Multimedia Tools and Applications (2023) 82:26411–26433
https://doi.org/10.1007/s11042-023-14352-x

Short text topic modelling using local and global word-context semantic correlation

Supriya Kinariwala[1] · Sachin Deshmukh[2]

Received: 21 October 2021 / Revised: 21 February 2022 / Accepted: 2 January 2023 /
Published online: 2 February 2023

Abstract

Nowadays, people use short text to portray their opinions on platforms of social media such as Twitter, Facebook, and YouTube, as well as on e-commerce websites such as Amazon and Flipkart to share their commercial purchasing experiences. Every day, billions of short texts are created worldwide in tweets, tags, keywords, search queries etc. However, this short text possesses inadequate contextual information, which can be ambiguous, sparse, noisy, remains a major challenge. State-of-the-art strategies of topic modeling such as Latent Dirichlet Allocation and Probabilistic Latent Semantic Analysis are not suitable as it contains a limited number of words in a single document. This work proposes a new model named G_SeaNMF (Gensim_SeaNMF) to improve the word-context semantic relationship by using local and global word embedding techniques. Word embeddings learned from a large corpus provide general semantic and syntactic information about words; it can guide topic modeling for short text collections as supporting information for sparse co-occurrence patterns. In the proposed model, SeaNMF (Semantics-assisted Non-negative Matrix Factorization) is incorporated with word2vec model of Gensim library to strengthen the word's semantic relationship. In this article, a short text topic modeling techniques based on DMM (Dirichlet Multinomial Mixture), self-aggregation and global word co-occurrence were explored. These are evaluated using different measures to gauge cluster coherence on real-world datasets such as Search Snippet, Biomedicine, Pascal Flickr, Tweet and TagMyNews. Empirical evaluation shows that a combination of local and global word embedding provides more appropriate words under each topic with improved outcomes.

Keywords Short text · Text mining · Topic modelling · Word embedding · Non-negative matrix factorization · Global corpus

✉ Supriya Kinariwala
sakinariwala@gmail.com

[1] Maharashtra Institute of Technology, Maharashtra, Aurangabad, India

[2] Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India

1 Introduction

In today's digital world, the global increase in the use of social media has resulted in a vast volume of unstructured data being spread around the WWW (World Wide Web). People rely on the internet daily for social and commercial activities. Recently, the Covid-19 pandemic has increased people's reliance on the internet. Information on the internet is being stored in dual forms, such as long-text and short-text. The text representing the details about the journal and news articles is termed as long-text, whereas the people's post on the websites such as Twitter, Instagram, and Facebook are known as short-text. Over the past years, the rapid accessibility of the internet has created the storage of massive digital information. The platforms and applications of social networks generate a vast amount of textual data in product reviews, tweets, comments, news headlines, blogs, and short-text communications [24]. Moreover, the relevant information extraction process has become challenging and requires new tools to effectively summarize and extract internet data.

With the massive amount of textual information, it is difficult for the users to obtain the needed information or the topic that is being discussed from the information content. The process of mining the text information is Text Analytics (TA), which involves the transformation of unstructured textual data into the structured form [30]. This analysis can help the users to understand the user opinions, evaluate the reviews of the product, and user feedback over the social media networking sites. The generation of short texts in the form of image tags, tweets, questions, headlines, ad keywords, and search queries is common every day. Finding significant knowledge from these texts has earned huge attention from academia and industry. Some limitations associated with short texts are ambiguity, noise, sparse, and limited contextual data. Thus, automatic learning procedures are required to tackle the problem of discovering the topic from short texts.

One of the effective techniques for TA is topic modelling which has been useful in various fields of data mining (DM), NLP (Natural Language Processing), and ML (Machine Learning). NLP integrates the ability of computer science, computational linguistics and AI (Artificial Intelligence), which enable machines to recognize, examine and construct the original meaning of the text on the internet [23]. Topic Modelling (TM) aims to discover the topics, keywords, tags, categories, semantics from the massive text data. NLP and ML algorithms are used to reduce information from text corpora. The term TM is expressed as a procedure of discovering topics for the word groups or document collection which represents the best information.

The traditional TM approaches are extensively used to automatically uncover the hidden thematic information from the document representing rich contents. The two forms of topic modelling are LDA (Latent Dirichlet Allocation), PLSA (Probabilistic Latent Semantic Analysis) and NMF (Non-negative Matrix Factorization) [4, 8, 20]. Classic algorithms such as LDA and PLSA act on word co-occurrence at the document level; they analyse lengthy text documents. NMF is a linear algebraic model used for dimensionality reduction. When the underlying factors are not negative, this strategy is appropriate, so it is utilized for topic modeling. It produces good clustering results for high-dimensional data [12, 13]. However, these traditional models achieved excessive success for considering the regular-sized documents and won't focus on collecting short texts. The short text analysis can be applied to various fields, including information retrieval, text classification, text clustering, recommender systems, user interest mining, and futuristic business perspectives [2, 19]. The short texts include only a few meaningful keywords, and it is difficult to capture the word co-occurrence data. In the real-world, the ubiquitous form of short text has emerged as a significant data source, and numerous techniques have been introduced to discover the latest topics for each document. Therefore, a new topic modelling strategy is introduced for short texts' topics.

1.1 Motivation

Today, the need and demand for the internet are widespread and has offered better commercial and shopping practices to customers around the globe. The vast amount of information and knowledge creates confusion and increases the overall searching time. Further, it is significant to extract the hidden and useful information from the online sources stored as text and written in natural language in social networking cites like Facebook, Twitter and LinkedIn. Text mining (TM) based topic modelling have been introduced to extract the data from the large source as it is a hard and inefficient operation. In order to avoid these challenges in manual text mining operation, various automatic TM approaches have been developed to extract the text or topics automatically from massive online sources. These approaches offer reliable outcomes in enormous text analysis domains. In recent years, machine learning and natural language processing algorithms like topic modelling have been commonly used to analyse the massive number of textual social media data available online. Hence, in this research work, an effective unsupervised clustering-based topic modelling is proposed, extracting more meaningful topics.

1.2 Contributions

The major contributions of the short-text topic modelling are:

- To propose an effective short-text topic model named G_SeaNMF by combining the word2vec model of Gensim library trained on Google News Dataset and SeaNMF model, which discovers the accurate topics from short-text documents.
- To incorporate both internal and external corpora to discover semantic relationships between words, resulting in more related terms under a single topic.
- To process qualitative SA (Semantic Analysis) that demonstrates the system's overall effectiveness by finding the meaningful topics and assigning appropriate labels to each topic using the document-term matrix.
- To perform sufficient quantitative assessments on different short text real-world datasets to illustrate the enhanced performance of proposed G_SeaNMF over existing models.

The remaining paper is organized as follows: Section 2 presents a literature survey of the recent relevant techniques related to short-text topic modelling, Section 3 covers the proposed G_SeaNMF methodology with principles and mathematical formulations, Section 4 includes the datasets, evaluation metrics, comparative analysis, and discussions. Finally, Section 5 presents the conclusion of the paper.

2 Related works

2.1 Some of the various short-text topic models are listed below

Rashid J et al. [28] developed a new FTM (Fuzzy Topic Modelling) to alleviate the data sparsity issue for short text data. This model is presented utilizing a fuzzy point of view to extract and discover the latent topics/themes from a short text. This approach computes the global and local term frequencies and eliminates the negative influence of high dimensionality on the global term weighting to find more relevant topics from the documents. The accuracy of FTM classification was computed with the SVM (Support Vector Machine) classifier on the dataset's questions and snippets. With SVM, for different number of topics such as 50, 75, 100, 125, 200 the accuracy obtained with questions were (0.73), (0.74), (0.70), (0.68), (0.78) whereas with snippets (0.95), (0.94), (0.91), (0.89), (0.87) respectively. The main drawback of FTM was random value setting for hyper-parameters and common topics. Yi et al. [35] introduced a regularized non-negative matrix factorization topic model (TRNMF) approach to solve the data sparsity issue. This major problem for short text has been done by exploiting the document correlation and word embedding, and it controls the pre-trained distributional vector representation of words. The TRNMF includes both sentence similarity regularization and word co-occurrence for short text in topic modelling. The performance computation was processed using topic coherence and classification. The datasets used for testing were snippet, Twitter and TMNews datasets. With topic coherence, a set of top words were displayed for each topic. The proposed TRNMF obtained higher classification results with accuracy 0.728, 0.798, and 0.523 on TMNews, Twitter, and snippet. The major drawback identified was difficulty understanding the semantic contents of short texts.

Yang et al. [34] developed a new topic, representative term discovery (TRTD) approach short text clustering to find a typical topic term by utilizing the intimacy and consequence of terms. In entire text documents, this method computes the intimacy of the typical topic terms by their symbiotic co-occurrence and estimates the consequence by global term events. The performance of the short text clustering was evaluated with the NMI, ARI (Adjustable Rank Index) and AMI (Adjusted Mutual Information). The dual datasets used for testing were tweet and title datasets. The proposed TRTD approach was compared with several baseline models and proved a better algorithm. The performance was computed with clustering accuracy using the metrics and gained 0.810, 0.842, 0.771 on the tweet dataset and 0.804, 0.828, 0.781, respectively. The challenge noticed was short texts with similar topics represent the repeated terms that affect the clustering accuracy. Liu et al. [17] developed a framework named CME-DMM (Collaboratively Modelling and Embedding-Dirichlet Multinomial mixture), which incorporates collaboratively Modelling and Embedding, and DMM to extract coherent hidden topics from a short text. This method combines word, and topic embedding's through the attention mechanism to enhance the latent topic quality. An embedding factor called SGD (Stochastic Gradient Descent) was used to optimize the words and topics. The datasets used for evaluation were news data set and Chinese short text data. The topic coherence metric signifies the performance of the latent topics using the generated PMI (Pointwise Mutual Information) score. The text classification was processed with SVM in Micro-P, Micro-R and Micro-F1 metrics. The performance obtained for k = 10 topics on CME-DMM were 0.884, 0.835 and 0.844, respectively. The drawback identified was noise issues and sparsity confronting the short texts.

Gao et al. [7] presented a new CRFTM (Conditional Random Field regularized Topic Model) to create a generalized solution and solve the sparsity issue by combining the short texts into documents. Moreover, it inspires the semantically related words to share a similar topic assignment. The performance of topic coherence was computed with the PMI index, and the quality of topics was evaluated with the accuracy measure. The datasets used for processing were Stack overflow and News dataset. By varying the number of topics as K = 40, 60, 80, the accuracy obtained using CRFTM on news dataset as 75.99%, 76.31%, 77.40% and with stack overflow dataset as 71.5%, 75.71% and 77.03%, respectively. Roccetti et al. [29] presented a model to analyse the importance of topic modelling with social media data. The major aim was concerning the patient data originating from various social media sources such as Twitter, Insta, Facebook etc. The data has been collected from Facebook with the time frame of 2011 October to 2015 August to examine the perspective of each patient on the medical prescription. The model allows rapid collection of huge amounts of data that can be analysed easily to gain awareness of medical therapy. Asmussen et al. [3] presented a smart literature review about topic modelling. An analysis was framed on the use of topic modelling with the 3 step procedure such as pre-processing the data, topic modelling, and post-processing. The pre-processing procedure aims to get the data and process it ready to run, whereas the LDA approach was used to execute topic modelling. At last, the post-processing was done to transform the LDA outcome into an exploratory review which can be further utilized to recognize the papers that have been used for reviewing the literature.

An enhanced approach called GLTM (Global and local word embedding based model) was proposed by Liang et al. [15] in the case of short texts. In order to attain local word embeddings, SGNS (Skip Gram model with negative sampling) was employed continuously, and the training of global word embeddings from huge external corpus were done in GLTM. GLTM can distill the semantically related information between words by utilising both local and global word embeddings. A new generation process was established for the collections of a short text that integrates spike and slab priors to analyse the topic number for every short document based upon the content. A GPU model was employed as the sampler in the inference process that forces the information between words. The major drawback in this research was inefficient robustness and degraded performance due to the maximization of topic number. Singhal et al. [32] presented an architecture for conducting scientific analysis of academic publications critical to observing research trends and determination of potential innovations. The proposed framework acquired and combined different natural language processing approaches, including word embedding and topic modelling. This research contributed two novel scientific publication embedding, including PUB-G and PUB-W. These processes could analyse semantic meanings of general and the specified words regarding domain in different fields of research. Hence topic modelling was utilized to determine clusters of research topics within these forms of huge research fields. The PUB-G and PUB-W model had attained superior topic coherence results when compared to the other baseline embeddings. The main advantage of this research was many word embedding processes that train over general text articles can capture the features related to domain specified texts available in scientific publications. Through this research, the best coherence score cannot be attained, and the PUB-G and PUB-W models promote poor generalization in case of limited data points.

The utilization of word embeddings for patent retrieval was explored by Hofstatter et al. [25], especially over the approaches on the basis of distributional semantics. Because of the limited efficiency of semantic models and the degraded effectiveness of specified word embeddings, the inherent constraints over the window context were not efficient in accessing the patent domain. To overcome the limitation of full complexity capture for patent domain, the local and global contexts for embedded learning were drawn altogether. The integration was done in this research on the basis of two ways: adoption of vectors using Skip-gram model through global retrofitting and filtering of similarities in the word through global context. The limitation faced here was that investigating retrieval domains with the same characteristics was difficult. A better tuning stage incorporated with the original corpus was proposed by Murakami et al. [9] by pre-trained word embedding. The neural topic models were processed for generating the semantically coherent and corpus specific topics. An enhanced study with eight neural topic models has been accomplished to check the efficiency of extra fine-tuning and pre-trained word embedding in establishing interpretable topics through simulation experiments with different benchmark datasets. The gathered topics are estimated through various metrics like topic coherence and topic diversity. The model performance with respect to classification and clustering tasks were also described.

The drawback faced here was the performance of downstream tasks, including classification in case of long texts were degraded. Qiang et al. [21] contributed to incorporating the correlation knowledge of external words into short texts to enhance the topic modelling coherence. On the basis of recent outcomes in word embeddings from a huge corpus, a novel methodology called ETM (Embedding based Topic model) was introduced to observe the latent topics from short texts. By aggregating short texts into long pseudo texts, the problem of restricted co-occurrence information of word can be solved. The k-means was implemented through a new metric called WMD (Word Mover's Distance) to evaluate the distance between two short texts. The correlated words were rendered by utilising the Markov Random Field regularized model to attain a better chance to be held in a similar topic. The model's effectiveness proposed in this research was validated by conducting experiments over real world datasets. The coherence of topic modelling can be enhanced better, but the overall effectiveness of the model is less.

2.2 Problem statement

In today's era, the posting of short text has popularized in people's everyday life due to the proliferating usage of social media. Short texts are being generated in an unlimited way in the form of tweets, tags, social network posts, phone messages, keywords, messenger conversations etc. Some of the most challenging issues associated with short texts are unique characteristics with many limited contents that make the text handling process difficult. In content analysis and text mining, this short text analysis has gained a significant advantage. However, the challenge is critical, focusing on extracting the exact topics from the large-scale short text documents.

Moreover, lots of documents are distributed on the internet from different sources. Internet websites contain different topics and data, but there is a lot of similarity between topics, contents and the entire quality of sources, which causes data repetition and provides the user with similar data. One more problem is data sparsity and ambiguity, as the length of the short text is limited, which provides poor outcomes to end-users. The baseline topic modelling approaches failed to obtain word co-occurrence patterns in topics due to the sparsity problem in short texts, such as social media like Twitter, text over the web, and news headlines. Therefore, this work proposes an efficient G_SeaNMF model based on the local and global word-context semantic correlation to solve these problems in short text topic modelling.

3 Proposed methodology

Topic modelling is a kind of unsupervised machine learning algorithm that helps discover latent topics in text corpora and annotate texts based on their subjects. Topic modelling is the technique for discovering the group of topics (words) from a massive document collection that illustrates the information better. It can also be defined as a text-mining procedure that gains recurrent word patterns in textual information. The steps involved in proposed short-text topic modelling are: Pre-processing G_SeaNMF algorithm-based topic modelling. The architecture of proposed G_SeaNMF model is shown in Fig. 1.

3.1 Pre-processing

This is the first step that is processed in the text to make the text into a suitable form, which can be predictable and analyzable for performing topic modelling. The initial stage is to assemble unstructured text data from various sources, such as tweets, product reviews, and news headlines. Pre-processing is also termed as the cleaning procedure that prepares the text for further processing. The text data collected from online sources usually includes noise, incomplete sentences, advertisements, and non-informative parts (scripts, HTML tags, numbers, symbols etc.). Using these following terms in the text creates a high dimensionality problem that affects the system performance. Pre-processing aims to minimize the noise present in the text and speed up the short-text topic modelling. The pre-processing techniques that are being assessed here to clean up the collected data are:

- *Lowercase conversion:* Transforming all the text data into lower case. This is a simpler procedure that treats all the words similarly. Lowercasing helps to maintain stability during text mining and NLP tasks.
- *Noise Removal:* Removing noise in a text can be defined as removing unwanted texts or digits that may affect the analysis of the text. Removal of noise is considered an important step in the data cleansing stage of text analysis. Noise removal or minimization of redundant information is the main aim of pre-processing.

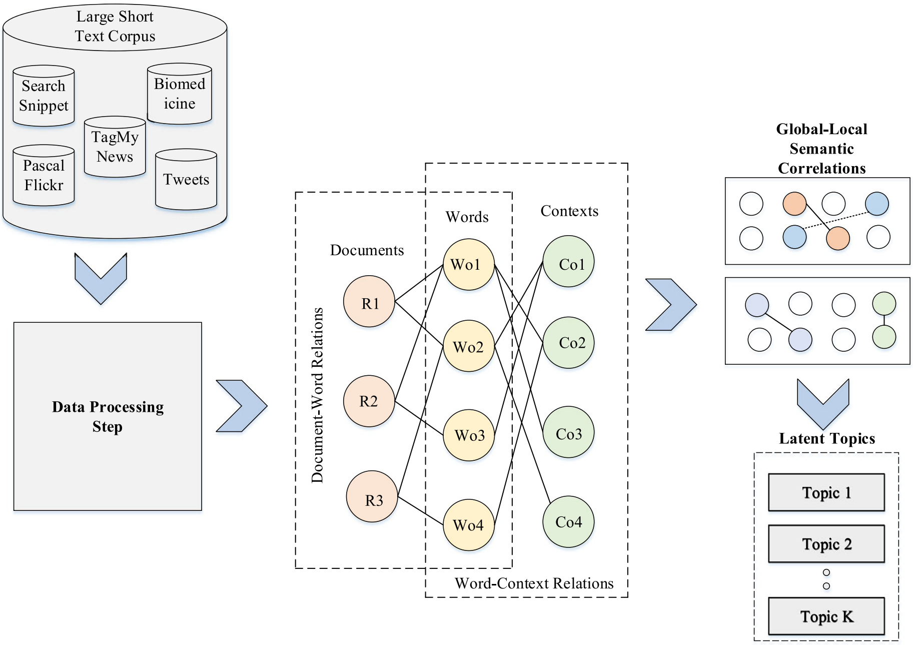


Fig. 1 Architecture of proposed G_SeaNMF model

- *Stop word removal:* In NLP, the useless data is called stop words. The most used stop words are 'the', 'a', 'is', 'an', 'are', 'in' etc. Removing these stop words gives more attention to the important data. Also, the elimination of non-essential terms increases focuses on important words.
- *Stemming:* The method of minimizing a word into a word stem is called stemming. It returns the root form of word, which aids in vocabulary standardisation.
- *Lemmatization:* The term lemmatization transforms the word into a relevant base word is called a lemma. It reduces the word to its base form by utilising a dictionary such as Wordnet. The different inflected word forms are grouped and analyzed as a single term that brings the context to words.
- *Normalization:* Transforms highly unstructured language to its canonic form. For example, the phrases 'gooood' and 'gud' are converted to good. Text normalization minimizes the amount of information and brings it into a standard form.
- *Text Enrichment:* Use POS (Parts of Speech) tagging to enhance information about the words. Enhancing the textual information with POS tagging helps recognize the grammatical meaning by verifying its contents.

3.2 G_SeaNMF algorithm based topic modelling

In this section, the proposed G_SeaNMF model, and BCD (Block Co-ordinate Decent) algorithm to measure short documents and terms latent representations are mentioned. The proposed G-SeaNMF technique leverages both internal and external corpora for weighting word-context matrix. The input is a short text corpus that has been pre-processed. Vocabulary is created, which is then used to construct a document–term matrix based on the terms contained in the given document. The matrix, which is represented using rows (documents) and columns (words), is called document–term matrix. The major challenge of the proposed model is to use the term correlations obtained from the pre-trained word representations on the large external corpus (Google News). This Google News dataset includes 3 M (Million) words in English that are embedded into 300d (dimensional) latent space by executing the W2V (word2vec) model on Google News corpus having 3B (Billion) running words [6].

Consider a given corpus that includes N number of distinct terms/words/keywords and M number of documents present in vocabulary Vo. The representation of term_document matrix is $T \in \Re_{+}^{N \times M}$ which is further approximated using dual lower-rank matrices such as $Wo \in \Re_{+}^{N \times K}$ and $R \in \Re_{+}^{M \times K}$, i.e., $T \approx WoR^{T}$, $K < < min\,(N, M)$ signifies the number of topics (latent factors). The latent matrix is represented as $Wo \in \Re_{+}^{N \times B}$. By applying the non-negative (NN) constraints on context and word vectors, it is symbolized as $c\overrightarrow{o} \in \Re_{+}^{B}$, $w\overrightarrow{o} \in \Re_{+}^{B}$. For a given keyword denoted as $wo_{i} \in \mathrm{X}$, then set $Wo_{(i,:)} = w\overrightarrow{o}_{i}$. The representation of word_context or semantic correlation matrix is described as,

$$Y \approx \boldsymbol{WoWo}_{c}^{T} \tag{1}$$

$$Y_{ij} = \left[log \left(\frac{\#(wo_{i}, co_{j})}{\#(wo_{i}) \cdot q(co_{j})} \right) - log\kappa \right]_{+} \tag{2}$$

The matrix Y can be acquired from skip-gram (SG) of the corpus. Thus, each element can be defined using Y_{ij}. The unigram distribution (UD) for sampling co_j of a context is $q(co_j)$. Hence, the expression for $q(co_j)$ is,

$$q\left(co_j\right) = \frac{\#\left(co_j\right)^{\gamma}}{\sum_{co_j \in \mathrm{X}} \#\left(co_j\right)^{\gamma}} \tag{3}$$

Where, the smoothing factor is represented as γ. The major problem with the basic SeaNMF model is that it reveals only the internal relationships among the context and its words. Moreover, SeaNMF is only related to short-text corpus, which is a critical problem to overcome. In this work, to boost the performance together with the internal corpus and additional external corpus is incorporated. Based on the internal and external corpora, a word-context matrix is generated using the following formula,

$$TC\left(wo_i, wo_j\right) = \lambda_E\left(wo_i, wo_j\right) + \lambda_L\left(wo_i, wo_j\right) \tag{4}$$

Where, the weight term computed for the words wo_i, wo_j using internal and external corpus is expressed as λ_E, λ_L. The calculation of similarity score with word 2vector model of Gensim library for both internal and external corpus is expressed as,

$$\lambda_E\left(wo_i, wo_j\right) = \begin{cases} x, \text{if} x > 0 \\ 0, \text{otherwise} \end{cases} \tag{5}$$

$$\lambda_L\left(wo_i, wo_j\right) = \begin{cases} 1, \text{if } wo_i, wo_j \in d_l \\ 0, \text{otherwise} \end{cases} \tag{6}$$

Where, the model similarity score is represented as x, and the document term is $d_l(1 \leq l \leq n)$. In G_SeaNMF, the semantic correlation matrix is calculated using the metric PMI. The objective function that relates the semantic information and term_document matrix can be expressed as,

$$\min_{Wo, Wo_c, R \geq 0} \left\| \begin{pmatrix} T^T \\ \sqrt{\alpha} Y^T \end{pmatrix} - \begin{pmatrix} R \\ \sqrt{\alpha} Wo_c \end{pmatrix} Wo^T \right\|_F^2 + \psi(Wo, Wo_c, R) \tag{7}$$

Where, $\alpha \in \Re_+$signifies the scale parameter and the penalty function is represented as $\psi(Wo, Wo_c, R)$. Here, an optimization algorithm named BCD is used to optimize the weight values. The other name of BCD is Nonlinear Gauss-Seidel (NGS). BCD is one of the simpler forms of iterative algorithms used for performing non-convex optimization, which updates the weight values at each iteration. The latent factor matrices of words (Wo), Context (Wo_c)and documents (R) are randomly initialized with non-negative numbers. Within every iteration, the coordinates are updated column-wise. The updated form of equations is characterized as:

Update Wo

$$Wo_{(:,k)} \leftarrow \left[Wo_{(:,k)} + \frac{(TR)_{(:,k)} + \alpha(YWo_c)_{(:,k)} - \left(WoR^TR\right)_{(:,k)} - \alpha\left(\boldsymbol{WoWo}_c^T Wo_c\right)_{(:,k)}}{\left(R^TR\right)_{(k,k)} + \alpha\left(Wo_c^T Wo_c\right)_{(k,k)}} \right]_+ \tag{8}$$

Update Wo_c

$$Wo_{c_{(:,k)}} \leftarrow \left[Wo_{c_{(:,k)}} + \frac{(YWo)_{(:,k)} - \left(Wo_c Wo^T Wo\right)_{(:,k)}}{\left(Wo^T Wo\right)_{(k,k)}} \right]_+ \tag{9}$$

Update R

$$R_{(:,k)} \leftarrow \left[R_{(:,k)} + \frac{\left(T^T Wo\right)_{(:,k)} - \left(RWo^T Wo\right)_{(:,k)}}{\left(Wo^T Wo\right)_{(k,k)}} \right]_+ \tag{10}$$

Where, $[z]_+ = max(z, 0), \forall z \in \Re$. Within every iteration, the coordinates are updated column-wise. After that, $Wo_{(:,k)}$ and $Wo_{c(:,k)}$are normalized to have a unit $l_2 - norm$. This iteration is repeated simultaneously till the algorithm converges. Therefore, consideration of word-context semantic relationship in the overall updating procedure yields highly correlated top keywords under each topic.

The flowchart of Proposed G_SeaNMF Model is illustrated in Fig. 2. The proposed G_SeaNMF based short text topic modelling is performed using the local and global word-context correlation. Initially, the data is being collected from the short-text data corpus. The data is pre-processed and cleaned using the steps such as lowercase conversion, stemming, stop word removal, lemmatization, normalization and text enrichment procedures. Net, vocabulary and term_document matrix is created. The word correlation matrix is generated by incorporating both local, global corpus. The semantic relations are learned among the words and their contexts using the skip-gram view of the corpus. The proposed approach offers substantial semantic terms under each topic by integrating the word2vec model of Gensim library trained on the Google News Dataset and SeaNMF model. Finally, the topics with top words are displayed. The evaluation of techniques based on DMM, self-aggregation approaches, and global word co-occurrence are considered in terms of purity and NMI. Five real-world datasets are used for experimentation: Search Snippet, Biomedicine, Pascal Flickr, Tweet and TagMyNews.

4 Results and discussion

This section explains the valid performance of the proposed G_SeaNMF model by performing substantial experiments on the various real-world datasets. A detailed explanation of the datasets used, baseline approaches, evaluation metrics and the results obtained are discussed in this section.

4.1 Details on the datasets

The performance of several short text topic modelling methods is assessed and contrasted in this section through testing on five distinct data sets. Five real-world datasets are used to test

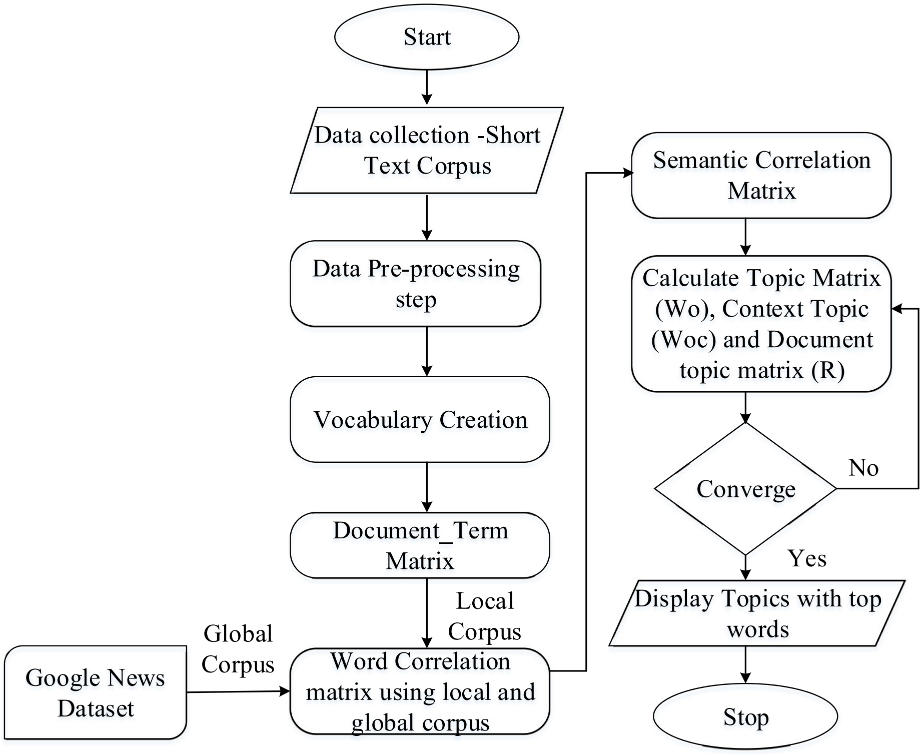

Start
Data collection -Short Text Corpus
Data Pre-processing step
Vocabulary Creation
Document_Term Matrix
Local Corpus
Global Corpus
Google News Dataset
Word Correlation matrix using local and global corpus
Semantic Correlation Matrix
Calculate Topic Matrix (Wo), Context Topic (Woc) and Document topic matrix (R)
Converge
No
Yes
Display Topics with top words
Stop

Fig. 2 Flowchart of Proposed G_SeaNMF Model

the performance of short text algorithms. The statistics about the datasets used are mentioned in Table 1.

Search Snippet: It comprises summaries of documents from eight different categories that were given as an outcome of a search engine query. Eight categories are business, computer, arts, engineering, health, education, politics, and sports. There is a total of 12,295 documents.

Biomedicine: There are 19,448 documents available on BioASQ's official website, covering a total of 20 topics.

Pascal Flickr: It has a collection of captions from 20 different domains in 4834 documents [27].

Tweet: It includes 2472 tweets with 89 topics. Tweets are collected from microblog tracks at Text Retrieval Conference (TREC) 2011 and 2012.

TagMyNews: The dataset contains 32,600 English news articles that were taken from the websites of three popular news publications. It includes news from seven distinct domains,

Table 1 Datasets Statistics

Dataset Name	Total Documents	Total Labels
Search Snippet	12,295	8
Biomedicine	19,448	20
Pascal Flickr	4834	20
Tweet	2472	89
TagMyNews	32,600	7

including sports, business, the United States, health, science and technology, the world, and entertainment. Experimentation with the title and description of news is being considered [1].

4.2 Evaluation metrics

The performance metrics used for computing the topic modelling are topic coherence or PMI (Point wise Mutual Information), Normalized PMI (NPMI), NMI, perplexity, and purity [10, 33, 39].

4.2.1 Topic coherence

PMI: This metric is also called topic coherence or UCI coherence. It is defined as the associative measure among the given word-pair (wo_i, wo_j), which is expressed as,

$$UCI(k) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI\left(wo_i, wo_j\right) \tag{11}$$

$$PMI\left(wo_i, wo_j\right) = log \frac{p\left(wo_i, wo_j\right)}{p(wo_i)p\left(wo_j\right)} \tag{12}$$

Where, the word probability wo_i and wo_j that co-occur in the similar document is defined as $p(wo_i, wo_j) = \#(wo_i, wo_j)/D$. The number of total word_context pairs can be denoted as $D = \sum_{wo,\ co \in Vo} \#(wo, co)$. The marginal probabilities are represented as $p(wo_i) = \#(wo_i)/D$ and $p(wo_j) = \#(wo_j)/D$.

Normalized PMI: This metric indicates the maximum correlation based on human judgment in evaluating topics. The expression to evaluate normalized PMI is

$$Normalized\ PMI\left(wo_i, wo_j\right)^{\gamma} = \left(\frac{log \frac{P\left(wo_i, wo_j\right)}{P(wo_i)\cdot P\left(wo_j\right)}}{-log\left(P\left(wo_i, wo_j\right)\right)}\right)^{\gamma} \tag{13}$$

Where, the greater values of γ offers more weights to normalized PMI values.

Perplexity: This is one of the evaluation metrics used to evaluate the performance of proposed topic modelling. The expression for the computation of perplexity is

$$P = exp\left(\frac{-\sum_{n=1}^{N} \sum_{i=1}^{L_m} log(p(wo_{ni}))}{\sum_{n=1}^{N} L_n}\right) \tag{14}$$

Where, the number of total documents is symbolized as N, the length of document is represented as L_n, and the likelihood of i^{th} word in the document is denoted as $p(wo_{ni})$.

4.2.2 Cluster evaluation

Purity: This metric is used to determine the consistency of a cluster, that is, the extent to which the cluster component belongs to a single class. Purity values range between 0 and 1. A high purity rating suggests that the cluster is of high grade [26]. For a given set of Clusters C, a given set of class labels *L*, and a certain total number of documents *N*. The expression for purity is defined as follows:

$$Purity(C,L)=\frac{1}{N}\sum_{c\in C}\max_{l\in L}|c\cap l| \tag{15}$$

NMI: This metric is used to compute the trade-off among the quality of clusters against the cluster numbers. NMI calculates the MI (Mutual Information) shared among clusters and cluster labels normalized via the mean of entropy of clusters and classes. The NMI range is 0 to 1 like purity metric. The formula to calculate NMI is,

$$NMI(C,L)=\frac{I(C,L)}{[H(C)+H(L)]/2} \tag{16}$$

$$I(C,L)=\sum_k\sum_j\frac{|co_j\cap lo_k|}{M}log\frac{M|co_j\cap lo_k|}{|co_j||lo_k|} \tag{17}$$

$$H(C)=-\sum_j\frac{|co_j|}{M}log\frac{|co_j|}{M},H(L)=-\sum_k\frac{|lo_k|}{M}log\frac{|lo_k|}{M} \tag{18}$$

Where, the number of total tweets is denoted as *M*, total tweets in the lo_k cluster is represented as $|lo_k|$, the overall tweets in co_jcluster is symbolized as $|co_j|$, the tweets that occur is both lo_k and co_j clusters are indicated as $|co_j\cap lo_k|$.

4.2.3 Classification accuracy

Accuracy: This metric evaluates the distribution of documents and topics. In topic modelling, the performance can be measured using text classification. Higher accuracy illustrates that the topics discovered are more representative and discriminative. The classification is performed with the package named LIBLINEAR, considering linear SVM. In this work, the classification accuracy can be evaluated with 5-fold cross-validation on the different datasets such as Tweet, Search snippet, Pascal Flickr, Biomedicine and Tag My News datasets.

4.3 Baseline comparison models

The performance of the proposed G_SeaNMF approach is compared with the following baseline methods. The conventional approaches to short text topic modelling can be broadly classified into three categories: DMM (Dirichlet Multinomial mixture), self-aggregation and global word co-occurrence.

4.3.1 Methods based on DMM

DMM [36] was introduced on the premise that each document is sampled from a single topic. This was a reasonable assumption for a short text document, as it comprises a small amount of words. Collapsed Gibbs sampling was used in the first DMM model. Twitter-LDA (Latent Dirichlet Allocation) was proposed by Zhao et al. [37] on the assumption that each tweet was related to a single subject. Recent work has incorporated word embedding into DMM. Regular LDA was predicated on assuming that a single document may include several topics. Gibbs Sampling by Dirichlet Multinomial Mixture (GSDMM) was a variant of LDA that assumes that each document has linked to a specific topic. GSDMM was capable of automatically determining the number of clusters and solving sparsity and high dimension problems. It uses Dirichlet distribution as the prior distribution for topic-word and document-topic distributions. It samples a topic for the document and creates all words in the document using a multinomial distribution.

Latent Feature Dirichlet Multinomial Mixture (LF-DMM) model relies on one-document and one topic assumption. This model substitutes the topic-to-word Dirichlet multinomial component with a combination of Dirichlet multinomial and latent feature components to produce the words [22]. The Generalized Polya Urn Dirichlet Multinomial Mixture (GPUDMM) focuses on the semantic resemblance between two words through a cosine similarity measure and pre-trained word embedding. It has been found that the assumption of individual documents for a single topic may be overly strict for some datasets. Model PDMM (Polya Urn Dirichlet Multinomial Mixture) has a topic number as a Poisson distribution [18]. This enables each short text to contain one to three topics. To improve the outcome, PDMM has extended the model as GPUPDMM by using prior knowledge about word semantic relations via GPUDMM. This method was not time-efficient [14].

4.3.2 Global word co-occurrence based method

When two words occur together in a document, it is referred to relevant words. Considering such a global word co-occurrence when learning latent themes from a corpus increases clustering performance. There was a need to standardise the window length used to extract word co-occurrences from individual documents. Because short text contains fewer words, many approaches treat each document as a sliding window. The Biterm Method (BTM) extracts topics from a short text by creating biterms directly [5]. Biterms are Unordered word pairs that appear in a corpus are called Biterem. BTM acquires topic knowledge by aggregating biterms over the entire corpus. This concept enables the handling of sparsity in an individual document [16].

The Word Co-occurrence Network (WNTM) uses standard Gibbs Sampling for latent topic discovery and learns distributions across the topics for each word from the word co-occurrence network instead of document topics WNTM less sensitive by the length of the document. In WNTM, the sliding window advances word by word, and the co-occurrence of two separate words in the same window is considered. It then creates a word network with the word as the vertex and the weight on the edges, representing the frequency of co-occurrence of two connected words. This simplifies and enhances the use of WNTM in real-world applications [38]. SeaNMF method was developed for determining the themes of a short text by revealing the meaningful relationship between keywords and their context. Skip gram was used to discover semantic associations between keywords and their context during the training phase of the word embedding approach. This helps to mitigate the sparseness of short text. Here each page was considered as a single sliding window [31].

4.3.3 Method based on self-aggregation

To address the issue of sparseness in short text, the self-aggregation method merges short texts into a pseudo-long document. The topic modelling technique is applied to a pseudo document [18]. Self-Aggregation based Topic Modelling (SATM) derives the underlying topic from the pseudo document using conventional topic modelling. It first evaluates the likelihood of a pseudo-document existence based on a short text document in the corpus. Then, each word estimates a pair of pseudo-document label and topic label. This method may lead to overfitting, which is computationally expensive [26]. Pseudo Document-based Topic Modelling (PTM) is predicated on the notion that a vast number of short texts are created from a small number of regular-sized latent documents, known as pseudo documents. Each short text corresponds to a single pseudo document. The distribution of short text over pseudo documents is modelled using a multinomial distribution [11].

4.4 Result analysis

In this section, the evaluation results are expressed with the graphical analysis and explanations. The efficacy of the proposed G_SeaNMF approach is analyzed on the five real-world datasets and compared with the baseline models. The fundamental task processed for evaluating topic models is verified in terms of topic coherence, purity, perplexity, and NMI. Table 2 illustrates the performance of proposed G_SeaNMF model evaluated using various evaluation measures.

Figure 3 represents the purity metric comparative analysis. The quality of clusters is measured with the purity metric to represent the effective characteristics of the proposed G_SeaNMF model. The baseline models used for the comparison are BTM, WNTM, DMM, GPUDMM, LF-DMM, PTM, SATM, and SeaNMF. With the datasets, the purity value estimated using the proposed G_SeaNMF are search snippet (0.82), biomedicine (0.47), pascal Flickr (0.34), TagMyNews (0.62), and tweet (0.9), respectively. The proposed G-SeaNMF outperforms BTM and WNTM in the word embedding group of methods. From the DMM family of algorithm, LF-DMM is the best algorithm for Search Snippet, Pascal Flickr, and TagMyNews. PTM has improved performance from Self Aggregation methods than SATM for all datasets. In general, when three types of topic modelling methods are compared, the self-aggregation technique is the least performer.

Figure 4 indicates the comparison of the NMI metric. The mutual information shared among the clusters and cluster set is computed using NMI measure. The baseline models used for the comparison are BTM, WNTM, DMM, GPU-DMM, LF-DMM, PTM, SATM, and SeaNMF. With the datasets, the NMI value estimated using the proposed G_SeaNMF are search snippet (0.57), biomedicine (0.37), pascal flickr (0.29), TagMyNews (0.45), and tweet (0.89), respectively. The performance of the approaches above has been observed to be dataset dependent. The proposed G-SeaNMF outperforms BTM, DMM, LF-DMM and GPUDMM on the tweet dataset. G_SeaNMF is the best algorithm for Search Snippet, Pascal Flickr, and TagMyNews. The baseline models WNTM, LF-DMM, PTM, and SeaNMF gained equivalent

Table 2 Outcomes of proposed G_SeaNMF model using various metrics

Method	Dataset	Metrics				
		Purity	NMI	Perplexity	NPMI	PMI
Proposed (G_SeaNMF)	Search Snippet	0.822	0.573	6728.3	0.023	1.6329
	Biomedicine	0.482	0.367	5223.6	0.034	1.8314
	Pascal Flickr	0.297	0.294	1853.7	0.028	4.1338
	TagMyNews	0.676	0.426	5642.7	0.030	3.6503
	Tweet	0.821	0.814	2639.2	0.038	4.2467

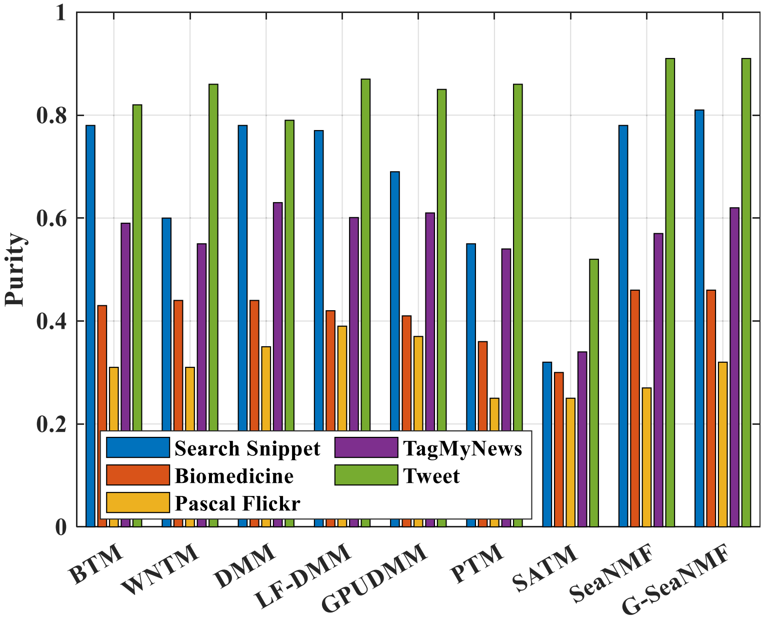


Fig. 3 Comparison of Purity

performance with the proposed G_SeaNMF approach for the tweet dataset. In general, when comparing proposed G_SeaNMF with three types of topic modelling methods, the self-aggregation technique is the least performer.

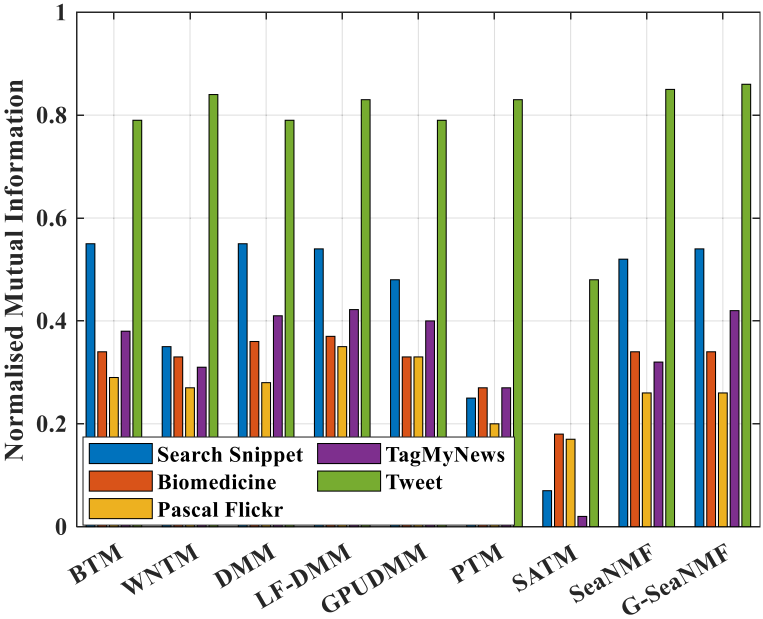
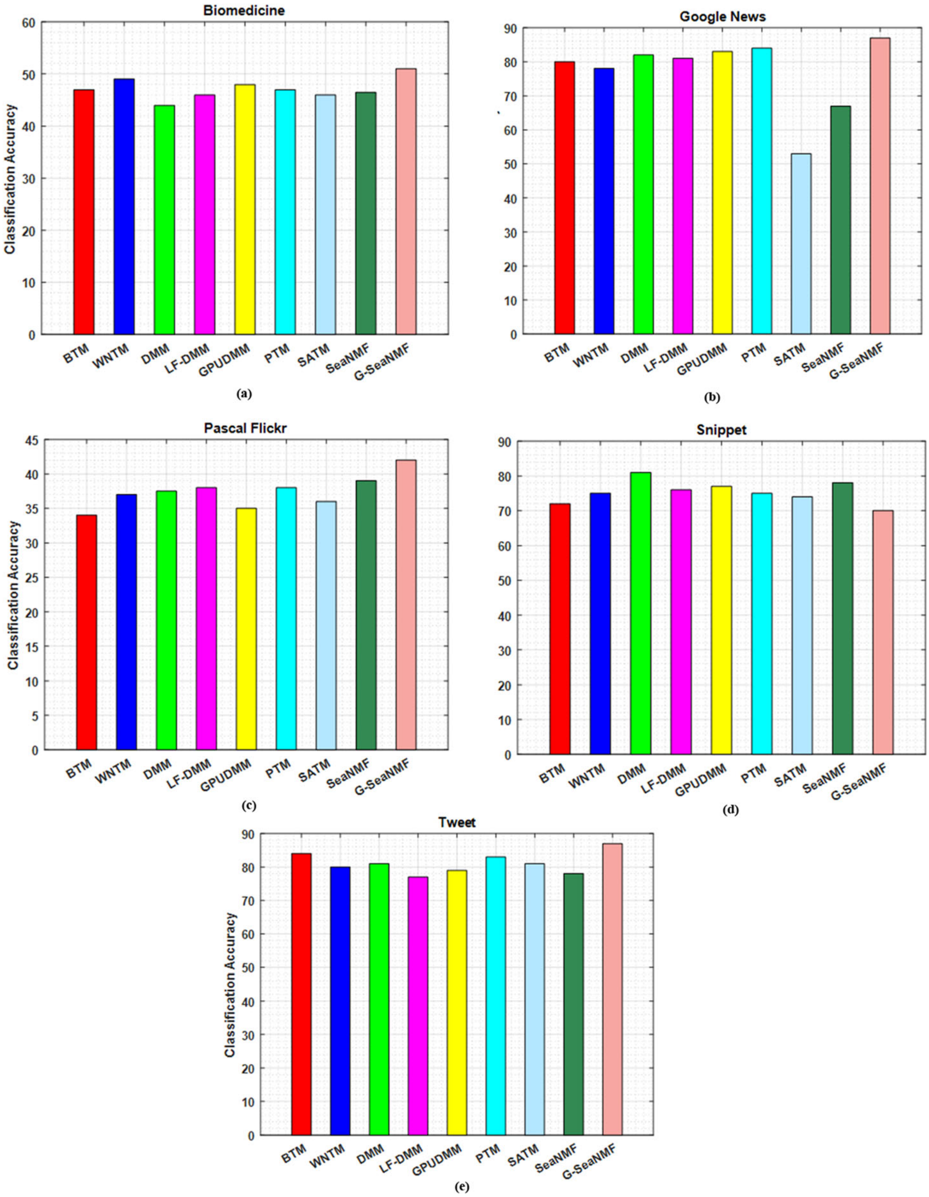


Fig. 4 Comparison of NMI

Figure 5 shows the comparison of classification accuracy. The performance is assessed using accuracy metrics on ten different models with the five datasets. It is observed that the performance of these techniques is dependent on the datasets. The proposed G_SeaNMF model using word embeddings outperforms the other models, mainly on Google News and Tweet datasets. The SATM method based on self-aggregation especially achieved lower accuracy because of generating pseudo-documents. The existing methods obtained lower



(a)

(b)

(c)

(d)

(e)

Fig. 5 Comparison of classification accuracy on different datasets

results since these models are extremely dependent on the dataset. The proposed G_SeaNMF model attains superior performance on all datasets, excluding SearchSnippets datasets.

Figure 6 displays the results of the proposed G_SeaNMF topic coherence results and all existing approaches on the TagMyNews dataset. For evaluating the outcomes, the top number of words per topic T = {5, 10, 15, 20} and number of total topics K = {10, 20, 30, 40} are set. From Fig. 6, it is observed that the proposed G_SeaNMF outperforms all the other existing methods on TagMyNews dataset. The performance of G_SeaNMF is superior when compared to SeaNMF and SATM by learning the topics from local and global co-occurrence of words. The proposed G_SeaNMF can obtain more readable topics from the outcomes.

Table 3 indicates the topics that are discovered from TagMyNews dataset. The seven randomly selected topics and their top-weighted terms for each topic are displayed. The two baseline models LF-DMM and PTM is compared with proposed G_SeaNMF with TagMyNews dataset. The contents of each topic are summarized using a topic label. The topic label offers information about the corpus. The top ten terms clustered under each topic in Table 3 were determined using LF-DMM, G-SeaNMF, and PTM, the best performing DMM, self-aggregation and global word co-occurrence algorithms, respectively on the TagMyNews dataset. The document-term matrix is used to assign topic labels. Appropriate labels are generated when words grouped in a topic share a common domain.

The TagMyNews dataset contains seven domains; LF-DMM was able to identify six of the seven topics; PTM was unable to identify specific topic labels because words within one

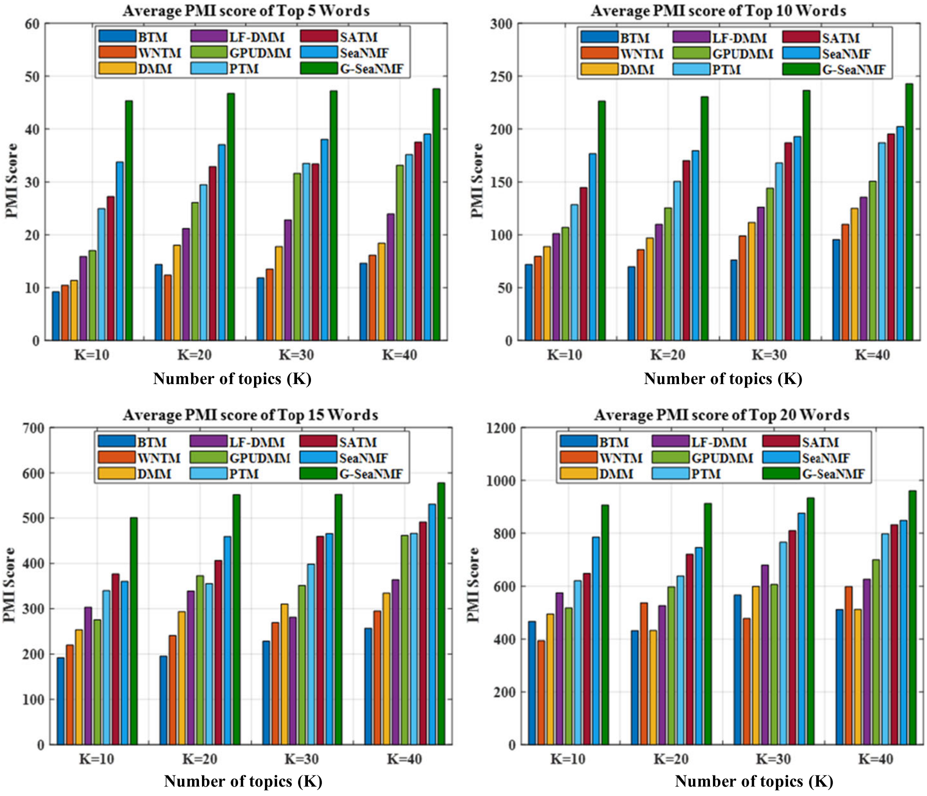


Fig. 6 Topic Coherence on TagMyNews dataset

Table 3 Topics revealed by proposed and baseline models on TagMyNews dataset

Methods/Topics	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
	(us)	(sci and tech)	(business)	(world)	(environment)	(sports)	(health)
G_SeaNMF (Proposed)	forces	film	profit	charges	river	win	study
	protests	movie	market	court	flooding	victory	researchers
	protesters	theatre	shares	accused	homes	beat	disease
	troops	tv	investors	judge	water	scored	patients
	syria	software	percent	case	residents	roundup	cancer
	opposition	nfl	growth	prosecutors	tornadoes	playoff	suggests
	military	online	sales	trial	caused	innings	risk
	protest	phone	prices	guilty	miles	inning	drugs
	leader	broadway	earnings	former	floods	points	health
	rebels	musical	inc	jury	tornado	game	finds
LF-DMM	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
	(us)	(sci and tech)	(sports)	(environment)	(world)	(business)	(*)
	trial	new	game	star	libyan	new	japan
	authorities	mobile	cup	new	president	drug	nuclear
	arrested	buy	season	world	said	risk	us
	said	company	final	take	killing	health	shares
	police	judge	playoff	show	leader	school	prices
	killing	software	win	sheen	killed	report	said
	state	nfl	knicks	today	afghan	republican	oil
	tornado	online	innings	movie	leader	many	billion
	tuesday	phone	league	tv	opposition	state	companies
	drug	game	coach	theater	gadaffi	federal	bank
PTM	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5(sports)	Topic 6	Topic 7
	(*)	(*)	(*)	(sports)		(*)	(*)
	us	new	said	first	first	said	new
	said	nfl	us	game	season	president	may
	japan	players	killed	league	win	people	would
	nuclear	says	forces	red	final	government	could
	billion	show	bin	two	open	wednesday	one
	year	tv	death	hit	coach	thursday	state
	company	video	laden	new	victory	state	make
	oil	game	city	yankees	game	obama	states
	inc	week	security	points	team	friday	like
	prizes	study	man	beat	nba	tuesday	likely

(*) The topic label cannot be predicted.

Table 4 Topics revealed by proposed and baseline models on Search Snippet dataset

Methods/Topics	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
	(politics)	(health)	(sport)	(education)	(culture)	(business)	(computer)
G_SeaNMF (Proposed)	party	health	games	edu	music	financial	intel
	democracy	disease	sports	research	film	business	computer
	culture	cancer	football	science	art	bank	device
	congress	healthy	league	resources	movie	trade	intel
	democratic	treatment	soccer	graduate	fashion	economic	software
	communist	drug	hockey	university	com	market	linux
	presidential	food	com	faculty	books	news	network
	war	physical	golf	center	arts	law	digital
	political	medical	game	exam	band	services	web
	philosophy	care	tennis	books	rock	stock	hardware
LF-DMM	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
	(politics)	(health)	(sport)	(education)	(culture)	(business)	(computer)
	culture	health	music	edu	art	market	computer
	political	cancer	news	research	film	business	software
	democracy	information	sports	theory	movie	information	programming
	culture	gov	com	science	com	news	web
	party	healthy	football	school	fashion	gov	memory
	wikipedia	medical	game	information	wikipedia	stock	wikipedia
	war	news	movie	journal	motor	services	intel
	republic	nutrition	wikipedia	university	arts	finance	com
	government	disease	tennis	computer	books	home	data
	information	hiv	games	physics	movies	com	internet
PTM	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
	(politics)	(health)	(sport)	(education)	(culture)	(business)	(computer)
	wikipedia	health	sports	science	movie	news	computer
	political	cancer	news	research	music	information	software
	wiki	information	com	edu	film	business	edu
	encyclopedia	medical	games	journal	com	services	research
	system	gov	football	art	news	market	school
	democracy	disease	game	resources	movies	stock	university
	party	news	soccer	information	reviews	trade	programming
	system	healthy	world	culture	video	gov	information
	war	nutrition	tennis	library	imdb	home	web
	house	hiv	culture	directory	intel	profit	system

cluster belong to distinct labeled documents; and G-SeaNMF was able to generate appropriate labels for all seven classes. The proposed G_SeaNMF can obtain a greater number of readable topics from the outcomes. The baseline models gained higher outcomes, but the topics are hard to understand, and the topics are not discovered in better form. The term (*) symbolizes that the topic label cannot be predicted. Thus, the proposed G_SeaNMF model can better learn the hidden semantic structure of short text collections.

Table 4 represents the topics modelled with the Search Snippet dataset. The latent topics analysed with the qualitative valuation of Search Snippet dataset includes seven topics as health, politics, education, sports, business, culture, and computer. Every topic is visualized by means of top 10 words. The table shows that the proposed G_SeaNMF model achieved equivalent coherence with topic coherence and obtained a greater number of coherent topics. Except for the LF-DMM, PTM, models, the proposed G_SeaNMF can identify more relevant topics. Thus, the existing approaches PTM, LF-DMM represented topics with more meaningless words compared with G_SeaNMF approach.

5 Conclusion

The widespread usage of social media in daily life has necessitated the analysis of the huge unstructured data scattered throughout the internet. In this work, a Gensim_SeaNMF model is introduced to find the topics for short texts. The proposed approach integrated local and global word_context semantic correlation, which completely overcomes the data sparsity issue. This article examined DMM, self-aggregation, and global word co-occurrence based on short text datasets. These approaches are tested using two topic coherence measures, Purity and NMI, on five datasets. LF-DMM, a member of the DMM family of algorithms, learns more coherent topics with less noisy and meaningful words but is highly sluggish due to optimising topic vectors. The proposed G_SeaNMF outperforms SeaNMF, BTM and WNTM when using Global Word Embedding-based algorithms. Moreover, the proposed G_SeaNMF employs a local and global word embedding to discover semantic associations between words, revealing meaningful words and improving results. PTM and SATM from the self-aggregation approach perform poorly compared to all other approaches. Because these methods infer latent themes from pseudo documents using traditional topic modeling, they raise the computing cost of determining word similarity. For the TagMyNews Dataset, G_SeaNMF, LF-DMM, and PTM yield purity values of 0.613, 0.601, and 0.527 and NMI values of 0.449, 0.422, and 0.283 respectively. The G_SeaNMF has the highest purity, indicating that topics are consistent and that most words come from a single category. This results in the right assignment of labels to each topic. The main limit of the proposed G_SeaNMF model is processing with only one larger external corpus dataset to obtain pre-trained word representations. Moreover, the processing speed is slow with the complexity in trade-off between implicit and explicit short-text relations. In future, this study may focus on improving the consistency of the topic model and the auto-creation of labels for each topic.

Data availability No data Availability.

Declarations

Conflict of interest Authors S.A. Kinariwala, S.N. Deshmukh declares that they have no conflict of interest.

References

1. A3 Lab. TagMyNews Dataset. http://acube.di.unipi.it/tmn-dataset
2. Albalawi R, Yeap TH, Benyoucef M (2020) Using topic modeling methods for short-text Dta: a comparative analysis. Front Artif Intell
3. Asmussen CB, Møller C (2019) Smart literature review: a practical topic modelling approach to exploratory literature review. J Big Data 6(1):1–18
4. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
5. Cheng X, Yan X, Lan Y, Guo J (2014) BTM: topic modeling over short texts. Knowledge Data Eng, IEEE Trans 26:2928–2941
6. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) LIBLINEAR: a library for large linear classification. J Mach Learn Res 9:1871–1874
7. Gao W, Peng M, Wang H, Zhang Y, Xie Q, Tian G (2019) Incorporating word embeddings into topic modeling of short text. Knowl Inf Syst 61(2):1123–1145
8. Hofmann T (1999) Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, ACM 50–57
9. Hofstätter S, Rekabsaz N, Lupu M, Eickhoff C and Hanbury A (2019) Enriching word embeddings for patent retrieval with global context. In European Conference on Information Retrieval, Springer, Cham, 810–818
10. Huang R, Yu G, Wang Z, Zhang J, Shi L (2013) Dirichlet process mixture model for document clustering with feature partition. Knowledge and data engineering. IEEE Trans 25(8):1748–1759
11. Jin O, Liu NN, Zhao K, Yu Y, Yang Q (2011) Transferring topical knowledge from auxiliary long texts for short text clustering. Proceed 20th ACM Int Conf Inform Knowledge Manag, ACM. 775–784
12. Kinariwala SA, Deshmukh SN (2020) Short text topic modeling with empirical learning. Indian J Comp Sci Eng 11:510–516
13. Kuang D, Choo J, Park H (2015) Nonnegative matrix factorization for interactive topic modeling and document clustering. In Partitional Clustering Algorithms, Springer:215–243
14. Li C, Wang H, Zhang Z, Sun A, Ma Z (2016) Topic modeling for short texts with auxiliary word embeddings. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM 165–174
15. Liang W, Feng R, Liu X, Li Y, Zhang X (2018) GLTM: a global and local word embedding-based topic model for short texts. IEEE Access 6:43612–43621
16. Ligutom C, Orio JV, Ramacho DAM, Montenegro C, Roxas RE, Oco N (2016) Using topic modelling to make sense of typhoon-related tweets. In 2016 international conference on Asian language processing (IALP). IEEE. 362-365
17. Liu Z, Qin T, Chen KJ, Li Y (2020) Collaboratively modeling and embedding of latent topics for short texts. IEEE Access 8:99141–99153
18. Mahmoud H (2008) P’olya urn models. Chapman and Hall/CRC
19. Mazarura JR (2015) Topic modelling for short text. University of Pretoria, PhD diss
20. Mikolov T, Chen K, Corradoand G, Dean J (2013). Efficient estimation of word representation in vector space. arXiv preprint arXiv:1301.3781.
21. Murakami R, Chakraborty B (2022) Investigating the efficient use of word embedding with neural-topic models for interpretable topics from short texts. Sensors 22(3):852
22. Nguyen DQ, Billingsley R, Du L, Johnson M (2015) Improving topic models with latent feature word representations. Trans Assoc Comput Ling 3:299–313
23. Nikolenko SI, Koltcov S, Koltsova O (2017) Topic modelling for qualitative studies. J Inf Sci 43(1):88–102
24. Qiang J, Chen P, Ding W, Wang T, Xie F, Wu X (2016) Topic discovery from heterogeneous texts. Tools with artificial intelligence (ICTAI). IEEE 28th iNternational Conference 196–203
25. Qiang J, Chen P, Wang T, Wu X (2017) Topic modeling over short texts by incorporating word embeddings. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Cham, pp 363–374
26. Quan X, Kit C, Ge Y, Pan SJ (2015) Short and sparse text topic modeling via self-aggregation. Proceed 24th Int Conf Artif Intell. 2270–2276
27. Radev D (2016) Effects of creativity and cluster tightness on short text clustering performance. Proceed 54th Annual Meeting Assoc Comput Linguistic 1:654–665
28. Rashid J, Shah SMA, Irtaza A (2019) Fuzzy topic modeling approach for text mining over short text. Inf Process Manag 56(6):102060
29. Roccetti M, Marfia G, Salomoni P, Prandi C, Zagari RM, Kengni FLG, Bazzoli F, Montagnani M (2017) Attitudes of crohn’s disease patients: Infodemiology case study and sentiment analysis of facebook and twitter posts. JMIR Public Health Surveill 3(3):e7004

30. Rolim V, De Mello RFL, Kovanovic V, Gaševic D (2019) Analysing social presence in online discussions through network and text analytics. In 2019 IEEE 19th international conference on advanced learning technologies (icalt). IEEE 2161:163–167
31. Shi T, Kang K, Choo J, Reddy CK (2018) Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. Proceedings of the 2018 world wide web conference on world wide web, international world wide web conferences steering committee. 1105–1114
32. Singhal T, Liu J, Blessing LTM, Lim KH (2021) Analyzing scientific publications using domain-specific word embedding and topic modelling. In 2021 IEEE international conference on big data (big data), 4965-4973
33. Yan X, Guo J, Lan Y, Xu J, Cheng X (2015) A probabilistic model for bursty topic discovery in microblogs. AAAI 29:353–359
34. Yang S, Huang G, Cai B (2019) Discovering topic representative terms for short text clustering. IEEE Access 7:92037–92047
35. Yi F, Jiang B, Wu J (2020) Topic modeling for short texts via word embedding and document correlation. IEEE Access 8:30692–30705
36. Yin J, Wang J (2014) A dirichlet multinomial mixture model-based approach for short text clustering. Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining 233–242
37. Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. Advan Inform Retri:338–349
38. Zuo Y, Zhao J, Xu K (2016) Word network topic model: a simple but general solution for short and imbalanced texts. Knowl Inf Syst 48(2):379–398
39. Zuo Y, Wu J, Zhang H, Lin H, Wang F, Xu K, Xiong H (2016) Topic modeling of short texts: a pseudo-document view. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM 2105–2114

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.