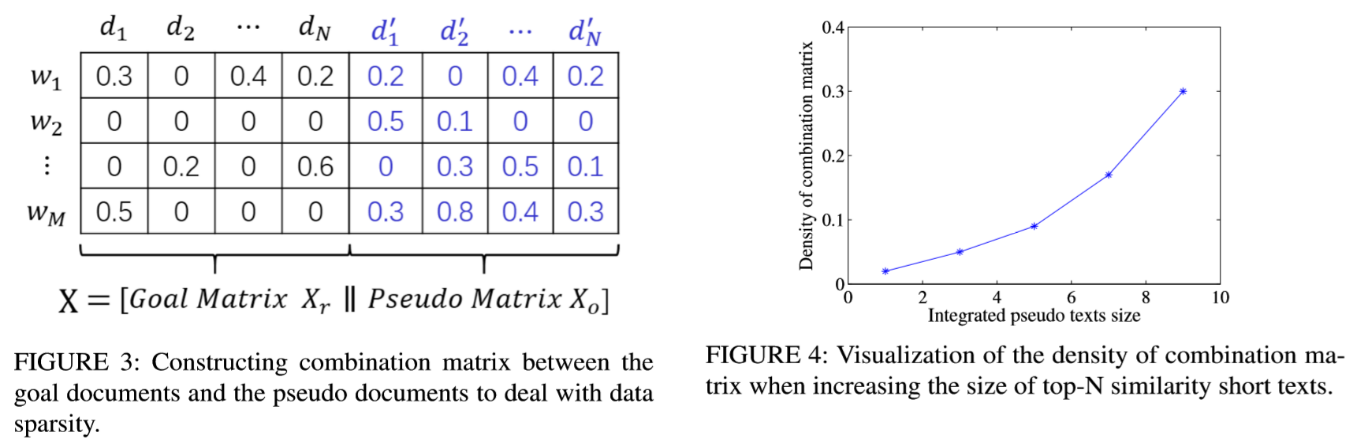
Due to the severe sparsity of the term-document matrix X, it is impossible for directly learning the optimal latent topic spaces for words and documents by relying solely on observed word co-occurrence entries. To alleviate the sparsity problem and improve the performance of topic inference, we leverage the global word embedding and document clustering information to constraint the objective function.

1. **MODELING PSEUDO-DOCUMENT MATRIX**

Topic modeling over short texts always suffers from the effect of data sparseness as only a few words appear in each text. For example, given a collection of 10,000 short texts in our case, only 0.04% entries of Xr have values. Our statistical results demonstrate that the word-document matrix Xr is over sparse. Thus, decomposing directly Xr only based on its own non-zero entries is not accurate enough for topic modeling on short texts. Furthermore, similar documents hold similar pattern both syntactic structure and semantic information, and also the similarities in observed spaces are consistent with the latent spaces. Based on the above observation and the similarity hypothesis, we aggregate short texts into long pseudo documents based on distributed vector representations for short texts. It is important that the strategy of aggregating short texts into long pseudo documents can alleviate the problem of data sparseness, and learning distributed vector representations of short texts can capture the semantics information to overcome many weaknesses of bag-of-words models. Specifically, we construct another word-document matrix Xo based on these long pseudo documents. Building long pseudo document matrix Xo has the following three steps. We ﬁrstly apply paragraph vector, an unsupervised algorithm that can model rich semantic information, to learn the ﬁxed length vector representations from variable-length pieces of short texts. After obtaining vector representation of each text, the similarity score in the distributed vector space is then calculated as the cosine similarity between short text vectors vp and vq as follows:

………………………(4)

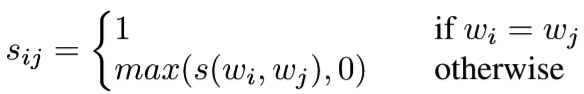
Once the similarity between short texts have been measured, we aggregate top-N (e.g., N: 2 - 5) short texts into a long pseudo document in term of similarity ranking for each text. As shown in Figure 3, pseudo word-document matrix Xo ∈ RM×N has the same structure as matrix Xr, while an entry Xo(i,j) = x’ represents the weight of a word wi that occurs in a document dj. Here, we also use the above-mentioned TF-IWF method to compute the weights between words and pseudo documents. Intrinsically, Xo is much denser than Xr, denoting the distributional similarity based on vector representations between short texts on the entire corpus. Figure 4 illustrates the impacts of the different size of windows on density. For instance, choosing top-N size to 3, then on-zero entries of Xo is about 0.5%. Therefore, decomposing Xr and Xo together can reduce the error of supplementing Xr.



1. **MODELING WORD EMBEDDINGS SEMANTIC**

MATRIX Studies reveal that short texts usually have shorter length than 100 characters. This cause the high semantic relatedness between two words less frequently co-occur in the context of short texts. On the other hand, word embeddings have been successfully applied in language models and many natural language process tasks. Conceptually, word embeddings involve a mathematical embedding from a space with one dimension per word to a continuous vector space with much lower dimension. Hence, we use word embeddings vector representation method to learn the linguistic and semantic similarity of the corresponding words (e.g., king - man + women = queen). The shallow neural network structure designed in these techniques is computationally effective on large text corpus. For instance, training skip-gram based word embeddings on a Google News corpus with 100 billion words takes less than one day on a modest computer 3. That is, general word semantic relatedness knowledge can be efﬁciently learned from a very large text corpus, in any language.

Fortunately, there are many pre-trained word embeddings learned from resources like Wikipedia, Twitter and Freebase, publication available on the Web 4. Formally, given pretrained word embeddings, a word wi is presented by a word occurrence vector (vi,1,...,vi,D), where D is the dimension of vector, and vi,j is decided by the co-occurrence of words wi and wj. We then measure the semantic relatedness between two words wi and wj by the cosine similarity between their vector representations in the latent space (i.e., the word embeddings of the two words). The semantic relatedness between the pair of words is denoted by s(wi,wj). Then a word semantic relation matrix S can be constructed, consisting of all word pairs whose semantic relatedness score is higher than 0, that is estimated with empirical probabilities as follows:

…………………………………(5)

After representing each word by the word co-occurrence vector, we then apply cosine similarity algorithm to measure the semantic correlation between any two words, resulting in the ﬁnal word correlation matrix S = [sij] ∈ RM×M. Motivated by previous works about graph clustering, we formulate this topic learning problem as ﬁnding a word-topic matrix U to minimize the following objective function:

……………………………(6)

where each column of the word-topic matrix U represents a topic by a vector of weighted words. This special formulation of non-negative matrix factorization is referred as the symmetric non-negative matrix factorization, which is suggested to be equivalent to kernel k-means clustering and spectral clustering.

1. **MODELING DOCUMENT CLUSTERING MATRIX**

Short texts refer to length-limited content documents unlike regular news articles. A length-limited sequence of word scan not provide sufﬁcient statistical factors to discover syntactic and semantic dependencies. Hence, the problem of data sparsity impedes the generation of discriminative document-topic distributions, and the resultant topics are less semantically coherent. We hypothesize that the resultant set of observed documents clustering is actually consistent with document topic representation produced by topic model with a smaller set of hidden variables. Therefore, we leverage clustering strategy to clustering semantic similarity document into the same group and resolves the problem of high dimension and data sparseness. To this date, a large number of clustering algorithms have been proposed with k-means and spectral clustering as some of the most famous ones. Among those, given that we want to tackle the problem of clustering short documents on large document collections, we focus on Dirichlet multinomial mixture clustering model that performs well on short text. This model holds that as the number of words in short documents is limited, and thus each word in the same document can be assigned to one topic. Then documents assigned to the same topic are in the same cluster. Recently, a collapsed Gibbs sampling algorithm for the Dirichlet multinomial mixture model for short text clustering (abbr. to GSDMM) was proposed.

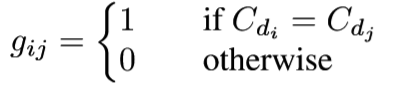
The algorithm has some good property for short text clustering problem, such as

(1) fast to converge;

(2) cope with the sparse and high dimensional problem of short texts;

(3) obtain the representative words of each cluster.

Therefore, we opt for this method as the clustering algorithm form document clustering matrix. The code used are publicly available. More speciﬁcally, we ﬁrst cluster short documents into different groups by exploiting GSDMM algorithm. Documents that are similar to one another within the same cluster and are dissimilar to documents in other clusters. We believe that if document dj is grouped to cluster Ck in observed spaces, it should be consistent with the latent space. Hence, once the clusters are created, in latent document feature space, a document should be more closed to the centroid of that cluster to which it belongs. In TRNMF model, we represent document-document clustering matrix as G ∈R2N×2N, which (i,j)-th entry deﬁned as

…………………………………….(7)

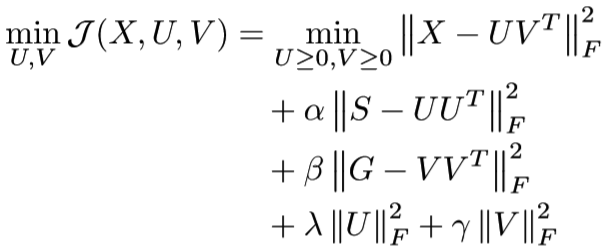
where Cdi and Cdj are the clustering labels of documents di and dj, respectively. This document clustering matrix G plays a role of making similar semantic documents to become more close each other. Each column of G corresponds to a K-dimensional topic distribution vector gk for each document by performing the vector representation algorithm of short texts. Under this scenario, assume that the given document corpus consists of K document clusters, we may arrive at the following document similarity cluster regularizer

…………………………………(8)

where V ∈ R2N×K is low rank latent factor matrix for document representations in the topic space.

1. **UNIFIED SHORT TEXT TOPIC MODEL**

Based on the above discussion, we demonstrate how to construct word co-occurrence matrix, word embeddings semantic matrix regularization and document-document similarity clustering matrix regularization, respectively. Now, we solve the optimization problem by combining J1, J2 with J:

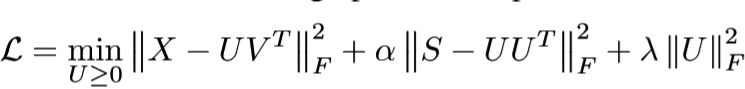
………………………………….(9)

where X ∈ RM×2N and V ∈ R2N×K; α ≥ 0 and β ≥ 0 are the parameters controlling word co-occurrence regularization and message clustering regularization on Ui and Vj, respectively.

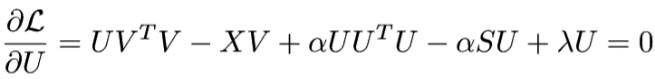
1. **MODEL OPTIMIZATION**

Easily observed in Eq. (9), where if V is ﬁxed, J is a convex optimization problem with respect to U and if U is ﬁxed, J is a convex optimization problem with respect to V. However, when both and are not ﬁxed, J is not convex. Therefore, the global optimal solutions of the objective function are difﬁcult to formalize.

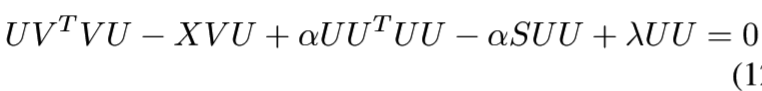
Nevertheless, local optimal solutions can be obtained by multiplicative update method. Solving the Word-Topic Matrix U. Hold document-topic matrix V = [v1,··· ,vN] ﬁxed, the updating rule of U equivalent to the following optimization problem:

………………………………..(10)

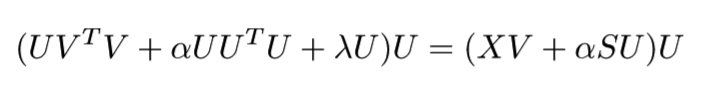
And then we take the gradient of (10) with respect to U and setting it to zero:

……………………………..(11)

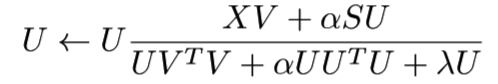
By multiplying (11) by U, (11) can be written as

2)

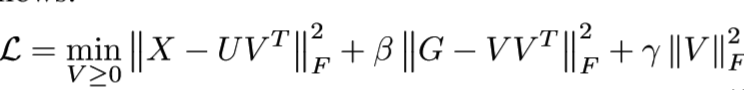
Where each element in X, S is non-negative and α, λ are non-negative as well, moreover, the initial values U and V are both non-negative; consequently, UVTV , XV , UUTU and SU are non-negative. Thus (12) can be written as follows

…………………..(13)

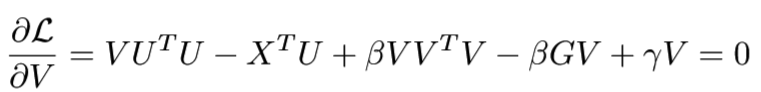
So the updating rules of U is deﬁned as follows:

………………………(14)

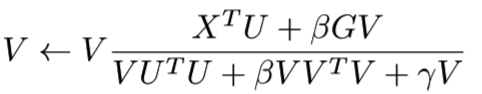
Solving the Document-Topic Matrix V. After learning matrix U, we then solve the document-topic matrix V . Similarly, the objective function based on matrix V is shown as follows:

……………………….(15)

And the gradient of (15) with respect to V and setting it to zero:

………………………..(16)

And the updating rules of V is deﬁned as follows:

……………………………..(17)

**Chapter 6**

**ALGORITHM**

**Require**: **X**, **U**, **V**, parameters α, β, γ, λ, topic number K

**Ensure**: Latent feature matrix **U** and **V**

1: Randomly initialize **U** and **V**;

2: **for** i = 1 : M **do**

3: **for** j = 1 : M **do**

4: calculate S(i,j) with (5);

5: **end for**

6: **end for**

7: **for** i = 1 : 2N **do**

8: **for** j = 1 : 2N **do**

9: calculate G(i,j) with GSDMM model;

10: **end for**

11: **end for**

12: **for** k = 1 : K **do**

13: **Uk** ← *Update* **U**(**X**,**Vk−1**);

14: **Vk** ← *Update* **V**(**X**,**Uk**);

15: **end for**

16: **return** **U**, **V**;

**Time Complexity:**

We assume that the number of iterations is T. The main computational cost of the proposed method lies in solving the updating rules of U and V on the latent spaces. Since X is very sparse, XV and UVTV take O (MNK) and O (max (M, N) K2) time, respectively, where the number of topics K is much smaller than the number of words M in vocabulary. We conduct the similar computation on V observed similar phenomena. Therefore, the overall complexity for our proposed method is T × O (MNK). It is easy to see that TRNMF model can be applied in large-scale real system to deal with incrementally increasing data. TRNMF model can be applied in the real system to deal with incrementally increasing data. It has been proved both storage and computational efﬁcient by solving the smoothly evolved factorized matrices.

**Chapter 7**

**EXPERIMENT**

In this section, an extensive quantitative comparison of TRNMF model approach against other state-of-the-art methods is presented. Afterwards, qualitative results containing high-quality latent topics identiﬁed by TRNMF methods are demonstrated, which would be otherwise difﬁcult to discover using other existing methods, from several real-world datasets.

1. **DATASETS**

We use several the corpus of short text documents as input for the topic modeling algorithms. These datasets are described as follows:

**TMNews.**

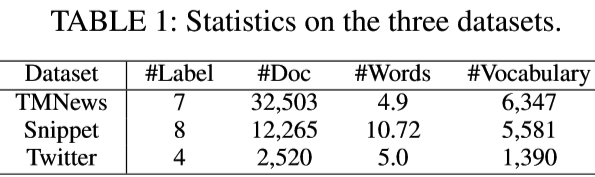
This dataset contains about 32,600 English news articles extracted from RSS feeds of three popular newspaper websites (nyt.com, usatoday.com, reuters.com). Categories include sport, business, U.S., health, sci & tech, world and entertainment. We retain news descriptions since they are typical short texts.

**Snippet.**

A collection of web search snippets, which are summaries of documents presented as results of a query by a search engine. The queries used are related to 8 different domains: business, computers, arts, education, engineering, health, politics, sports.

**Twitter.**

This corpus consists of 5,513 hand-classiﬁed tweets. These tweets were classiﬁed with respect to one of 4 different topics: Apple, Google, Microsoft, Twitter. All datasets were preprocessed before the expansion step by making all the text lower case, removing non-alphabetic characters and stop words in a standard list. We also removed 425 words shorter than 3 characters, and words appearing less than 10 times in Snippet and under 5 times in the TMNews and Twitter datasets. Table 1 lists statistics of the data set used in this paper. Note that we have few words per document for all datasets (column #Words).



1. **BASELINE METHODS**

We compare TRNMF model against the following various state-of-the-art topic models speciﬁc to short texts.

**Biterm Topic Model (BTM).**

BTM explicitly learns the word co-occurrence patterns in the whole corpus for learning latent topics to solve the problem of sparse word co-occurrence patterns. However, a biterm for this model is an unordered word pair co-occurred in a short context.

**Word Network Topic Model (WNTM).**

WNTM constructs a word co-occurrence network to discover latent word group and learns distribution over topics for words rather than topics for documents. The model successfully enhances the semantic density of data space, and makes topic inference less sensitive to the document length or heterogeneity of the topic distribution.

**Latent Feature model with DMM (LF-DMM).**

LFDMM integrates latent feature word representations into Dirichlet Multinomial Mixture by replacing the topic-to-word Dirichlet multinomial component with a mixture of two components: a Dirichlet multinomial component and a word embedding component.

**Generalized Polya Urn with DMM(GPU-DMM).**

GPUDMM also extends the Dirichlet Multinomial Mixture model by incorporating the pre-trained word embeddings learned from large text corpus through the generalized Polya urn model in topic inferences. This model can promote the semantically relevant words together, even if they share very limited or no co-occurrences in the current corpus.

**Pseudo-document-based Topic Model (PTM).**

PTM assumes huge volume of short texts are generated from much less yet regular-sized latent documents, and introduces the concept of pseudo document to implicitly aggregate short texts against data sparsity without using auxiliary contextual information.

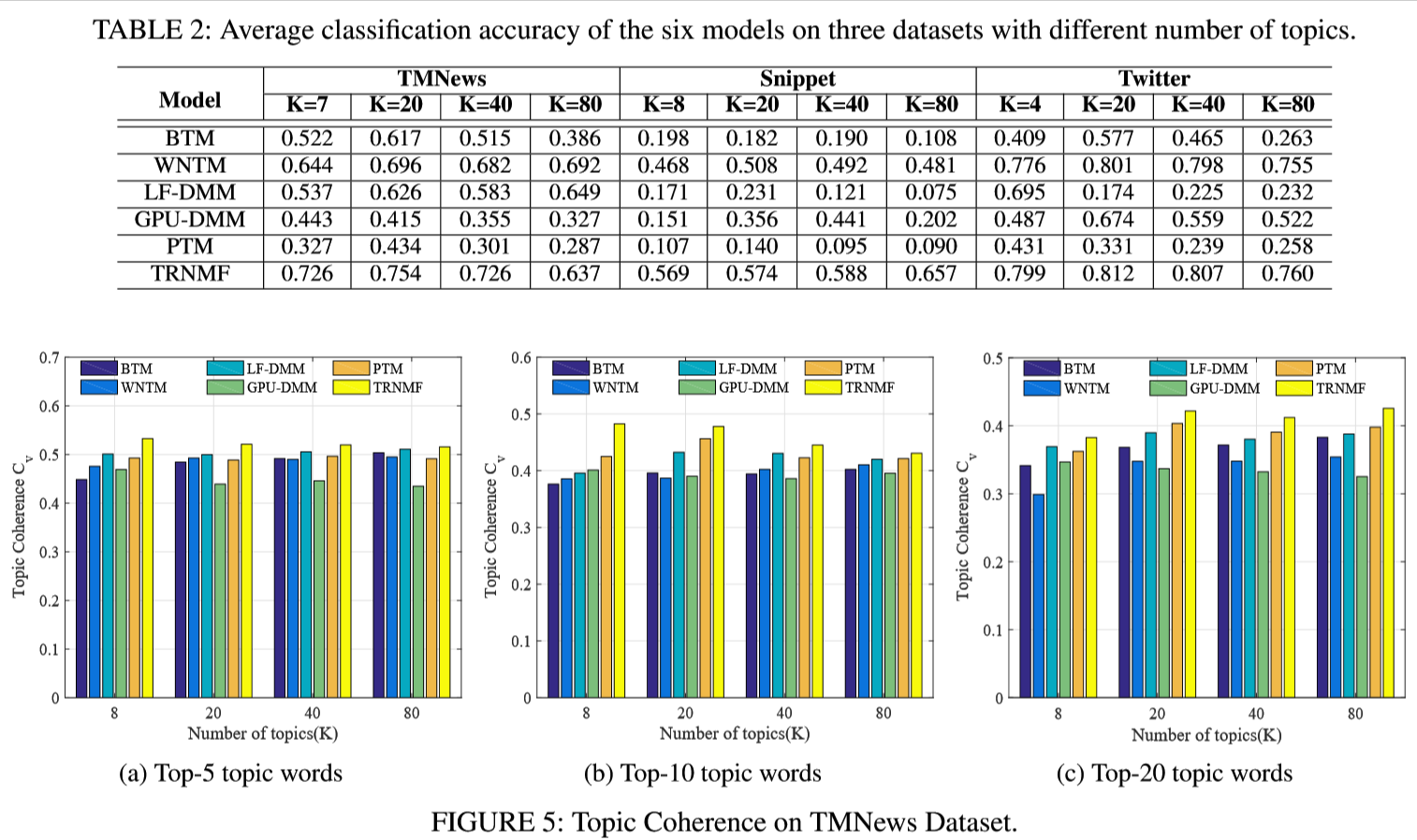
Regarding these topic models, BTM, WNTM, and GPUDMM share two main hyper-parameters: α = 50/K, β = 0.01. Speciﬁcally, we set to 10 with length of the sliding window for WNTM.

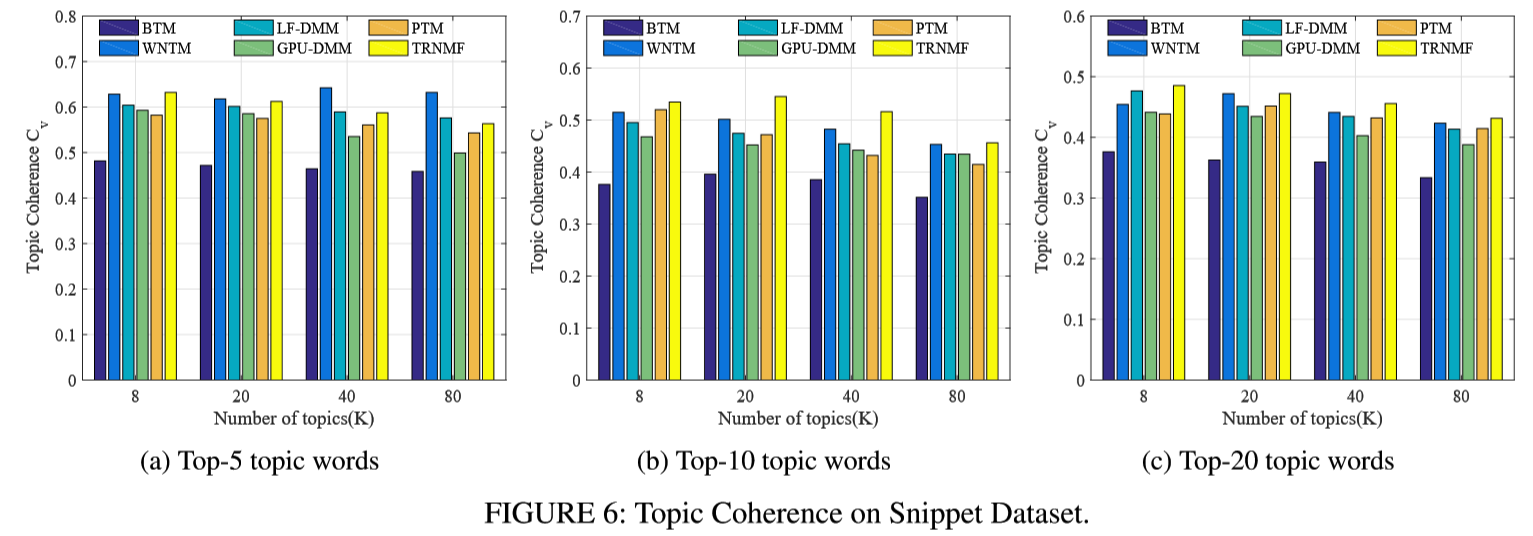
For LF-DMM, we use the authors recommended settings with λ = 0.6, α = 0.1, β = 0.01.

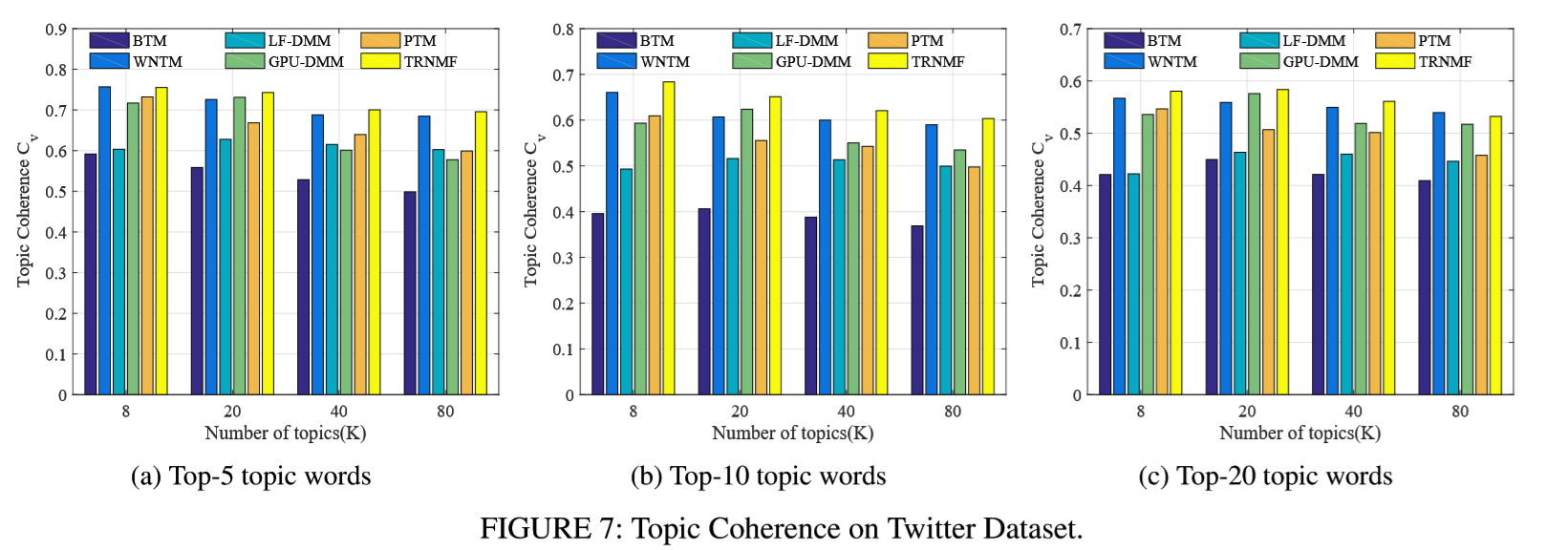
For PTM, we set α = 0.1, λ = 0.1 and β = 0.01.

For model TRNMF, we use cross-validation method to the optimize parameters. Through extensive evaluations, we set α = 0.1, β = 0.1, λ = 0.01 and γ = 0.01 in the following comparison experiments.

In all experiments, the number of latent topics is set to {20,40,80} and the number of ground truth categories, and each model’s Gibbs sampling is run for 1,000 iterations. The results of document classiﬁcation here are reported base on the average over 5 rounds.







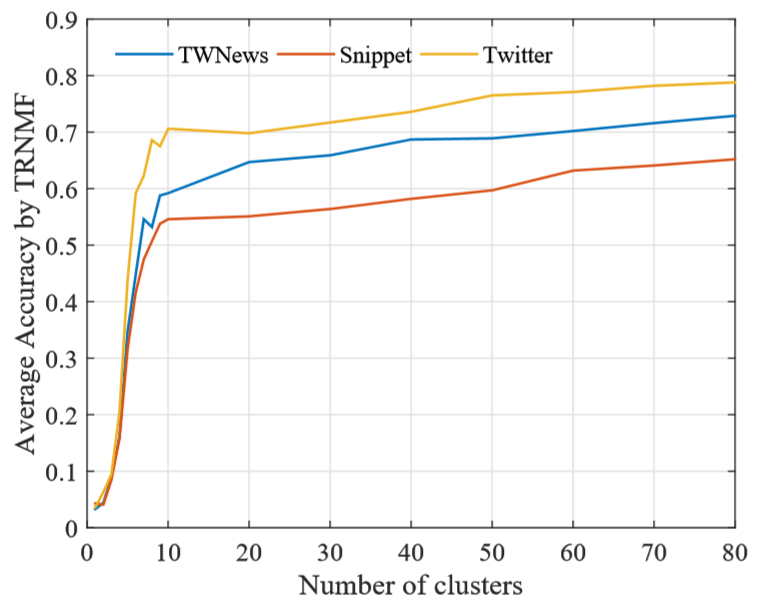


FIGURE 8: Average accuracy achieved by TRNMF with different values of K on the three datasets.

**Chapter 6**

**ADVANTAGES AND CHALLENGES**

**ADVANTAGES**

We can see from the results of the experiment that

1. TRNMF method consistently outperform all other baseline methods on all datasets;
2. TRNMF can discover more readable topics;
3. TRNMF model can better discover the hidden semantic structure of short text collections.
4. Word-weighting schema is much more effective for capturing semantic meaningful features from the input of short texts.
5. The stable number of clusters found by TRNMF in FIGURE 8, is near the ground true number of groups which indicates TRNMF can well discover the number of clusters with K increasing.
6. Besides, we can also observe that TRNMF model has a good convergence property when increasing the number of clusters.
7. The superior performance of TRNMF method as compared to other is because it employees global co-occurrence of words and it uses clustering of documents, which guarantees the quality of topics.
8. It deals with data sparsity.

**DISADVANTAGES**

(2) On TWNews and Snippet datasets, TRNMF does not performs best whereas, PTM performs the best among baseline methods.

**Chapter 7**

**CONCLUSION**

In this paper, we propose a new method based on non-negative matrix factorization model for latent topic inference, aiming at short texts under data sparseness. By leveraging global word-word co-occurrence knowledge to help extract better topics over short texts, TRNMF gains advantages in learning topic distributions using auxiliary contextual information. On the other hand, the proposed model alleviate the sparsity problem and improve the performance of topic modeling by exploiting document clustering relationship. The experimental results show that TRNMF model outperforms existing conventional alternatives in terms of effectiveness and efﬁciency.

**FUTURESCOPE**

The TRNMF model can also be improved by using the graph-based word embedding technique in future.

**REFERENCES**

1. Feng Yi, Bo Jiang, And Jianjun Wu, “Topic Modeling for Short Texts via Word Embedding and Document Correlation”. IEEE,2019
2. T. Shi, K. Kang, J. Choo, and C. K. Reddy, “Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations,” in Proceedings of the 2018 World Wide Web Conference. International World Wide Web Conferences Steering Committee, 2018, pp. 1105–1114.
3. P. Pinoli, D. Chicco, and M. Masseroli, “Latent dirichlet allocation based on gibbs sampling for gene function prediction,” in Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on. IEEE, 2014, pp. 1–8.
4. Y. Wang, J. Liu, J. Qu, Y. Huang, J. Chen, and X. Feng, “Hashtag graph-based topic model for tweet mining,” in Data Mining (ICDM), 2014 IEEE International Conference on. IEEE, 2014, pp. 1025–1030.
5. L. Jiang, H. Lu, M. Xu, and C. Wang, “Biterm pseudo document topic model for short text,” in Tools with Artiﬁcial Intelligence (ICTAI), 2016 IEEE 28th International Conference on. IEEE, 2016, pp. 865–872.
6. J. Wood, P. Tan, W. Wang, and C. Arnold, “Source-lda: Enhancing probabilistic topic models using prior knowledge sources,” in Data Engineering (ICDE), 2017 IEEE 33rd International Conference on. IEEE, 2017, pp. 411–422.
7. J. Wang, L. Chen, L. Qin, and X. Wu, “Astm: An attentional segmentation-based topic model for short texts,” in 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018, pp. 577–586.
8. 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.
9. X. Quan, C. Kit, Y. Ge, and S. J. Pan, “Short and sparse text topic modeling via self-aggregation.” in IJCAI, 2015, pp. 2270–2276.
10. Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, and H. Xiong, “Topic modeling of short texts: A pseudo-document view,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 2105–2114.
11. O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, “Transferring topical knowledge from auxiliary long texts for short text clustering,” in Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011, pp. 775–784.
12. J. Tang, M.Zhang, and Q. Mei, “One the mein all views: modeling consensus topic sin multiple contexts,” in Proceedings of the 19th ACMSIGKDD international conference on Knowledge discovery and datamining. ACM, 2013, pp. 5–13.
13. 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. ACM, 2013, pp. 1445–1456.
14. 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. Association for Computational Linguistics,2010, pp. 100–108.
15. J. H. Lau, D. Newman, and T. Baldwin, “Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality.” in EACL, 2014, pp. 530–539.
16. M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in Proceedings of the eighth ACM international conference on Web search and data mining. ACM, 2015, pp. 399–408.
17. Bouma, “Normalized (pointwise) mutual information in collocation extraction,” in Proceedings of the Biennial GSCL Conference, vol. 156, 2009.
18. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of machine Learning research, vol.3, no. Jan, pp.993–1022,2003.
19. And many more.