Artificial Intelligence Review (2023) 56:5133–5260
https://doi.org/10.1007/s10462-022-10254-w

Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis

Belal Abdullah Hezam Murshed[1,2] · **Suresha Mallappa**[1] · **Jemal Abawajy**[3] · **Mufeed Ahmed Naji Saif**[4] · **Hasib Daowd Esmail Al-ariki**[5,6] · **Hudhaifa Mohammed Abdulwahab**[7]

Accepted: 17 August 2022 / Published online: 26 October 2022

Abstract

Social media platforms such as (Twitter, Facebook, and Weibo) are being increasingly embraced by individuals, groups, and organizations as a valuable source of information. This social media generated information comes in the form of tweets or posts, and normally characterized as short text, huge, sparse, and low density. Since many real-world applications need semantic interpretation of such short texts, research in Short Text Topic Modeling (STTM) has recently gained a lot of interest to reveal unique and cohesive latent topics. This article examines the current state of the art in STTM algorithms. It presents a comprehensive survey and taxonomy of STTM algorithms for short text topic modelling. The article also includes a qualitative and quantitative study of the STTM algorithms, as well as analyses of the various strengths and drawbacks of STTM techniques. Moreover, a comparative analysis of the topic quality and performance of representative STTM models is presented. The performance evaluation is conducted on two real-world Twitter datasets: the Real-World Pandemic Twitter (RW-Pand-Twitter) dataset and Real-world Cyberbullying Twitter (RW-CB-Twitter) dataset in terms of several metrics such as topic coherence, purity, NMI, and accuracy. Finally, the open challenges and future research directions in this promising field are discussed to highlight the trends of research in STTM. The work presented in this paper is useful for researchers interested in learning state-of-the-art short text topic modelling and researchers focusing on developing new algorithms for short text topic modelling.

Keywords Big data · Social media · Short text topic modeling · Data streaming · Coherence · Sparseness · Deep learning topic modeling

✉ Belal Abdullah Hezam Murshed
belal.a.hezam@gmail.com

Extended author information available on the last page of the article

1 Introduction

The massive advancement of communication and information technologies has resulted in a huge volume of data of multiple varieties from multiple sources. In recent years, the advent of internet media, e.g. (blogs), information websites, e.g. (Wikipedia), and the invincible Social Media Platforms (SMP) e.g. Facebook, Weibo, Twitter, etc. rapidly, have become the major sources of such massive information (Ediger et al. 2010). Extracting and analyzing this information is one of the pivotal areas of study for various research and business entities (Ruths and Pfeffer 2014). The analysis of this data is useful for various aspects such as product/service opinion mining, market analysis, trend analysis, event detection, and so on (Malleson and Birkin 2012). Moreover, extracting topics can be helpful for detecting events such as natural disasters to act rapidly and mitigate the disaster's impact (Kraft et al. 2013; Earle et al. 2011; Oh et al. 2010), supporting political parties (Tumasjan et al. 2010), companies and organizations to figure out customers opinions about their brands, and to ameliorate content marketing by better understanding customer requirements (Ren and Wu 2013). Topic Modeling (TM) is the process of automatically discovering the latent/hidden thematic structure from a set of documents/short text and facilitates building new ways to browse and summarize the large archive of text as topics (Nikolenko et al. 2017). It also aids in organizing, understanding and summarizing the large collection of data into specialized topic labels.

Many extensive research studies have focused on long text TM. The traditional long text TM models such as Latent Dirichlet Allocation (LDA) (Blei et al. 2003), Latent Semantic Analysis (LSA) (Deerwester et al. 1990), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999), and Non-negative Matrix Factorization (NMF) (Lee and Seung 2001) are popular for discovering latent semantic topic structures in long texts. They do not require prior annotation or labelling processes. These traditional topic models visualize each text and document as a mixture of different themes (topics) which are distributed over the text. They employ statistical approaches such as variational techniques and Gibbs sampling to infer the trending topics of each document through complex order word co-occurrence patterns (Ostrowski 2015). Recently, the Short Text (ST) in social media networks is gaining more interest because of common people's raw and direct thoughts. However, modelling such short text topics is quite challenging mainly due to their short length and the limited number of words. The scarce content in the short text makes it more challenging to find the co-occurrence of topic patterns. The application of long text TMs for short text is not promising in terms of performance due to the absence of word co-occurrence in short texts (Abdel-Hafez and Xu 2013). Therefore, the research community focuses on addressing the problem of data sparsity in short texts such as Twitter data to provide efficient topic modelling.

Earlier studies on short text TM employed the traditional topic models but with added external metadata sources for knowledge extraction to formulate the word co-occurrences. Some studies employed the strategy of extracting latent topics from long documents and inferred them to extract short text topics. However, these models failed to achieve the expected results both due to the limited availability of external sources for metadata and their expensive deployment costs. Hence, the research community started considering the use of specialized designed short text topic models instead of conventional topic models. Therefore, we are motivated to aggregate such studies and provide a holistic taxonomy which can help the research community to develop new efficient TM for short text on social media.

This article aims at presenting a holistic survey, comparative analysis, and an expanded taxonomy for the most recent and ever-growing efficient TM approaches in social media. It mainly focuses on most aspects of Traditional Short Text Topic Modeling (TSTTM) models such as (probabilistic models, matrix factorization (non-probabilistic), Exemplar based, clustering techniques, dynamic-based categories, data source based, labelling based, word types, application-based,, and Frequent Pattern Mining (FPM) techniques) and Advanced Short Text Topic Modeling (ASTTM) based models such as (Dirichlet Multinomial Mixture (DMM) based models, Global word co-occurrences based models, and Self-aggregation based models, Deep Learning Topic Modeling (DLTM) based models). Despite the fact that various surveys have already been proposed in the literature of short text topic modelling, our work focuses on a more comprehensive and holistic survey and taxonomy along with qualitative and quantitative analysis. Further, a comparative empirical analysis of the most efficient recently proposed TM approaches.

1.1 Paper contributions

In general, our contributions in this taxonomy and analysis article can be stated as follows:

- A comprehensive review to investigate the various existing STTM models.
- A taxonomy of existing TSTTM and ASTTM models.
- A qualitative analysis of the STTM models based on their strengths and weaknesses.
- Review and quantitative analysis of the utilized datasets by STTM.
- Review and summarize the useful software tools and open-sources libraries for STTM.
- A quantitative analysis of the STTM models based categories, publication year, and evaluation Metrics.
- A comparative study of ten TSTTM and ASTTM models to evaluate their performance through an experimental analysis based on two real-world datasets: the Real-World Pandemic Twitter dataset (RW-Pand-Twitter) and Real-World Cyberbullying Twitter dataset (RW-CB-Twitter).
- An overview of the open challenges and the future research directions in this promising field.

This study boosts the existing surveys by providing a detailed and up-to-date comprehensive review and taxonomy, which helps researchers in understanding and utilizing the key elements of STTM. Moreover, it aids in finding out the limitations of currently available STTM techniques, open research issues, and challenges by which they can decide their future research direction.

1.2 Paper organization

This paper is organized as the following: The recent proposed surveys on topic modelling and short text topic modelling are discussed in Sect. 2. The Problem Formulation of STTM is formulated in Sect. 3. The topic modelling process flow and the prominent traditional topic models applied for short texts and the advanced STTM methods are described in Sect. 4. Then, Sect. 5 reviews the existing datasets utilized by STTM along with qualitative analysis. Section 6 summarizes and reviews the tools and open-source library for topic modelling. Quantitative analysis of the literature is presented in Sect. 7, while Sect. 8 describes the utilized dataset for experiments, common evaluation metrics, and the experimental results of the prominent methods. The overall observations noted from the qualitative and quantitative analysis, as well as the comparative analysis, are discussed in Sect. 9. Section 10 highlights the open challenges and suggestions for future directions of STTM. Finally, Sect. 11 concludes the article.

2 Existing surveys

Many existing surveys and review articles were proposed for long text TM approaches and short texts. In this section, the recently proposed surveys on topic modelling for short text are reviewed, analyzed, and compared with our suggested taxonomy. To this extent, (Li and Lei 2021; Vayansky and Kumar 2020; and Kherwa and Bansal 2020) are the most recent articles that provide a survey and analysis of all prominent research studies on the topic models. Alghamdi and Alfalqi (2015) conducted a survey of the TM in text mining. Only common methods in TM are presented. Jelisavčić et al. (2012) presented a review of the eminent probabilistic TM for providing prospective inspiration for research direction. Xia et al. (2019) provided a survey of TM by classifying the models into the traditional (probabilistic) evaluation and Hybrid topic modelling. However, as positioned in recent research, short text topic modelling is efficient through specialized strategies and techniques, which make them different from long text models.

Several studies have focused on the evolution of short text topic models. Most studies in sentiment analysis and event detection from social media data were related to topic modelling. Stieglitz et al. (2018) presented an elaborate analysis on social media analytics, including the topic discovery models. The authors addressed the research gap in data discovery, collection and processing and also elaborated upon the analytics techniques for social media data processing models with larger volumes. Hasan et al. (2018) conducted a survey on real-time Twitter event detection techniques that analyses most of the recent techniques that are similar to the functioning of topic models.

Ibrahim et al. (2018) surveyed topic detection models for tweets and evaluated their performance. The authors divided the techniques into five classes based on their functioning properties and discussed the most prominent research techniques. Although this article covered all major techniques, the survey focused mainly on the topic models that are popular for tweet streams. However, the advanced STTM techniques, data source based, labelling based, word types, and application-based models are not covered in this survey. Mottaghinia et al. (2020) explored various techniques to extract and discover topics of tweets. Then, they grouped the topic detection techniques into four classes of categories. Kaur and Bansal (2019) surveyed topic extraction techniques on Twitter. The authors reviewed the topic extraction techniques suggested by experts for the reliable collection of information. However, it considered only the techniques that use attribute-based extraction, which seems to be efficient only in limited applications. Likhitha et al. (2019) also conducted a detailed survey on the topic models for short texts in documents through semantic structure detection. Many different strategies were reviewed, but the survey is not so impressive since most techniques are the same as those of the long text topic models.

Qiang et al. (2020) conducted a comparative survey on short text topic modelling techniques and analysed their performances and applications. The authors categorized topic models into three categories and evaluated them on benchmark datasets. They also developed a comprehensive open-source library for short text topic modelling in Java language that combines all the survey methods and their evaluations with ample scope for including future techniques. However, compared to our survey, this survey did not cover all the ASTTM techniques; it only covered 8 of 62 methods. Also, a quantitative analysis of the literature did not perform in this survey.

Nugroho et al. (2020) provided yet another significant survey of topic models for Twitter data for different feature-based strategies and evaluated them on different datasets. The authors identified the main algorithms and provided an extensive analysis of Twitter streams. They focused mainly on the content, social and temporal features-based topics. Similarly, Dutta et al. (2020) conducted a survey that exploited the spatiotemporal topic models for Twitter data. This survey has provided brief descriptions about almost all the recent topic models for Twitter, focusing on spatiotemporal topic analysis. However, this study contains only a limited number of reviews due to the sparse number of relevant articles. Burkhardt and Kramer (2019a) conducted a survey of topic modelling based on multi-label methods by grouping the methods according to various variants dimensions. The authors summarized the most widely multi-label topic modelling models from a variety of studies that have been conducted for this purpose. Albalawi et al. (2020) provided a survey covering the topic modelling and its tools, applications, and methods. The authors compared and tested five common TM methods such as NMF, LSA, LDA, Principal Component Analysis (PCA), and random projection and applied them to the short-text data to show their superiority in extracting topics.

For nearly two decades, TM has been a successful text analysis technique. Integrating this technique with Deep Neural Networks (DNN) has opened a new and growing research area. The Neural Topic Models (NTM) have emerged with various models and a wide range of applications for NLP such as text generation, summarization etc. Zhao et al. (2021) presented a taxonomy and comprehensive survey for NTM; this taxonomy classifies NTM based on their backbone framework to facilitate researchers in exploring this fast-growing research area. Recently, NTM has attracted more attention since it takes the benefits of both the probabilistic topic models and neural networks. Several works related to this approach have been proposed in the literature and showed their effectiveness compared to traditional probabilistic models. Doan and Hoang (2021) presented an empirical analysis for some of the state-of-the-art NTM in various aspects over large, diverse datasets in terms of the set of metrics. Specifically, they analyzed the performance of these models in different three tasks: (1) uncovering cohesive topics, modelling the input documents, and representing them for downstream classification. The evaluation results illustrated that the neural topic models are effective in both tasks the first and third, whereas in many cases, the traditional probabilistic models are still effective in the second task. However, these findings and insights enable researchers to easily select off-the-shelf topic modelling toolboxes in various contexts. Table 1 depicts a comparative analysis of the existing surveys on short text topic modelling.

It is notable that, as indicated in Table 1, there are numerous existing reviews and surveys in the state of the art for short text topic modelling. But they did not take into account all the elements of STTM, as done in this taxonomy; it is distinct from other existing surveys by being holistic, extensive, and up-to-date. It comprehensively reviews and investigates both TSTTM, and ASTTM models along with their different sub-categories; it analyzes them based on their performance, respective strengths and weaknesses. This taxonomy presents the survey and qualitative analysis of the literature datasets used to evaluate the STTM models. Besides, it summarizes and reviews the useful software tools and open-sources libraries for STTM. Further, it also provides a quantitative analysis of the STTM models based on publication year, sub-categories, and platform. It also provides the quantitative analysis of datasets utilized by STTM based on Language and data sources.

Table 1 Comparative analysis on the existing surveys for STTM

Survey	TSTTM models												ASTTM models				DsA	TSC	QLA	QNA	CS
	PM	MF	EB	US	DB	DS	SLB	MLB	WT	AB	HTB	FPM	DMM	GWC	SAB	DLTM					
Jelisavčić et al. (2012)	✓				✓				✓												
Alghamdi and Alfalqi (2015)	✓	✓			✓																
Ibrahim et al. (2018)	✓	✓	✓	✓								✓							✓		✓
Likhitha et al. (2019)	✓	✓											✓		✓						
Burkhardt and Kramer (2019a)							✓	✓											✓		✓
Xia et al. (2019)	✓	✓									✓										
Vayansky and Kumar (2020)					✓									✓	✓				✓		
Kherwa and Bansal (2020)	✓	✓								✓									✓		
Mottaghinia et al. (2020)				✓			✓	✓											✓		✓
Albalawi et al. (2020)	✓	✓																	✓		✓
Qiang et al. (2020)													✓	✓	✓				✓		✓
Li and Lei (2021)	✓																			**T**✓	
Zhao et al. (2021)																✓					
Doan and Hoang (2021)																✓					✓
Proposed	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

PM probabilistic models, *MF* matrix factorization (non-probabilistic), *EB* exemplar based, *US* unsupervised techniques, *DB* dynamic-based, *DS* data source-based, *SLB* single labelling based, *MLB* multi-label based, *WT* word types, *AB* application-based, *FPM* frequent pattern mining, *GWC* global word co-occurrences based models, *SAB* self-aggregation based models, and *DMM* Dirichlet multinomial mixture based models, *DLTM* Deep learning topic modeling, *DsA* datasets analysis, *TSC* tools and source code, *QLA* qualitative analysis, *QNA* quantitative analysis, **T**✓ quantitative analysis only for traditional, *CS* comparative study

Moreover, it compared some of the prominent techniques and evaluated their performance on two real-world Twitter datasets (RW-Pand-Twitter and RW-CB-Twitter datasets). Finally, it highlights the challenges and open issues related to STTM techniques which facilitate in finding the inefficient and efficient topic modelling techniques and their merits and demerits by researchers. This survey is intended to improve the efficiency of learning cutting-edge methods and to identify potential research gaps, allowing researchers to choose their research directions. To the best of our knowledge, it will broaden their minds and pave the way for future proposing new techniques in the future.

3 Problem formulation

Let $\mathcal{T}_x$ denotes a topic detection algorithm. Where, the subscript x is used to indicate various algorithms, i.e., when $x = ALG$, the $\mathcal{T}_{ALG}$ represents any STTM model. Let $\mathcal{Q}$ represents the optimal quality of discovered topics by $\mathcal{T}_{ALG}$. Assume that A represents the optimal topic accuracy. $\mathcal{P}$ represents the optimal precision of discovered topics by $\mathcal{T}_{ALG}$. Let $\mathcal{R}$ represents the optimal recall of discovered topics by $\mathcal{T}_{ALG}$. Let us assume that algorithm ALG is available and always finds the optimal quality of discovered topics. The short text topic modelling problem is formulated as follows:

Given a set of m short text social media documents (posts), $D = \{d_1, d_2, \dots, d_m\}$, a vocabulary W with size V and K pre-defined latent topics, where each short text $d_i \in D, 1 \le i \le m$ is unstructured and unlabelled or labelled posts. The size of each short text $d_i \in D$ is given as:

$$S_i = \sum_{k=1}^{n} \sum_{j=1}^{z} c_{kj} \tag{1}$$

Note that each short text (post) $d_i \in D$ is represented by a set of n words, $d_i = \{w_{i1}, w_{i2}, \cdots, w_{in}\}$ such that each word $w_{ik} \in d_i, 1 \le k \le n$ is composed of z characters $w_{ik} = \{c_{k1}, c_{k2}, \dots, c_{kz}\}$.

A multinomial distribution represents the topic φ in a specific collection D over the vocabulary W that means $\{p(w|\varphi)\}_{w \in W}$. A multinomial distribution over K topics represents the representation of topic of a documents or short text d θ_d that means $\{p(\varphi_k|\theta_d)\}_{k=1,\dots,k}$.

The STTM problem is to represent short texts D as a set of K topics $\{\varphi_{k=1,\dots,k}\}$ under the following constraints:

$$S_{min} \le |S_i| \le S_{max} \tag{2}$$

$$|d|_{\mathcal{W}} \approx \mathcal{T}_{ALG,\mathcal{W}}, \mathcal{W} \in \{\mathcal{Q}, A, \mathcal{P}, \mathcal{R}\} \tag{3}$$

The primary objectives of the STTM for a given of m short text posts D can be stated as follows: (a) Learn how topics φ are represented in words, and the sparse topic representation of short texts (posts) θ. Constraint (2) indicates that the size of each tweet or post should not be less than the minimum characters limit (S_{min}) and should not be greater than the maximum characters limit (S_{max}) in short text data (post). In our case, S_{max} contains 280 characters, including blank space. The minimum size can be set depending on the quality of the short texts received. The constraint (2) is formulated specifically for Twitter social media data. Constraint (3) deals with the optimality of the four metrics, namely accuracy, recall, and precision, and quality of the discovered topics.

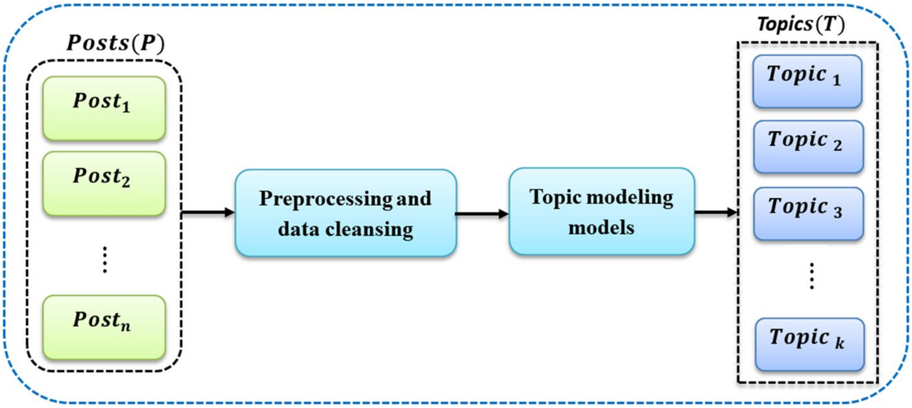
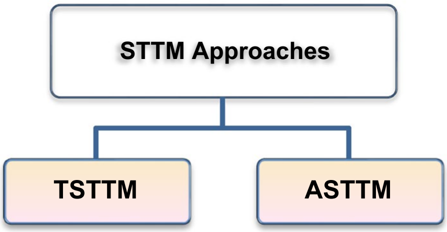


Fig. 1 Topic modelling process flow

Fig. 2 Taxonomy on STTM models



4 Topic modelling

This section presents the general topic modelling process flow, as depicted in Fig. 1 and the taxonomy of STTM models, where STTM models are broadly classified into two main categories: Traditional Short text Topic Modeling (TSTTM) and Advanced Short Text Topic Modeling (ASTTM) models, as depicted in Fig. 2. The TSTTM models are discussed in Sect. 4.2, and ASTTM models are discussed in Sect. 4.3.

4.1 Topic modelling process flow

Social media becomes a significant information source; such information comes in the form of tweets or posts which are short in nature. Discovering the potential topics from such tweets and posts is important for many natural language processing (NLP) tasks, such as automatic summarization, document classification, content analysis, emerging topic detecting, question answering, sentiment analysis, recommendation systems, information retrieval and etc. However, Topic modelling (TM) is a key technique that has been used for extracting knowledge and latent topics from a short text in social media. Generally, TM can be defined as the process of automatically extracting and identifying topics from short texts.

A typical approach for creating topics from short text involves three key sub-tasks: (i) data acquisition, (ii) pre-processing, and (iii) topic modelling method. The initial phase, as shown in Fig. 1, is to collect unstructured and semi-structured data from data sources such as Twitter. The dataset consists of tweets collected from multiple topics of interest. The Twitter dataset could be collected using the Twitter streaming API using Python language with tweepy package. Following that, the baseline pre-processing step is applied to the dataset to clean up the data using toolkits such as the NLTK python package (Anil Phand and Chakkarwar 2018) that provides stop-word and punctuation removal, tokenization, lemmatizing, stemming, identifying n-gram procedures, and lowercase transformation and also other pre-processing and data cleansing steps (Murshed et al. 2021). Finally, the TM method is applied to extract a set of recurrent themes/topics that are explored throughout the collection of posts and the extent to which each post reflects those themes. The output of TM is a set of topics which can be further used to explore, visualize, and summarize posts and tweets.

4.2 TSTTM models

The generalized topic models are developed primarily for extracting latent topics from long text documents. However, many studies applied them for STTM (Niyogi and Pal 2019; Shirolkar and Deshmukh 2019; and Hidayatullah et al. 2019). Others attempted to combine short texts into long text documents and using traditional TM (Al-Sultany and Aleqabie 2019; and López-Ramírez et al. 2019). Some others have designed by modifying the strategies of long text models to be applied for short texts, especially in news and tweets data (Zhao et al. 2011; Quercia et al. 2012; Fang et al. 2017; Sharath et al. 2019; Wang et al. 2012; and Han et al. 2020). Figure 3 shows the overall classification of TSTTM that can be used for short texts. This part of taxonomy classifies the TSTTM models into eleven sub-categories: probabilistic models, matrix factorization (non-probabilistic), Machine learning-based (unsupervised and supervised) techniques, dynamic (strategies) based categories, exemplar-based, data source based, word types, application-based, Frequent Pattern Mining (FPM) techniques, and Hybrid based. All the models in these categories can be used for both long text and short text. However, this article focuses on short texts models.

4.2.1 Probabilistic based models

Probabilistic TMs are a suite of models that apply statistical solutions to extract and uncover the latent thematic structure in a massive collection of documents and decompose and deconstruct its documents into topics. One of the most important assumptions of these probabilistic models is that the generative process follows a bag-of-words (BOW) assumption, which means that each token (word) is independent from the token that came before it. This section presents the probabilistic topic modelling models with their extensions, which classifies as follows:

4.2.1.1 Latent Dirichlet Allocation (LDA) based models LDA is a probabilistic statistical approach based on the de Finneti's theorem for extracting the significant statistical structure in a text document. It is one of the most popular utilized techniques for topic discovery and extraction models (Blei et al. 2003). The basic idea in LDA is that each document is represented as a probability distribution over hidden topics, while each topic is characterized as a probability distribution over a number of words. The generative process of the LDA model for each document or short text $d_i \in D$ in a dataset D is written as following.

A. Draw each topic parameter $\beta_k \sim$ Dirichlet (φ), for $k \in \{1 \dots K\}$

B. For each document:

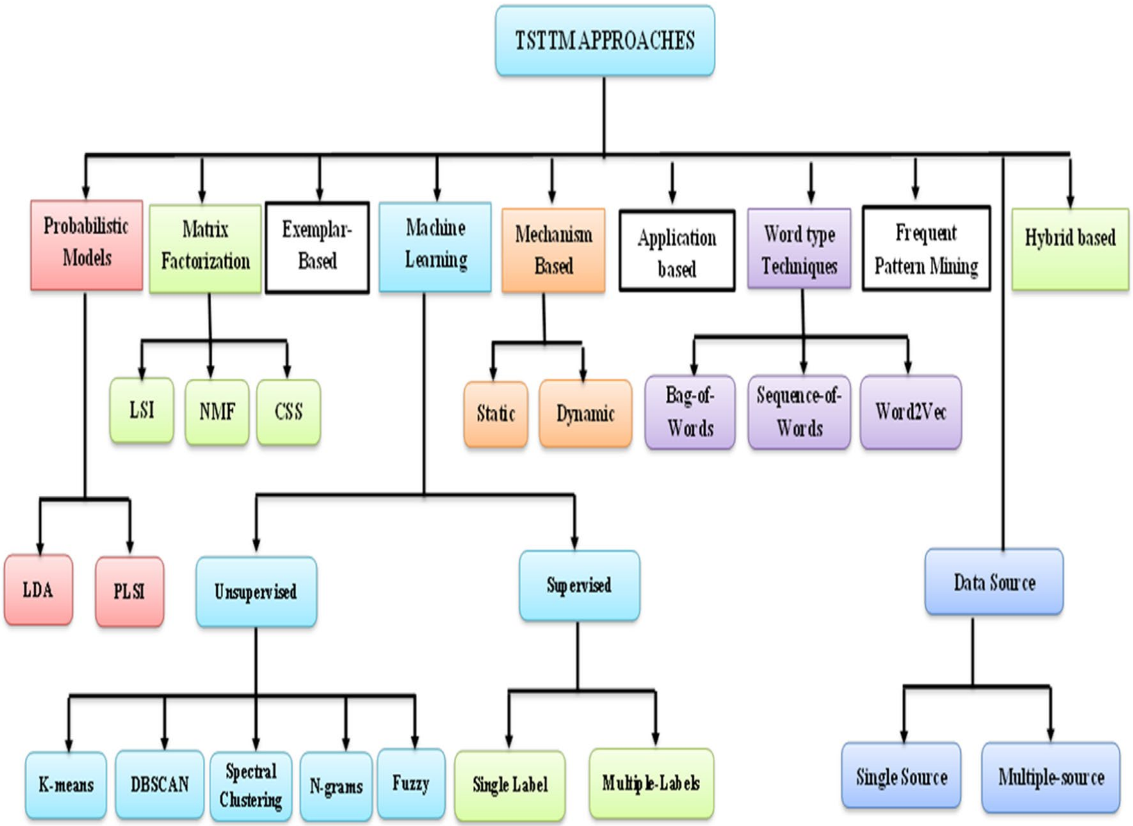


Fig. 3 Taxonomy on TSTTM models

1. Choose the topic distribution $\theta_m \sim$ Dirichlet (α)
2. For each of the N words w_n:
   i. Select a topic $z_{mn} \sim \text{Multinomial}\left(\theta_m\right)$
   ii. Select a word $w_n \sim \text{Multinomial}\left(\beta_k\right)$ from $p(w_n|z_{mn}, \beta_k)$

Where the dataset level parameters are represented by the parameters β and α, which are sampled just once in the procedure of generating the dataset, K denotes the number of topics, and the word-level variables that are taken just once for every word in every document are denoted by z_{mn} and w_{mn}. The document-level variable that is sampled just once per short text (document) is denoted by the parameter θ_m. Finally, the word probability distribution for the topic k is denoted by φ_k. The time complexity of LDA algorithm is $O\left(N_{itr}KN\bar{I}\right)$, where N_{itr} denotes the number of iterations, K is the number of hidden topics, N denotes the number of documents in the dataset, and $\bar{I}$ is the average length of each document in D.

Many studies apply LDA for topic modelling of both short texts and long texts. Hoffman et al. (2010) proposed a new model named Online LDA (OLDA), which is based on online stochastic optimization with a natural gradient step. Online LDA can handle and analyze the enormous number of documents, including streaming document collections. The time complexity of Online LDA is $O\left(N_{itr}K|N^{(t)}\bar{I}^{(t)}\right)$, where N_{itr} denotes the number of iterations, K is the number of hidden topics, N denotes the number of documents in the dataset, $\bar{I}$ is the average length of each document in D, and the superscript of t represents the latest time-slice or version. Al-Sultany and Aleqabie (2019) utilized LDA for enriching tweets topic modelling by merging the tweets into long documents and linking them to Wikipedia for latent topic discovery. The procedure of merging tweets into long text, Twitter Name Entity Recognition (TNER) was suggested to categorize the short text tweets, extract entities, and linking the entities for each short text tweet with Wikipedia to construct a new Twitter dataset. Then, the pre-processing tasks are conducted. Finally, the topic modelling is applied by LDA. Similarly, Niyogi and Pal (2019) used LDA to discover conversational topics associated with India Demonetization tweets. Further, López-Ramírez et al. (2019) used LDA for extracting the geographical collection of topics. Shirolkar and Deshmukh (2019) also used LDA for finding topic experts in the Twitter dataset. Hidayatullah et al. (2019) utilized LDA for extracting the weather and climate condition topics on Twitter. Chan (2020) developed real-time social big data analytics using LDA topic model. LDA is based on the BOW (Bag-Of-Words) for extracting the features of interest topics but fails to combine the short text features and interest attributes. Further, the standard LDA doesn't seem to reflect the dynamic and hierarchical trend of microblogs users' interest.

As the limitations of LDA seem to impact its' performance for short texts, the focus of the research community was shifted towards a modification of the traditional LDA. Along this direction, Chen and Kao (2017) developed an improved topic model using Re-Organized LDA (RO-LDA), which seems to resolve the lack of local word co-occurrence of LDA. However, this model has limitations in terms of handling redundant data. Fang et al. (2017) utilized Time-Sensitive Variational Bayesian inference LDA (TSVB-LDA) for extracting trending topics with higher levels of accuracy. However, this model has limitations in terms of inference of news tweets. Ni et al. (2018) presented the hot event detection model using Background Removal LDA (BR-LDA) topic modelling, which removes the background words from tweets.

Sharath et al. (2019) designed the Corpus-based Topic Derivation(CTD), which combines LF-LDA (Latent Feature-LDA) using an asymmetric topic model and Timestamp-based Popular Hashtag Prediction (TPHP) to discover Twitter topics based on corpus semantics. Korshunova et al. (2019) developed the Logistic LDA for discriminative topic modelling of tweets which can extract topics from unlabelled data through unsupervised exploitation of the data group structure. However, Logistic LDA is closer to LDA and in which it has limited inference abilities. Tajbakhsh and Bagherzadeh (2019) designed the Semantic knowledge LDA with topic vector extracting tweet topics based on the co-occurrence of word for recommendation system. Zheng et al. (2019) developed the Three-layer Interest Network LDA (TIN-LDA) model to discover topics from tweet data through interest attributes. The TIN-LDA significantly extracts the semantic correlation among the keywords and thus enhancing the coherence of the topics. Wang et al. (2012) presented Temporal LDA (TM-LDA) for latent temporal topic extraction from tweets. It has been applied to a huge volume of data and updated when the new temporal stream data comes. Although high accuracy was achieved, it uses time only for collecting tweets. In some of the current topic representations, there is no attention about how to choose terms with better differentiation for representing topics, Han et al. (2020) developed a topic representation model based on user behaviour analysis and LDA, namely (MBA-LDA) for analyzing the data of microblogging. Topic-word distribution is obtained by the LDA model. This model addressed the problem and considered the words distribution and the user behaviour information to re-appraise the significance of words for topic representation. Some topic models like LDA generate incoherent topics with noisy terms. Such these topic models suffer from a lack of semantic data, data sparsity, and the binary weighting of words, so Akhtar et al. (2019b) developed a new model based on fuzzy document representation with LDA, namely (FBLDA) to handle these issues. Besides, Ozyurt and Akcayol (2021) Suggested a new topic modelling, namely Sentence Segment LDA SS-LDA model, to extract the aspect product of user reviews in attempt to overcome data sparsity. Rahimi et al. (Rahimi et al. 2022) proposed a novel model named LLDA, which concentrated on local word relationships and encoded the word orders using overlapping windows. The authors suppose that a document consists of overlapping windows of fixed size, and a novel generative process is formulated appropriately. According to the inference process, every word is only sampled just once in a single window while influencing the sampling of its co-occurring counterparts in the other windows. The LLDA model alleviates the sparseness problem and generates more coherent topics.

4.2.1.2 Twitter LDA Twitter LDA or Tweet-LDA is an extended version of LDA specifically designed for tweet topic modelling with additional data pre-processing techniques and data interpretation from sparse tweets. Zhao et al. (2011) designed Twitter-LDA for extracting topics from noisy tweets. Since the general LDA has one topic label for each word, it may not work well and is not suitable with Twitter due to the tweets are very short and noisy in nature, and every single tweet is more likely to be a single topic. Therefore, the Twitter-LDA is proposed to fill this gap. The generative process of the Twitter-LDA is written as follows:

```
A. Draw $\varphi^B \sim$ Dirichlet (β),$\pi \sim$ Dirichlet (γ)
B. For each topic $t = \{1, \dots, T\}$
  (a) draw $\varphi^t \sim$ Dirichlet (β)
C. For each user $u = 1, \dots, U$
    (i) draw $\theta^u \sim$ Dirichlet (α)
    (ii) for each tweet $s = 1, \dots, N_u$
      (a) Draw $z_{u,s} \sim$ Multinomial($\theta^u$)
      (b) For each word $n = 1, \dots, N_{u,s}$
        (i) Draw $y_{u,s,n} \sim$ Multinomial($\pi$)
```

(ii) Draw $w_{u,s,n} \sim \text{Multinomial}\left(\varphi^{B}\right)$

If $y_{u,s,n} = 0$, and $w_{u,s,n} \sim$ Multinomial($\varphi^{z_{u,s}}$) if $y_{u,s,n} = 1$.

Formally, let T denote the topics in Twitter data, each is defined by word distribution. The word distribution for the background words and topic t are denoted by φ^{B} and φ^{t}, respectively. A Bernoulli distribution that controls the selection among topic and background words is indicated by π and the topic distribution of the user u is denoted by θ^{u}. When composing a tweet, a user initially selects a topic based on his topic distribution. Then, he selects a BOW (Bag of Word) individually based on the selected background model or topic.

Other studies use the LDA model, with some modifications of the original model for Twitter data. Quercia et al. (2012) also designed another Tweet-LDA by modifying the Labelled LDA as a supervised topic classification model for tweets with a Support Vector Machine (SVM). Both methods provided better topic coherence when compared to the LDA model. Akhtar (2017) also used Twitter-LDA for the hierarchical summarization of news tweets. Yu et al. (2019) designed Twitter Hierarchical LDA (TH-LDA) for discovering topics in tweets for On-Line Analytical Processing (OLAP). It mines hierarchical dimensions automatically and uses the word2vec model to analyse the semantic relationships. However, it focuses only on direct relationships while ignoring the other indirect social relationships. Table 2 provides the comparative analysis of probabilistic based TSTTM models.

4.2.1.3 PLSI based models PLSI is another extensively utilized document model, which stands for probabilistic latent semantic indexing (Hofmann 1999). PLSI was introduced as an improvement to the LSI model by providing a more solid statistical foundation and identifying an appropriate generative data model compared to LSI. Besides the merits of this model compared to the conventional LSI model, since it depends on the probability principle, it can also make use of statistical models for model fitting, immediately reducing word perplexity, model combination as well as overfitting control. it seems to offer a probabilistic interpretation. The main aim of the PLSA model is to use the co-occurrence matrix for discovering the topics and exploring the documents as a set mixture of topics. It extracts the latent statistical class model as a mixed decomposition through co-occurrences among words and documents.

Formally, let D be a dataset consisting of a large collection of documents, which is represented in the document-term-matrix indicating how many times every word occurs in every document. Assuming the latent variables model is defined as the following: (1) Documents are the observed variables, $d \in D = \left\{d_1, d_2, d_3, \dots, d_m\right\}$ where m denotes the number of documents or short texts in the dataset. (2) Words: are the observed variable, where $w \in W = \left\{w_1, w_2, w_3, \dots, w_n\right\}$, and n is the number of words in the dataset. (3) Topics are hidden (Latent) variables, $z \in Z = \left\{z_1, z_2, z_3, \dots, z_K\right\}$, where K denotes the number of topics, and it should be identified priori. The generative process of the PLSI for a document in a dataset is written in the following steps.

1. Select a document d_m with the probability $p(d)$.
2. For each word w_i, where $i \in \{1, \dots, n\}$ in the document d_m:
   a. Choose a topic z_i from a multinomial conditioned on the selected document with the probability $p(z|d_m)$.
   b. Choose a word w_i from a multinomial conditioned on the selected topic with probability $p(w|z_i)$

Table 2 Comparative analysis of probabilistic based TSTTM models

References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Zhao et al. (2011)	Twitter-LDA	Assigns topics at the tweet level,, Better coherence over short texts	Does not treat both words and hashtags separately	Twitter, NYT	EN	Twitter, News	Specific	Java[a]	TC
Wang et al. (2012)	TM-LDA	Temporal feature-based topic extraction, high accuracy	Considers time only for tweet collection interval	Collect tweets	EN	Twitter		Matlab	Perplexity
Chen and Kao (2017)	RO-LDA	Resolves lack of local word co-occurrence	Difficult handling redundant data	Twitter2011, ETtoday News	EN, CHI	Twitter, News	Specific	JAVA, Python	NMI, Purity, AC, RI
Fang et al. (2017)	TSVB-LDA	High accuracy	Limitations in the inference of news tweets	Twitter	EN	Twitter	Specific	NA	TC, MD,ERR
Ni et al. (2018)	BR-LDA	Removes background words	Less ability to handle large data	Twitter, Sina Microblog	EN, CHI	Twitter, Sina Weibo	Specific	NA	R, P, F-M
Kumar and Vardhan (2019)	PLSA-ICA	Accurate sentiment analysis	Bad in extracting the deep meaning of words	Tweets	EN	Twitter	Specific (Google hashtag)	NA	AC, F-Measure
Sharath et al. (2019)	CTD	Enhance the purity and F-of topic derivation, Latent features for high accuracy	High run time	Tweets (200 M)	EN	Twitter	Generic	Java	R, Purity, P, F-M

Table 2 (continued)

References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Korshunova et al. (2019)	Logistic LDA	Extract topics from unlabeled data	Limited inference	Collects 2 sets(Tweets), NG20[c]	EN	Twitter, News	Generic	Python	AC
Zheng et al. (2019)	TIN-LDA	Overcomes the high-dimensional sparsity problem, explores the interest of microblog users, Enhances coherence	High time complexity	Sina microblog	CHI	Weibo	Generic	NA	Perplexity, KLD,TE
Yu et al. (2019)	TH-LDA	Hierarchical dimensions for high semantics	Only direct relationships considered	Collected Tweets	EN	Twitter	Generic	NA	Perplexity, PMI
Akhtar et al. (2019b)	FBLDA	Handle lack of semantic inf., data sparsity, and the binary weighting of words	Fails to determine all the topics in the dataset, trade-off between computational costs of LDA and performance based on the size of the term set	Web Snippet, Reuters	EN	Website	Generic	Java	PMI, NMI, Purity

Table 2 (continued)

References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Han et al. (2020)	MBA-LDA	Solve the problem of choosing words with better differentiation to represent a topic which not been addressed before	The performance is not good with a smaller dataset	Sina Weibo	CHI	Weibo	Generic	Python	R, P, F-M
Ozyurt and Akcayol (2021)	SS-LDA	Unsupervised does not require any annotated training data, alleviates data sparsity and resolves a lack of co-occurrence patterns	NA	UserReview, SemEval-2016 Review	TUR	e-commerce web site[b]	Specific (Smart Phone Resturant)	Java	R, P, F-M

TC topic coherence, *P* precision, *R* recall, *F-M* F-measure, *AC* accuracy, *MD* topic mixing degree, *ERR* estimate error, *NA* not mention, *EN* English, *CHI* Chinese, *TUR* Turkish

[a] https://github.com/minghui/Twitter-LDA

[b] www.hepsiburada.com

[c] http://qwone.com/~jason/20Newsgroups/

PLSI model has been used by several research works for TM. This sub-section presents recent studies of the PLSI based models. Hennig (2009) designed a topic model based on multiple document summarization using PLSA by combining both query and thematic features. Yirdaw and Ejigu (2012) also used a similar topic model with PLSA for Amharic language text summarization. However, PLSA is considered to be very vague for short texts because these methods require multiple features. PLSA was also used for certain other major applications, such as sentiment analysis in Twitter data. Kumar and Vardhan (2019) utilized PLSA in combination with the Independent Component Analysis (ICA) for aspect-based sentiment analysis of tweets. Likewise, many studies have utilized PLSA for sentiment analysis and event detection from short texts. However, there are some limitations attached to PLSA, which suffers from the problem of data sparsity when applied to short text topic modelling. Table 2 shows a summary of the probabilistic-based TM based on their perspective objectives, weaknesses, the dataset used, source, domain, and platform.

4.2.1.4 A Bayesian graphical model In this sub-section, we review the probabilistic topic modelling based on the Bayesian model. The Probabilistic topic models, which can detect latent topics or themes in the documents such as LDA, PLSI, etc., have been investigated extensively. These models learn only from a single document dataset. Therefore, several real-world applications necessitate an in-depth comprehension of the relationships between numerous document datasets. To fill this gap, (Hua et al. 2020) suggested a novel model named Common and Distinctive Topic Modelling (CDTM), which can detect and discover the common and distinctive topics from several datasets simultaneously, where the common topics (global mixtures) are the topics which are shared among all the multiple datasets, while the distinctive topics(specific topics) represent the unique features or characteristics locally in every respective dataset. The proposed model is the first model based on the Bayesian graphical approach.

Formally, let us suppose that a set of l datasets, represented by $S = \{D_1, D_2, \dots, D_l\}$. In this set of datasets, each dataset is a collection of m documents in the dataset D and indicated by $D = \{d_1, d_2, \dots, d_m\}$. Each document or short $d_i \in D, 1 \le i \le m$, is represented by n of words $d_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$. The vocabulary V of a set of datasets (S) consists of the words from all the considered datasets. The local topics (distinctive topics) is indicated by a K_d- dimensional Dirichlet variable θ_d, where the number of distinctive topics for every respective dataset is denoted by K_d. The Global topics (common topics) is indicated by a K_c- dimensional Dirichlet variable θ_c, where the number o global topics (common topics) is represented by K_c. Utilizing these variables and definitions, the distinctive and common topics are learnt as the estimation for posterior distributions of variable θ_d and θ_c.

4.2.2 Matrix factorization based models

This section presents the matrix factorization based models utilized for TM and shows their extension. Table 3 summarizes the Matrix factorization based TSTTM models.

4.2.2.1 LSI based models Latent Semantic Indexing (LSI) or Latent Semantic Analysis (LSA) is a traditional and popular text mining method that extracts the hidden semantic structure of the words from a collection of documents or short text (Deerwester et al. 1990). The existing text mining techniques before LSI were unable to retrieve data based on concepts and queries; LSI was the first technique to be introduced. The demerits of LSI model

Table 3 Comparative analysis of matrix factorization based models

Reference	Models	Objective	Limitations	Datasets		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Valdez et al. (2018)	LSA	High accuracy in a large dataset	Negative approximation values	Collected corpus of debates	EN	Politico	Specific (Politics)	NA	R, P
Karami et al. (2018)	FLSA	Avoids the negative impact of redundant data	Difficult to handle complex structures	Twitter, MuchMore,[a] Ohsumed,[b] nursing[c]	EN, GER	Twitter, others	Specific	R and Matlab	AC, F-M
Kim et al. (2020)	W2V-LSA	Alleviate the problems of sparseness and high dimensionality, High accuracy and high coherence	Entirely dependent on data for user-defined parameters	blockchain-related doc	EN	Others	Generic	NA	TC, NMPI, KMS
Yan et al. (2013b)	NMF	High term correlation and stability,	Worst mixtures	Tweets,[d] (title data)News,[e] Question[f]	EN, CHI	Twitter, website, Sogou Lab	Generic	Matlab, C++	Perplexity, NMI,AC, Purity
Belford et al. (2016)	Ensemble NMF	produces a more definitive and stable solution, High accuracy	NA	Collect 2 Twitter sets	EN	Twitter	Specific	NA	NMI
Yan et al. (2012)	N-cut-weighted NMF	Alleviate sparsity problem with short text, improve purity, ARI and NMI	Consuming High computational time	Tweets, (title data) News	EN	Twitter,	Generic	NA	NMI, ARI, Purity
Iskandar (2017)	RED-NMF	High purity and high correlation	Data sparsity problem	Collected Tweets	EN	Twitter	Specific (Pandemic)	NA	TC, Perplexity

Table 3 (continued)

Reference	Models	Objective	Limitations	Datasets		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Chen et al. (2020b)	NFM-LTM	Manifests human-like learning behaviours and considerable potential for big data topic modelling., Better term semantics	High complexity, worst mixtures, demands task arrangement plan in NFM-LTM in order to improve the ability to adapt to new tasks more swiftly	Tweets, Snippet, Reviews	EN	Twitter, Google, Amazon	Generic	C++/Java	TC,NDCG
Farahat et al. (2015)	CSS	Faster, best recall and coherence	Reduced throughput and consumed a lot of time	[Reuters-21578, MNIST-4 K, PIE-20 and YaleB-38],[g] [Reviews, LA1][h]	EN	Others	Generic	Matlab,	AC

P precision, *R:* recall, *F-M* F-measure, *AC:* accuracy, *NA* not mention, *GER* Germanize, *EN* English, *CHI* Chinese

[a] http://muchmore.dfki.de/resources1.htm

[b] http://disi.unitn.it/moschitti/corpora/ohsumed-first-20000-docs.tar.gz

[c] http://physionet.org/

[d] http://trec.nist.gov/data/tweets/

[e] http://www.sogou.com/labs/dl/tce.html

[f] http://zhidao.baidu.com

[g] http://www.cad.zju.edu.cn/home/dengcai/Data/data.html

[h] http://trec.nist.gov

are that the detected topics are hidden and ambiguous. Besides, it has the issue of negative values in its decomposed matrices which cannot be interpreted.

Let assume that X is a data matrix (documents $m \times n$ terms), and the LSI model factorizes the matrix X into the product of 3 matrices $\boldsymbol{U\Sigma V^T}$**.** This process is called Singular Value Decomposition (SVD). Figure 4 shows the process of SVD of LSA topic modelling. It can be formulated as given in Eq. (4).

$$X = U\Sigma V^T \tag{4}$$

where X denotes the data matrix with size $(m \times n)$, m represents documents and n terms, U denotes to $(m \times r)$ matrix, r represents concepts. V indicates to $(n \times r)$ matrix, V^T denotes the coefficients of terms in new space, Σ denotes a diagonal $(r \times r)$ matrix all values are equal to zero those in the diagonal.

Many recent studies have employed LSI for topic modelling and different applications of text mining. Valdez et al. (2018) presented a topic modelling framework for US election tweets using LSA and provided thematic patterns embedded in a large tweet dataset with a high degree of accuracy. However, Qomariyah et al. (2019) compared LSA and LDA for Twitter topic discovery and concluded that the LDA is better than LSA. Though it is evident that LDA is better than LSI, some studies tried to develop LSI with certain improved features to overcome its limitations in handling larger semantic structures. Karami et al. (2018) developed Fuzzy LSA (FLSA) topic discovery in health news tweets, which avoids the negative impact of redundant data. Kim et al. (2020) presented Word2Vec-based Latent Semantic Analysis (W2V-LSA) for trending topic discovery in expert tweets. However, this model is entirely dependent on data for optimizing user-defined parameters, which tends to reduce its efficiency in large scale modelling. However, the LSA model is not very suitable for short texts owing to their usage of approximation results in negative matrix values. Yet, this model is suitable for multiple applications such as document clustering (Magerman et al. 2010), language modelling (Yeh et al. 2005), etc.

4.2.2.2 NMF based models Non-Negative Matrix Factorization is a statistical and linear-algebraic model that reduces the dimensions of the input dataset. Internally, it utilizes the factor analysis model to assign relatively less weight to the less coherent words. The NMF-based methods formulate the input dataset as a matrix and learn themes or (topics) by immediately splitting the term-document matrix, which represents a text dataset as a bag-of-words matrix, into two low-rank factor dimensions matrices to extract the trending topics. It discovers the latent topical structures of the data by identifying the factor matrices. Although this is one of the efficient topic models, it is mostly considered only after the LDA

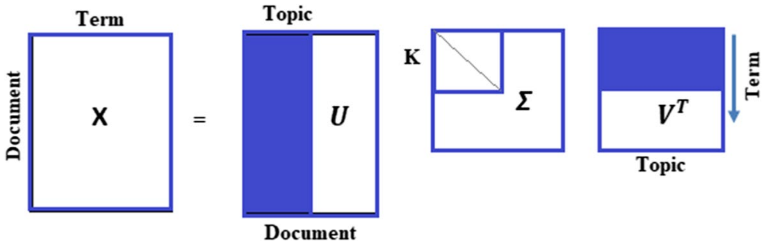


Fig. 4 SVD of the Latent semantic indexing TM

model. This is because LDA adds a Dirichlet prior on top of the data collection process, while NMF qualitatively leads to worse mixtures. As mentioned in the LSI section, the LSI has an issue with negative values in its decomposed matrices; then, the NMF was suggested to alleviate this issue.

Formally, let X be a data matrix of size $\mathrm{m} \times \mathrm{n}$, which represents the term-document matrix. Where this model of TM decomposes X matrix into the projecting of two lower-dimensional matrices named W with the size of $m \times k$ and H with the size of $k \times n$ spanned using a set of hidden or latent topics. The interpretation of these factorized matrices is that each column of the matrix W denotes the weight of every word received in every sentence and every row of matrix H is a word embedding. Since the values of all the elements of W and H factorized matrices are Nonnegative, given that all elements of X matrix are non-negative. The general process of the NMF is shown in Fig. 5 and formulated as the Eq. (5).

$$X \approx W \times H \tag{5}$$

Many recent researchers adapted NMF for topic modelling. To this extent, this section presents the most recent NMF based models for topic modelling. Belford et al. (2016) developed the ensemble topic modelling using NMF for annotated tweet data and achieved high stability and accuracy. The proposed model combines and integrates multiple unstable topic modelling methods to form one ensemble topic model with the ability to produce a stable and informative solution. Sitorus et al. (2017) utilized NMF for sensing trending urban topics on Twitter near the Greater Jakarta area. However, as described above, the limitations of NMF were observed to have a negative impact on overall performance. Hence many studies tried to modify and improve versions of NMF. Yan et al. (2013b) presented an approach using NMF on the term correlation matrix. This approach enhances both term correlation and stability. Yan et al. (2012) developed a new model, named Ncut-weight term weight (N-cut-weighted NMF) topic model, which assesses the discriminability of terms based on word co-occurrences. Murfi (2017) utilized separable NMF for topic extraction with a higher level of accuracy than LDA. Iskandar (2017) developed a regular expression discovery (RED) algorithm based NMF (RED-NMF) for disease outbreak topic extraction on Twitter. Shi et al. (2018) developed a semantics-assisted non-negative matrix factorization (SeaNMF) approach which enriches with local word-context correlations for extracting the latent topics and improves topic coherence and accuracy of classification. Lahoti et al. (2018) utilized joint NMF for learning ideological topics on Twitter with 90% purity and higher correlation. Casalino et al. (2018) designed an intelligent topic modelling framework using NMF and applied subtractive clustering to detect trending topics. Chen et al. (2020b) developed the Affinity regularized NMF named NFM-LTM for lifelong topic

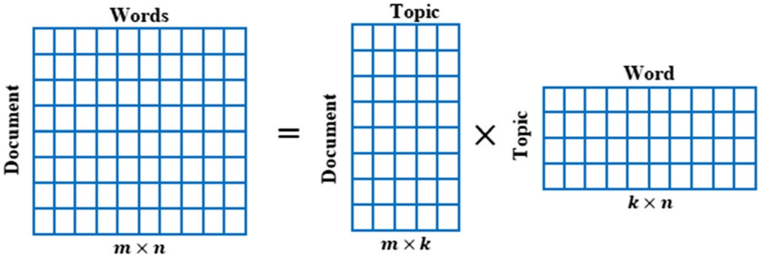


Fig. 5 NMF process

modelling in short text big data. Although this model is efficient, it demands many adjustments to support short texts due to the problem of data sparsity.

4.2.2.3 Column Subset Selection (CSS) models The CSS issue can be characterized broadly as the selection of the most representative columns from *X* data matrix in a general manner. The column subset selection generalizes the challenges of selecting features problem and representative data instances. Farahat et al. (2015) suggested an accurate and novel greedy method named CSS for tweet topic modelling, which can choose a subset of columns of a matrix to reduce the error. It is the fastest and most efficient approach for topic modelling with greedy selection. The primary goal of this method is to minimize the rate of reconstruction error of X matrix by employing the features (S) that have been selected. The approach is based on an efficient recursive formula for computing the error reconstruction of the data matrix depending on the chosen criterion for the subset of columns at every iteration. This concept was utilized in this research for the purpose of topic extraction on the Twitter dataset. Specifically, each picked data point is regarded as a representative of a particular topic. However, the greedy approach significantly reduces the throughput and consumes a lot of time for settling at optimal coherent topics.

4.2.3 Machine learning-based TSTTM models

Another important category of TSTTM is the Machine Learning (ML) based models. This section reviews some the TSTTM based on ML. Generally, the ML models can be classified into two main categories: Unsupervised and Supervised TSTTM where the supervised is classified into Single-Label and Multi-Label. The comparative analysis of the machine learning-based TSTTM models is summarized in Table 4.

4.2.3.1 Unsupervised models This section presents the clustering techniques used for topic modelling. The Clustering algorithms are generally a part of topic modelling algorithms. Most times, the topics extracted by the LDA and other topic models are clustered using any clustering algorithm for post-processing. However, in some cases, the researchers have tried employing clustering algorithms directly as a topic extraction model. Li et al. (2013) utilized incremental clustering for improved topic detection from Chinese microblogs. Yang et al. (2013) developed the hot topic detection method using CURE hierarchical clustering algorithm. Along the same lines, Fang et al. (2014) proposed Multi-View Topic Detection (MVTD) to detect hot topics from Twitter with high levels of coherence and accuracy. However, it does not consider the retweet–reply and geographical relations of tweets. Nur'aini et al. (2015) proposed a model which integrates the K-means clustering method and Singular Value Decomposition (SVD) for tweet topic extraction. Muliawati and Murfi (2017) designed a topic model using Eigen space-based Fuzzy C-Means (EFCM) clustering in which Singular Value Decomposition (SVD) is used for reduction in data dimension. Prakoso et al. (2018) extended this work by introducing Kernel Eigen space-based Fuzzy C-Means (KEFCM) for sensing detecting trending topics. Trupthi et al. (2018) utilized Probabilistic fuzzy C-means topic modelling for analysing user sentiments. Lim et al. (2017) developed Clustop by merging the concepts of clustering-based topic modelling word networks of n-grams and part-of-speech tagging. However, it models only those topic labels that are based on Wikipedia articles.

Capdevila et al. (2017) developed Tweet-SCAN, which is based on DBSCAN and used a hierarchical Dirichlet process and Jensen-Shannon distance for event discovery from

Table 4 Comparative analysis on machine learning based TSTTM models

	References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
					Name	Lang				
Unsupervised based Models	Li et al. (2013)	Incremental clustering	High stability and purity	Increased run time	Sina Weibo	CHI	Weibo	Specific	NA	MR, FR
	Yang et al. (2013)	CURE hierarchical clustering	Better topic features relationships	Complex architecture	Sina Weibo	CHI	Weibo	Generic	NA	Na
	Fang et al. (2014)	MVTD	High coherence and accuracy	Does not consider the retweet–reply and geographical relations	Collect tweets	EN	Twitter	Specific	NA	F-M NMI, Entropy
	Muliawati and Murfi (2017)	EFCM	High clustering accuracy	Topic labelling is based on Wikipedia articles	Collect tweets	EN	Twitter	Specific (politics	NA	Topic recall
	Prakoso et al. (2018)	KEFCM	Less time complexity	Topic labelling is based on Wikipedia articles	Collect 3 tweets DS	EN	Twitter	Specific(politics, sports,	NA	Topic recall, AC

Table 4 (continued)

References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Capdevila et al. (2017)	Tweet-SCAN	Accurate event detection	Difficult handling a large number of sparse texts	Collect 2 tweets DS	EN	Twitter	Specific	NA	F-M, Js distance
Mustakim et al. (2019)	DBSCAN	High clustering accuracy	Difficult handling a large number of sparse texts	Collect tweets	EN	Twitter	Generic	NA	Silhouette index
Rashid et al. (2019b)	FTM	Effective dimension reduction for local and global term frequencies, mitigate sparsity problem with FTM	The time is increasing with a various number of topic but here somewhat is stable	Snippets, Tweets, BaiduQA	EN, CHI	Weibo, Twitter, website	Generic	Matlab	NPMI, enropy, AC, purity, F-M
Indra and Pulungan (2019)	BN-grams and Doc-pivotal	Accurate and stable	Data sparsity problem	Collect 6 twitter dataset	EN	Twitter	Generic	NA	Topic recall, AC, Entropy

Table 4 (continued)

		References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
						Name	Lang				
		Abou-Of (2020)	IFCM	Reduce running time twice without losing data	Not suitable for right-to-left style language like Arabic	NG20(mini)	EN	News	Specific	Matlab	Entropy, F-M
		Ozyurt and Akcayol (2021)	SS-LDA	Unsupervised does not require any annotated training data; alleviate data sparsity	NA	UserReview, SemEval-2016 Review	TUR	e-commerce web site[a]	Specific (Smart Phone Resturant)	Java	R, P, F-M
	Single-Label	Mcauliffe and Blei (2008)	SLDA	Single label and provided high coherence	Single labels are almost out-of preference	Movie reviews DIgg	EN	News	Specific	NA	–
		Mao et al. (2012)	SSHLDA	Single label and provided high coherence	Single labels are almost out-of preference					NA	Perplexity, F-M

Table 4 (continued)

		References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
						Name	Lang				
		Lacoste-Julien et al. (2009)	DiscLDA	Single label and provided high coherence	Single labels are almost out-of preference	NG20	EN	News	Specific	NA	
		Bhattacharya et al. (2014)	Labeled LDA	High purity	Manual labelling is required	NA			NA	NA	
Supervised based models		Li et al. (2015)	DF-LDA	High accuracy	Required additional storage space	Yahoo, RCV1v2 (Lewis et al. 2004)	EN	Yahoo	Specific	NA	F-M, AUC
	Multi-label TM	Yu et al. (2017)	MS-LDA	Identifies the multi-level interests hidden in massive Twitter data for OLAP, Better use of mining features	The major limitation of this model is that it consumes a lot of running time	Collected Twitter data	EN	Twitter	Generic	NA	R, P, F-M
		He et al. (2020b)	Bi-Labeled LDA	Highly cohesive	Supports only two labels	Twitter dataset	EN	Twitter	Generic	NA	DCG, Hit No

Table 4 (continued)

	References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
					Name	Lang				
	Pang et al. (2019)	WLTM, XETM	Alleviate sparsity problem	Need to extend non-parametric model to deal with data streams where the No. of topics is difficult to determine manually	SemEval, ISEAR:	EN	News	Generic	NA	AC, HD, AP
	Wilcox et al. (2021)	SLDAX	More reliable and accurate estimation	NA		EN		Specific	R Lang	
	Wang et al. (2021a)	SL-LDA, AL-LDA	Overcomes the short-coming overfitting on noisy labels	NA	Yahoo art,[b] Health,NG20, Reuters	EN	Yahoo, News, others	Specific, Generic	Matlab	R, P, F-M

P precision, *R* recall, *F-M* F-measure, *AC* accuracy, *HD* Hellinger distance, *DCG* discounted cumulative gain, *NA* not mention, *EN* English, *CHI* Chinese, *TUR* Turkish, *MR* miss rate, *FR* false rate of topics

[a] www.hepsiburada.com

[b] https://github.com/timothyrubin/DependencyLDA

geo-located tweets. Mustakim et al. (2019) employed DBSCAN algorithm for clustering of trending tweet topic pilkada pekanbaru. However, DBSCAN has limitations in handling a large number of sparse texts. Rashid et al. (2019b) designed the Fuzzy Topic Modelling (FTM) approach, where the frequencies of the global and local terms are generated by the bag-of-words and high dimensions removed by Principal Component Analysis (PCA). Indra and Pulungan (2019) developed a trending topic detection model using BN-grams and Doc-pivotal. Abou-Of (2020) developed an incremental trending topic detection model named Incremental FCM (IFCM), which integrates the incremental semantic metrics and Fuzzy C-Mean clustering to increase the accuracy of trending topics. IFCM aims to solve such issues as discovering the semantic relatedness when different titles present the same event and to tackle incorporating semantically similar topics from various sources. Yang et al. (2019) proposed a Topic Representative Terms Discovering (TRTD) for short text, which alleviates and addresses the noise and data sparsity problem. Though analysis reveals that the clustering based topic models are largely accurate and stable, they seem to lack the effectiveness of STTM techniques owing to their dependence on topic features for clustering even in cases of sparse tweet texts.

4.2.3.2 Supervised (Label) based topic models In this section, we present the single-label and multi-label-based topic modelling techniques. Single label topic models are mostly used for classifying tweets based on a single topic. While most studies fail to undertake such labels, some researchers have already performed such tasks.

Supervised LDA (Mcauliffe and Blei 2008), SSHLDA (Mao et al. 2012), and DiscLDA (Lacoste-Julien et al. 2009) are some of the most recent single-label topic models. Most authors seem to prefer multi-label topic models such as Labeled LDA (Bhattacharya et al. 2014), LF-LDA (Zhang et al. 2018c) and DF-LDA (Li et al. 2015). Yu et al. (2017) developed the Multi-layered Semantic LDA (MS-LDA) for mining topics from unstructured tweet data with high recognition of accuracy. However, this model has a major handicap in that it consumes a lot of running time. Jun He et al. (2020b) designed the Bi-Labeled LDA for automatic detection of interest tags from Twitter topics using the social relationship between popular and non-popular users. The topic discovery in this model is highly cohesive and, therefore, best suited for inference applications. However, the model supports only two labels, thus reducing the multi-label relationships. Slutsky et al. (2014) designed a new model named Hash-Based Stream LDA, a multi-label topic model, which is one of the most commonly used topic models in recent years. Wilcox et al. (2021) introduced a new supervised LDA with covariates (SLDAX) model, which integrates both a measurement (latent variable) model of text and a regression model to enable the hidden themes and other manifest variables to be used as predictors of results. The big issue of short text is that it suffers from the data sparsity in the feature vector, Pang et al. (2019) developed two supervised topic models, namely a Weighted Labeled Topic Model (WLTM) and X-term Emotion-Topic Model (XETM), to address this issue and to discover emotions toward specific topics.

4.2.4 Exemplar based topic model

Conventional models for detection topics concentrate on representing topics utilizing words, which are negatively impacted by Twitter's character limit and lack of context. Besides, one of the limitations is the scalability of the processing models required to handle the enormous volume of tweets created daily. Therefore, Exemplar-based techniques are suggested by Elbagoury et al. (2015) to fill this gap. An Exemplar-based technique selects the representative Exemplar tweets instead of the set of words to represent the extracted topics based on the variance of the similarity among other tweets and exemplars tweets. The proposed model mitigated the aforesaid issues and adapted for an easy interpretation and understanding of the meaning of retrieved topics.

Formally, let T be a collection of tweets with the size of m, and the main objective is to extract and detect the hidden topics from the T collection and represent every topic utilizing just one tweet (Exemplar). The criterion for selection should be able to identify a tweet for every topic so that every tweet describes a single topic and distinguishes it from other topics simultaneously. The criterion utilized in the proposed model is as follows: A tweet t_i which is similar to a collection of tweets and yet different from the other tweets, is an excellent theme (topic) representative. It can be expressed by formulating a similarity matrix $S_{m \times m}$ in where $S_{i,j}$ represents the similarity among tweets t_i and t_j. Consequently, the sample variance of its similarity distribution will be considerable. The formula of the sample variance for every tweet t_i can be calculated as in Eq. (6).

$$var\left(S_{:i}\right) = \frac{1}{m-1} \sum_{j=1}^{m} \left(S_{i,j} - \mu_i\right)^2 \tag{6}$$

where $S_{i,j}$ is the similarity between two tweets t_i and t_j, $\mu_i = \frac{1}{m} \sum_{j=1}^{m} S_{i,j}$ is the average of the similarities of the tweet t_i.

Similarly, other recent research used Exemplar based technique for detecting topics from tweets. To this extent, Shi et al. (2019b) used the exemplar approach for event detection from tweet topics with higher degrees of accuracy. Liu et al. (2020a) also employed an exemplar approach for event evolution strategy with high stability and purity. Although this model is highly efficient and seems to provide a better balance between the term-recall and term precision, it has limited control over the dynamically changing number of topic labels. The analysis of these models is furnished in Table 5.

4.2.5 Dynamic topic models

Static topic models were the most widely used models in many applications. But these models are limited to static vector modelling of tweets and thus unsuitable for representing dynamically changes of Twitter streams. This category includes traditional topic models such as LDA, LSA, and PLSA. On the other side, Dynamic Topic Modeling (DTM) generally pursues to detect the hidden topics from sequentially reached data blocks in order to catch the evolving trends of these topics. One of these models (Blei and Lafferty 2006) suggested DTM based on the assumption that all sequentially arranged short text or documents have the same concentrated themes with smooth variations. Continuous-Time Dynamic Topic Models (CDTM) was proposed by (Wang et al. 2008), which models latent topics through a successive set of documents by employing Brownian motion. The CDTM has several advantages, the most important of which is the ability to use sparse variational inference for rapid model comparison. These models DTM and CDTM were scaled by (Bhadury et al. 2016) on big data utilizing a form of Gibbs Sampling, which combines the Metropolis–Hastings sampler and Stochastic Gradient Langevin Dynamics. Saha and Sindhwani (2012) and Vaca et al. (2014) presented a dynamic NMF approach with temporal regularisation to understand and learn emerging and evolving topics in social media

Table 5 Comparative analysis for dynamic based, exemplar-based, application-based, word type based, and frequent pattern mining based models

Category	References	Models	Objective	Limitations	Datasets		Source	Domain
					Name	Lang		
Dynamic topic models	Blei and Lafferty (2006)	DTM	High adaptability and accuracy	High computations required scalability and sparsity issues	Corpus articles	EN	JSTOR	Specific (Science)
	Yao and Wang (2020)	DTM			Tweets (Disaster)	EN	Twitter	Specific (Disaster
	Liang et al. (2016)	DCT	cope with dynamicity, alleviate the data sparsity problem	Scalability issues	Tweets2011	EN	Twitter	Generic
	Ghoorchian and Sahlgren (2020)	GDTM	Alleviate data sparsity problem, overcome the scalability issues, account for dynamicity of DTM	Only applied over short texts	Collected Twitter Dataset	EN	Twitter	Specific
Exemplar based	Elbagoury et al. (2015)	Exemplar	High event detection accuracy, high stability and purity	Limited control over dynamically changing topic labels	Collected 2 Twitter D	EN	Twitter	Specific (Politics)
	Shi et al. (2019b)	Exemplar			Collected 4 small Twitter DS	EN	Twitter	Generic
Application based models	Wang and Iwaihara (2015)	Bilingual LDA	It supports two language tweets	Requires external sources	Collected Twitter DS	EN, JPN	Twitter	Specific (Comic)
	Pu et al. (2016)	Wiki-LDA	Better cohesion	It depends solely on Wikipedia articles	Collected Twitter DS	EN	Twitter	Generic
	Feng (2018)	ED-LDA	High accuracy for environmental topics	Lack of effective semantic structures	Collected Twitter DS	EN	Twitter	Generic

Table 5 (continued)

Category	References	Models	Objective	Limitations	Datasets		Source	Domain
					Name	Lang		
Word type techniques	Koike et al. (2013)	SOW	High efficiency than BOW	Less performance for complex structures	News, Twitter DS	EN, JPN	Newspaper Twitter	Specific
	Sasaki et al. (2014)	Twitter-TTM			Collected tweets	EN	Twitter	Generic
	Dey et al. (2018)	SOW-LSTM (T-PAN)	Detect user stance with respect to given topics on Twitter. It is easy to implement, robust, and reusable	High computation complexity	SemEval (Mohammad et al. 2016)	EN	Twitter	Specific
	Vargas-Calderón and Camargo (2019)	Word2vec	High correlated topics	Complexity in handling large data	Twitter DS	EN	Twitter	Generic
Frequent pattern mining techniques	Guo et al. (2012)	FPM	high stability and accuracy with less overhead	Required extensive training	Twitter DS	EN	Twitter	Specific Pandemic
	Peng et al. (2018a)	ET-TPM	High coherence		RawSet, SinaSet	CHI	Weibo	Generic
	Choi and Park (2019)	ET-HUPM			FA Cup Final (FA), Super Tuesday (ST), US Elections[a]	EN	Twitter	Specific

[a] https://www.socialsensor.eu/results/datasets/72-twitter-tdt-datasets/

networks in order to better capture freshly emerging and fading topics. Cotelo et al. (2014) used the Dynamic topic-related tweet retrieval approach to extract efficient topics. Liang et al. (2016) introduced a new Dynamic Clustering Topic (DCT) model, which is capable of tracking words over topics and the time-varying distributions of topics over documents. The data sparsity issue of the short text and the dynamic nature of topics across time was solved by this model, but the scalability issue was not solved. Yao and Wang (2020) also used DTM for tracking urban geo-topics from user tweets. Similarly, many studies developed DTM models such as Biterm Topic Modelling (BTM) (Cheng et al. 2014), Pseudo-document-based Topic Modelling (PTM) (Zuo et al. 2016a), etc. Finally, Ghoorchian and Sahlgren (2020) developed a Graph-based Dynamic Topic Modelling (GDTM) which combines a language representation technique and an incremental dimensionality reduction method with a graph partitioning method to solve dynamicity and scalability problems and utilized a rich language method to address a data sparsity problem.

4.2.6 Single and multi-source topic models

Topic models are generally based on single-source documents. Even the STTM models are mostly based on single-source data such as Twitter, Facebook, Weibo, etc. On the other hand, multi-source data-based topic modelling seems to be gradually increasing in recent years. Because it is highly beneficial to extract topics, sentiments and events from multiple sources instead of extracting them from single sources. The only possible limitation in employing these multiple source data is their inability to handle complex semantic relations. This section describes the single and multi-source-based topic models and analyzes them in Table 5. Hong et al. (2011) developed a time-dependent topic model for multiple text streams from Twitter and Yahoo news data and achieved high accuracy. Cao et al. (2017) employed a domain-aware LDA topic model to extract topics from multiple data sources, namely Twitter, news and PubMed. Gupta et al. (2019) described Multi-view and Multi-source data transfers from PubMed, AG news, and Twitter using predefined topics and word embedding.

4.2.7 Application-based models

Researchers developed application-based models by modifying the existing topic models for specialized applications of topic modelling. For example, Wang and Iwaihara (2015) designed the Bilingual LDA topic model for cross-lingual tweet recommendation by assuming the cross-lingual tweets as similar to linguistic Wikipedia articles. Pu et al. (2016) presented the Wiki-LDA model in which the tweets are merged as documents and the Wikipedia labels, which is applied to extract the topics. Feng (2018) presented the Environmental Data LDA(ED-LDA) specifically designed for environmental tweet datasets through probabilistic learning.

4.2.8 Word type models

Most topic models such as LDA, LSA and PLSI and their extensions use the Bag-Of-Words (BOW) approach for topic representation. Sequence-of-Words (SOW) approach is simpler than BOW, but it is used less frequently due to its unpopular representation in tweets. Table 5 presents the comparative analysis of word type models. Koike et al. (2013) developed the SOW based document representation model for time series topic detection from correlated news and Twitter. Sasaki et al. (2014) proposed Twitter Topic Tracking Model (Twitter-TTM), which used the sequence-of-words approach for the online trending topic model for Twitter. Further, Dey et al. (2018) also used a sequence-of-words approach with LSTM for topical stance detection. Word2vec is yet another representation model that is being used widely in Twitter topic modelling for detecting a specific group of user profiles. Vargas-Calderón and Camargo (2019) used word2vec and latent topic analysis for extracting topics and portraying of citizens. Similarly, word2vec is also used for sentiment analysis based on user categories.

4.2.9 Frequent pattern mining based models

Frequent Pattern Mining (FPM) was initially suggested in mining transactions with an aim to locate items that occurred simultaneously in the transactions. The FPM can also be used for Twitter topic modelling, as suggested in (Aiello et al. 2013). In the context of social media, the item refers to any word w included in the post or tweet (except punctuation token and stop words). The transaction refers to the post, and all the posts or tweets appear in the slot of time T_i are denoted by transaction sets. The frequency with which a particular set of words occurs in a certain time slot is referred to as its *support*, and any combination of words (itemset) that meets minimal support is named a *pattern*. This approach consists of two processing stages: Frequent Pattern detection(FP-detection) and Frequent Pattern ranking (FP-ranking), where the FP-detection is utilized to discover and detect the frequent patterns and FP-ranking is utilized to rank the patterns. This method uses the FP-growth algorithm, which consists of the following phases, to identify frequent patterns.

- Compute the frequency of every word and disregard those words that fall below a given threshold.
- Arrange the patterns based on their frequencies and co-occurrences.
- Construct association rules on the transaction set using the following form:$\{w_1, w_2\} \rightarrow p_i = \{w_3, w_4, w_5, \ldots.\} withsupport(p_i)$

Then, this method ranks the frequent patterns after identifying them and gives the top k frequent patterns as the discovered and extracted themes (topics). Guo et al. (2012) suggested an approach for extracting hot topics from Twitter streams with high stability, accuracy, and with lesser overheads using the Frequent Pattern stream mining (FP-stream) algorithm. The FP-Stream technique can get results that are sensitive to time; indeed, it has the ability to differentiate between new and old new transactions. Therefore, the differentiation between the old and new transactions is an important part of discovering the hot trending topics on Twitter. Kim et al. (2012) presented a probabilistic topic model using FPM and improved the coherence of topics. Other versions of FPM are have also been used in topic modelling. Peng et al. (2018a) developed Emerging Topic detection based on Emerging Pattern Mining (ET-EPM) model using High Utility Item-set Mining (HUIM) algorithm. Likewise, Choi and Park (2019) detected Emerging Topics in the Twitter stream using High Utility Pattern Mining (ET-HUPM). Since the FPM algorithms require an extensive learning process to extract hot topics on Twitter, they are not preferred in STTM. The summary and comparison of FPM techniques are provided in Table 5.

4.2.10 Hybrid topic modelling

Apart from the standard topic modelling techniques, many studies have designed hybrid models to utilize the benefits of multiple models. This section reviews the hybrid TSTTM models. To this extent, Huang et al. (2017) and Ge et al. (2019) presented a hybrid topic model by combining TF-IDF and LDA with high coherence and perplexity. Zhang et al. (2019) developed a hot topic detection model using deep learning and LDA from limited-words, noisy tweets. It integrates image data using deep learning with the short text information using LDA from twitter to match topic words using fuzzy matching. However, the news tweets are not effectively classified in this hybrid model due to a lack of feature training.

Zhang and Eick (2019) developed a topic model by combining LDA and density–contour-based Spatio-temporal clustering for event detection. Rashid et al. (2019a) designed the Fuzzy K-means Latent Semantic Analysis (FKLSA) model for trending medical tweet topics. The frequencies of the global and local terms are extracted by the BOW model, and PCA is utilized for dimension reduction. FKLSA handles the problem of redundancy effectively to extract medical topics from the tweet health dataset. Zhang and Zhang (2020) developed a model based on Long Short-Term Memory (LSTM) to discover new topics for incremental short text. The short text was transformed into a word vector using word embedding (word2vec) in the first stage. Then, two models were designed based on LSTM. Lastly, hierarchical clustering was utilized to get the number of new topics. Pornwattanavichai et al. (2020) presented a tweet recommendation system using hybrid topic modelling of supervised and unsupervised strategies. It combined LDA and matrix factorization-based neural networks for discovering topics and providing recommendations. Though these hybrid models are highly efficient, they are also prone to certain complexities problems in most scenarios, which prove to be an obstacle. Yi et al. (2020) developed a novel regularized NMF topic model for short texts called TRNMF. It combines the extended NMF and clustering mechanism by presenting topic regularization and document regularization, respectively, to mitigate the data sparsity issue in the short text. Shahbazi and Byun (2021) proposed a model to anticipate the topics and knowledge discovery that integrates deep learning such as Artificial Neural Network (ANN) and LSTM with topic modelling and machine learning. This model overcomes data sparsity, data limitation, and word relationship issues. Ha et al. (2019) presented an approach which combines dropout into many learning models to learn LDA. The purpose of dropout assists in preventing the overfitting of probabilistic topic models on noisy and short text. The Hybrid based topic models for short text are compared and summarized in Table 6, along with the respective merits and demerits of each method.

4.3 ASTTM models

This section presents the second part of the taxonomy, which mainly categorizes ASTTM models into four categories, namely Dirichlet Multinomial Mixture (DMM) based models, Global word co-occurrences based models, and Self-aggregation based models (Qiang et al. 2020), and Deep Learning Topic Modeling (DLTM) based models. Figure 6 shows the taxonomy of the ASTTM models.

Table 6 Comparative analysis for hybrid TSTTM of short text

References	Models	Objective	Limitations	Datasets		Source	Platform	Domain	Evaluation
				Name	Lang				
Huang et al. (2017)	T-LDA	It is superior to LDA at runtime. high coherence and perplexity, identified topics, Precision, F-Measure, and Recall rate	The topic density leads to lose the central topic and impacts the performance	Weibo	CHI	Weibo	Java	Generic	Perplexity, R, P, F-M
Ge et al. (2019)	TF-IDF and LDA	Alleviate the data sparsity issue caused by the length restriction on the microblog, high coherence	High complexity	Sina Weibo	CHI	Weibo	NA	Generic	R, P, F-M
Zhang et al. (2019)	deep learning and LDA	High accuracy and less time	News tweets are not classified effectively	Collected Tweets	EN	Twitter	NA	Specific	R, P, F-M
Rashid et al. (2019a)	FKLSA	Resolves redundancy problems, execution time is stable with various numbers of topics	Multiple iterations are required, which increases the computational cost	Tweets, Ohsumed, MuchMoe, Synthetic WSJ	EN	Twitter, others	Matlab	Specific (Medical Health news)	P, R, F-M, AC
Zhang and Zhang (2020)	LSTM-HC	Discover new topics from the incremental short text	accuracy of the outcome is not well enough	Reuters21578, Twitter	EN	News, Twitter	NA	Specific (Tweet Product Company	NMI, AC., R, F-Measure,

Table 6 (continued)

References	Models	Objective	Limitations	Datasets		Source	Platform	Domain	Evaluation
				Name	Lang				
Pornwattan-avichai et al. (2020)	LDA and matrix factorization-based neural network	High accuracy	prone to certain complexities problems in most scenarios	Tweets	EN	Twitter	NA	Generic	MAE, Prediction coverage
Yi et al. (2020)	TRNMF	Overcome the data sparsity, dealing large-scale increasing data and employs the real system	Needs more time for handling large data	TMNwes, Twitter, Snippet:	EN	Twitter, Website, News	C++/ Java	Specific	TC, P, R, AC, F-M
Shahbazi and Byun (2020)	RL-SeaNMF	Mitigate data sparsity problems and reduce repetitive information problem	NA	Tweets, Tag-MYNews, DBLP, Yahoo Ans., Goofle-News		Twitter, Yahoo, Website, others	python	Specific	TC, R, P, F-M
Shahbazi and Byun (2021)	Hybrid Model	Mitigate data sparsity, Overcome word relationship issues	NA						

TC topic coherence, *P* precision, *R* recall, F-M F-Measure, *AC* accuracy, *NA* not mention, *EN* English, *CHI* Chinese

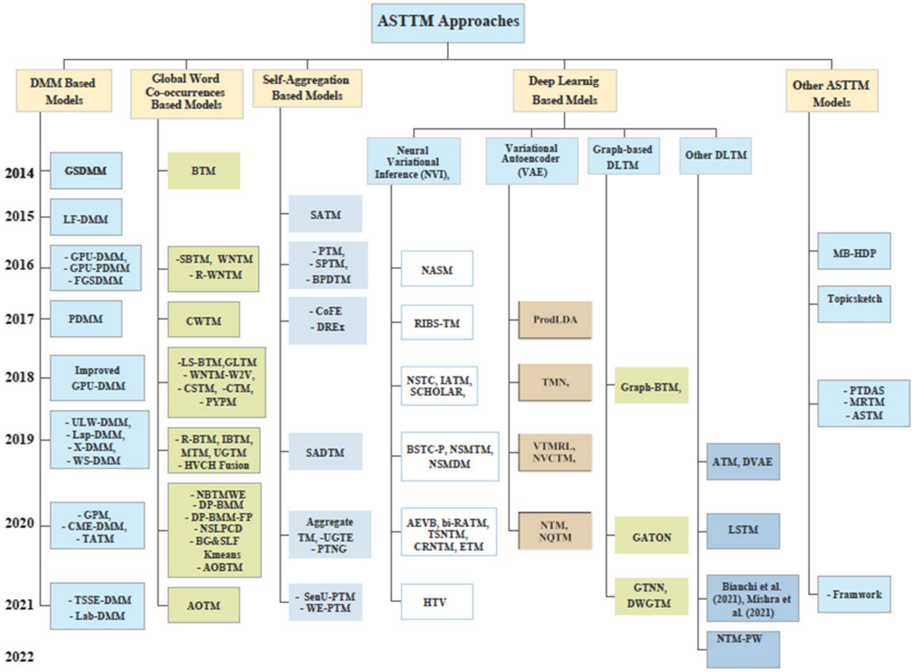


Fig. 6 Taxonomy of the ASTTM models

4.3.1 DMM based models

The DMM model was initially suggested by (Nigam et al. 2000), which has been applied to infer the hidden topics over short texts (Qiang et al. 2020). Nigam et al. (2000) suggested an Expectation–Maximization (EM) based method for DMM. Apart from the fundamental EM, various inference models such as Gibbs sampling and variation inference were utilized to estimate the parameters. It is based on a strategy of a simple assumption that one short text or tweet is sampled by only one hidden topic, which is much suited for the short texts in comparison with the more complicated assumption. This adopts the use of LDA in which every document or text is formed on a collection of topics (Quan et al. 2015; Yan et al. 2015). Some of the models are based on variational inference algorithms (Huang et al. 2013), such as the DMAFP model, which has been suggested by (Yu et al. 2010) and other models proposed by collapsed Gibbs sampling algorithm for DMM.

This sub-section presents the DMM based models which can discover and extract the Latent topics from a short text. Hence, many studies incorporating the DMM models for STTM followed. Yin and Wang (2014) proposed a Gibbs Sampling algorithm for Dirichlet Multinomial Mixture (GSDMM), which utilized DMM for short text topic clustering and achieved higher efficiency. Further, they introduced a Fast GSDMM (FGSDMM) (Yin and Wang 2016), which acclimatized an online clustering method for initialization. The time complexity of the GSDMM algorithm is $O\left(KN\bar{I}\right)$ where N denotes number of documents in the dataset, K is the number of pre-defined hidden topics, and $\bar{I}$ is the average length of each document in D. However, DMM has two key drawbacks when handling short texts. First, DMM supposes that each short text has only one topic, which is reasonable, not always true due to the users' approach to topic collection. This reduces its overall effectiveness. Likewise, the DMM does not possess the background knowledge of short texts. To address the first limitation, Li et al. (2017) designed an improved DMM model known as Poisson DMM (PDMM), which is based on modelling the topic number as the Poisson distribution with auxiliary word embedding. To resolve the second limitation, Li et al. (2016a) developed the Generalized Pólya Urn (GPU) model for semantic relatedness in the sampling process of DMM to develop GPU-PDMM and GPU-DMM. However, the promotion weight of the topics is fixed in this model. These models seem to outperform both the DMM and individual PDMM methods but also involve high computation costs. The time complexity of the GPU-DMM and GPU-PDMM algorithms are $O\left(KN\bar{I} + NI\zeta + KV\right)$ and $\left(\mathrm{N}\bar{I}\sum_{\mathrm{i}=1}^{\varsigma-1}\mathrm{C}_{\mathrm{K}}^{\mathrm{i}} + \mathrm{NI}\zeta + \mathrm{KV}\right)$, respectevely. where N denotes number of documents in the dataset, K is the number of pre-defined hidden topics, and $\bar{I}$ is the average length of each document in D. V denotes number of words in the vocabulary, ζ denotes the time and cost of considering GPU mode, ς is the naximum number of topics allowable in a short text, and c is size of sliding window. Zhang et al. (2018b) enhanced the GPU-DMM by including context information and word embedding to obtain the semantic similarities of the word pairs in order to improve topic coherence. However, this model also has limitations in terms of handling large collections of data. Mazarura et al. (2020) designed the Gamma-Poisson Mixture (GPM) topic model using an improved DMM concept and collapsed Gibbs sampling. It provided a high convergence and high coherence with better flexibility in topic extraction but offered limited performance on complex short texts. Nguyen et al. (2015) presented the Latent Feature vector based DMM known as LF-DMM by improving the feature word representations and incorporating word-topic mapping. However, LF-DMM increased noise interference due to the dependence only on external word expansions. The time complexity of the LapDMM algorithm is $O\left(O\left(2KN\bar{I} + KVU\right)\right)$ where N denotes number of documents in the dataset, K is the number of pre-defined hidden topics, and $\bar{I}$ is the average length of each document in D. V denotes number of words in the vocabulary, and U is the number of dimensions in word embeddings. Yu and Qiu (2019) developed ULW-DMM as an extension to DMM by combining DMM with user-LDA for potential words representation using feature vectors. ULW-DMM model increases topic coherence in short texts and reduces noise interference by considering both external and internal word expansion. Li et al. (2019c) developed the Laplacian DMM (Lap-DMM) topic model with Variational Manifold Regularization to improve the topic classification by 20%. However, it is based on the document similarities and hence, involves complexities.

Li et al. (2019a) developed a highly efficient text clustering model known as X-DMM utilized to reduce the complexity of sampling utilizing the Metropolis–Hastings method and presents an uncollapsed Gibbs sampler to parallelize the training model for scalable topic clustering. However, this model produced wrong mixtures in some cases due to contrasting sampling. Xiao et al. (2019) developed the Word Sense embedding based DMM (WS-DMM) for topic extraction and item recommendation through time-aware probabilistic modelling of user profile presence score. But this model suffers from the problem of error propagation that negatively impacts accuracy. Liu et al. (2020b) designed the Collaboratively Modelling and Embedding based DMM known as CME-DMM, incorporating topic and WE for extracting latent topics with high coherence. Although this model correlated different source data, the complexity of handling large data continues to be an issue. He et al. (2020a) developed Targeted Aspects-oriented Topic Modelling (TATM) using the

Enhanced DMM process (E-DMM) and different angle target aspect extraction. Though it overcomes the problems of standard DMM with efficient time management, it is incapable of handling complex structured short texts. Garcia and Berton (2021) introduced an effective method to explore a huge number of posts or tweets in both the USA and Brazil countries with the death and spreading by COVID-19, which uses a sentiment analysis and topic modelling. Li et al. (2021) suggested two models, namely Lab-DMM and OLabDMM, which can handle a set of massive short texts and try to alleviate the data sparsity problem. The time complexity of the LapDMM algorithm is $O\left(TDK\overline{N} + T\hat{T}DKR + D^2\right)$ where D represents the numbers of short texts, K denotes the topics, $\overline{N}$ represents the average document length of a corpus, T denote the numbers of outer iteration and $\hat{T}$ denotes the number of inner iterations in LapDMM. Mai et al. (2021) suggested a TM model called TSSE-DMM over short texts to improve the interpretability and coherence themes utilizing the topic subdivision and mitigating data sparsity problem utilizing the semantic improvement mechanism. Table 7 summarizes and analysis the existing DMM based ASTTM models qualitatively.

4.3.2 Global word co-occurrences based ASTTM models

Generally, there is insufficient word co-occurrence information in a short text. To cope with this issue, some models attempt to leverage the wealthy Global word co-occurrence patterns from the original dataset to infer hidden topics (themes) such as (Cheng et al. 2014; Zuo et al. 2016b). The short text sparsity problem is alleviated to some extent using the Global word co-occurrences due to its adequacy. These models require configuring a sliding window in order to extract word co-occurrences. Generally, if the length of each short text is greater than ten, they employ a sliding window and fix its size to 10; otherwise, If the length is less than 10, they can just treat the short text as a sliding window. This sort of model can be classified into two categories based on the utilization strategies of global word co-occurrences. The first one of the categories can infer the latent topics immediately by utilizing Global word co-occurrences. For example, the BTM model (Cheng et al. 2014) assumes that the two words comprising a biterm have the same topic, which is derived from various topics on the entire dataset. Whereas the second one of these categories, such as (Zuo, Zhao, et al. 2016b), creates a word co-occurrence network based on Global word co-occurrences and then figures out hidden topics from the constructed network, where the weight of each edge in the network represents the empirical likelihood of co-occurrence between the two connected words and each term or word represents a node of the constructed network.

This sub-section presents the STTM models based on Global word co-occurrences. The most important methods in this category are the Word-Network Topic Model (WNTM) and Biterm Topic Model (BTM). BTM is one of the most effective STTM. Cheng et al. (2014) first proposed BTM to extract topics from short texts by generating word co-occurrence patterns. BTM is easy to implement and can learn higher quality topics, and can offer better topic structure extraction. However, BTM can lose several possible prominent and coherent word co-occurrence patterns that can't be remarked in the corpus. It also suffers from noise and extraction of more irrelevant biterms. The time complexity of the BTM algorithm is $O\left(KN\bar{l}c\right)$, where N denotes number of documents in the dataset, c is size of sliding window, $\bar{l}$ is the average length of each document in D, and K is the number of pre-defined hidden topics. Pang et al. (2016) developed Sentimental BTM (SBTM) through the resemblance between words and documents with

Table 7 Comparative analysis of existing DMM based models

References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Yin and Wang (2014)	GSDMM	High coherence for short tweets with reduced sparsity problem	No background knowledge and assumption of single topic is not efficient	TweetSet[a] Google News	EN	Twitter, Others	Generi, Specific	Python	NMI, ARI, AMI
Nguyen et al. (2015)	LF-DMM	Better learning of latent features	Increased noise interference	Twitter[b] (Qiang et al. 2020), NG20 TagMyNews	EN	Twitter, Others	Specific	Java[c]	PMI, NMI, Purity, F-M, TC
Yin and Wang (2016)	FGSDMM	Low time and space complexity	Dirichlet process cannot capture the power-law phenomenon of the distribution of cluster	TweetSet NG20,[d] R52[e]	EN	Twitter, Others	Generic	Python	NMI, H, C
Li et al. (2017)	PDMM	Topic number determined based on auxiliary word embedding	Unable to handle complex structures	Snippet (Phan, Nguyen, and Horiguchi 2008), BaiduQA	EN, CHI	Popular Chinese Q, A website, Web search snippet	Generic	Java	TC, AC
Li et al. (2017)	GPU-DMM	Resolves background problem in DMM	Promotion weight of the topics is fixed	Snippet (Phan et al. 2008), BaiduQA	EN, CHI		Generic	Java	TC, AC
Li et al. (2017)	GPU-PDMM	Reduces the complexity of PDMM	High computation cost	Snippet (Phan et al. 2008), BaiduQA	EN, CHI		Generic	Java	TC, AC

Table 7 (continued)

References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Zhang et al. (2018b)	Improved GPU-DMM	Context information and word embedding for improved coherence	Unable to handle large data collections	Snippet (Phan et al. 2008), Amazon Reviews (McAuley and Leskovec 2013)	EN	Weibo, Amazon,Web search website	Generic, Specific(business)	NA	TC
Yu and Qiu (2019)	ULW-DMM	Increased coherence, less noise interference	Unable to handle large data collections	Sina Weibo	CHI	Weibo	Specific (News)	NA	TC, PMI
Li, Zhang, and Ouyang (2019)	Lap-DMM	Increased classification performance by 20%	The construction of document graphs consumes time when dealing with large and streaming data	Trec,[f] Snippets, StackOverFlow	EN	Others	Generic	C++	NMI, TC, AC
Li et al. (2019a)	X-DMM	Reduced time complexity	Wrong mixtures in outcome	NG20, ohsumed, QA,[g] Reuters	EN	Others	Generic	NA	Perplexity, NMI
Xiao et al. (2019)	WS-DMM	High coherence and less time	Error propagation problem	Sina Weibo (Zhang et al. 2017)	CHI	Weibo	Generic	NA	P

Table 7 (continued)

References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Mazarura et al. (2020)	GPM	High convergence and high coherence with better flexibility	Low performance on complex structures	TweetSet (Yin and Wang 2014),Pasca lflickr[h](Eve ringham et al. 2010), Search snippets (Phan et al. 2008)	EN	Twitter, Web search Snippets	Generic	Python	PMI, TC, AC, F-M
Liu et al. (2020b)	CME-DMM	High coherence even in different source data	Time-consuming for large data	Collected from Weibo, News collected by Sogou Lab	CHI	Weibo, Others Sogou Lab	Specific	Python	PMI, T C, R, P, F-M
He et al. (2020a)	TATM	Reduced time complexity	Unable to handle complex structures	TweetSet, NewsTitle	EN	Twitter, Others	Specific	Java	PMI, TC, AC, Purity
Mai et al. (2021)	TSSE-DMM:	Improve the interpretability and coherence themes and mitigate data sparsity problem	NA	Snippets, sogouCA	EN, CHI	Twitter, CHI website	Specific	Python[i]	NA
Li et al. (2021)	Lab-DMMT OLapDMM	Handles the massive short texts, and dealing with data sparsity problem	NA	Tweets, Trec, Snippets, StackOverFlow, BaiduQA	EN, CHI	Twitter, CHI website,	Generic	C++[j]	PMI,TC, Entropy

Table 7 (continued)

References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Garcia and Berton (2021)	GSDMM, sentiment Analysis	High accuracy for short tweets and alleviate the data parsity	Possible to miss tweets which did not include in the keywords of retrieving the content related to covid-19	Collected 2 Twitter Dataset	EN, POR	Twitter	Specific (Covid-19)	Python, Java	P, R

TC topic coherence, *P* precision, *R* recall, F-M F-Measure, *AC* accuracy, *H* Homogeneity, *C* completeness, *NA* not mention, *EN* English, *CHI* Chinese, *POR* portuguese

[a] https://github.com/jackyin12/GSDMM/

[b] http://www.sananalytics.com/lab/index.php

[c] https://github.com/datquocnguyen/LFTM

[d] http://qwone.com/~jason/20Newsgroups/

[e] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[f] http://cogcomp.cs.illinois.edu/Data/QA/QC/

[g] https://www.cs.cmu.edu/~ark/QA-data

[h] https://github.com/qiang2100/STTM

[i] https://github.com/PasaLab/TSSE

[j] https://github.com/li-ximing/LapDMM

sentimental relations to tackle the problem of irrelevance. Li et al. (2018a) developed Latent Semantic augmented BTM (LS-BTM) to include the latent semantic details for topic extraction and improve the performance of BTM. However, this model has its limitations due to the usage of more topic-irrelevant bi-terms. Li, A. Zhang, et al. (2019) developed Relational BTM (R-BTM) to overcome the loss of coherence of the BTM model through a similarity list of related words using word embedding. However, R-BTM does not consider other problems of BTM, such as meaningless biterms extraction and noise reduction.

Li et al. (2016b) designed the hybrid model of K-means clustering and BTM for micro-blog topic discovery with less noise. Xu et al. (2018) proposed a Chinese Topic Modelling (CTM) based on LDA and BTM to extract the topic distribution of short text from the corpus and the document itself. Zhu et al. (2019b) developed a joint topic model using Incremental BTM (IBTM) and extended LDA over streaming Chinese short texts. The BTM model tends to overlook the document-topic semantic data and dearth of the exact intents of users with sparsity problems. This joint model overcomes this limitation by using extended LDA for retaining document-topic information and IBTM to alleviate the issue of sparsity. However, in this model, some words fail to have semantic relevance and tend to reduce efficiency.

Wu and Li (2019) established the Multi-term Topic Model (MTM), which extracts the length of variable and multiple correlative word patterns from short texts to infer the latent trending topics. This model overcomes the limitations of BTM, such as extracting many irrelevant and useless bi-terms. However, MTM has smaller limitations of complexity due to separate handling the multi-terms and topics. WNTM is the second type of Global word co-occurrences based topic model, which utilizes Global word co-occurrence to build up word co-occurrence network and learns the distribution over topics using LDA. Zuo, Zhao, et al. (2016b) developed WNTM and used it for topic clustering from short and imbalanced texts. However, WNTM fails to express the deep meaning among words due to a lack of semantic distance measures. Further, WNTM has a lot of high irrelevant data in word-word space. The time complexity of the WNTM algorithm is $O\Big(KN\bar{I}c(c-1\Big)$, where K is the number of pre-defined hidden topics, N denotes number of documents in the dataset, c is size of sliding window, $\bar{I}$ is the average length of every document in D. Wang et al. (2016) developed the Robust WNTM (R-WNTM) as an extension for Short Texts. As the irrelevant data in the word-word space building procedure of WNTM is high, the R-WNTM that filters the unrelated data during the sampling process is presented. Jiang et al. (2018) presented WNTM with Word2Vector (WNTM-W2V) to extract deep meaning between words to enhance topic coherence as well as increases the accuracy of relationship among words.

Another word co-occurrence based model, namely, the Couple-Word Topic Model (CWTM) was presented by (Diao et al. 2017) to tackle the problem of data sparsity and incomplete description in topic extraction using couple word co-occurrences. CWTM is the first model to incorporate a couple of words, but it also suffers from difficulties in handling complex structures short texts and noisy microblogs data. Akhtar and Beg (2019a) developed the User Graph Topic Model (UGTM) by extending the author topic model through semantic relationships of contextual data. This method is highly efficient for topic extraction in a dynamic manner. Liang et al. (2018) designed the Global and Local word embedding-based TM (GLTM), which trains global embedding with a suitable encoding of continuous Skip-Gram model with Negative Sampling (SGNS) for getting local word embedding. This process enhances the semantic relatedness based topic discovery in short texts. Li, Wang, et al. (2018c) presented the Common Semantics Topic Model (CSTM) using unigrams for filtering noise in short text topic discovery. But, this model has limitations in terms of settings priority and determining the number of topic labels.

Chen et al. (2020a) developed two models, namely; Dirichlet Process Biterm-based Mixture Model (DP-BMM), which can alleviate the sparsity issue and handle the topic drift in the short text stream; the second method is an enhancing model of DP-BMM with forgetting property named (DP-BMM-FP) which removes the biterms of antiquated documents efficiently by eliminating clusters of antiquated batches. Moreover, Singh and Singh (2020) proposed a novel algorithm, namely Significance-based Label Propagation Community Detection (NSLPCD), which is capable of detecting and identifying topics promptly after happening from the Twitter dataset in a faster manner without compromising accuracy. Due to the issue of the data sparsity associated with short text and Twitter data, traditional topic discovery typically face difficulties of unintelligible and incoherent representation of topics. Thus, Liqing et al. (2019) presented a new model named (Hot topic detection based on the VSM Combined HMBTM) HVCH fusion to resolve this problem. Also, Hadi and Fard (2020) proposed an approach called Adoptive Online BTM (AOBTM) to solve the data sparsity issue and takes into account the statistical data for an optimal number of previous time-slices. The time complexity of the AOBTM model is $O\left(N_{itr}K\left|N^{(t)}\right| + vKW\right)$ Where N denotes the number of documents in the dataset, K is the number of pre-defined hidden topics, v denotes to the number of available time-slices, and W represents the total number of words. Moreover, a new model called Noise BTM Word Embedding (NBTMWE) was developed by (Huang et al. 2020) to tackle data sparsity. NBTMWE combines noise BTM and WE from external corpus to ameliorate the coherence of the topic. Wu et al. (2020a) invented a clustering algorithm for short texts based on the (BG & SLF–Kmeans) technique. This research proposed to discover the hot topics from short text microblogs. The pre-processed short texts were modelled using the BTM and GloVe approach. The similarity of the text based on the BTM vector was estimated using the JS divergence, and the similarity of the text based on the GloVe vector was estimated using the Improved Word Mover's Distance (IWMD). Lastly, the K-means clustering was realized with the use of the distance function obtained from the linear fusion of the two similarities. Yang and Wang (2021) presented propose a new TM called AOTM (Author co-Occurring Topic Model) for extracting the topics from short user comments and normal text. By taking authorship into account, AOTM provides each author of short text with a probability distribution over a collection of themes exemplified solely short texts. It explores clean user preferences and alleviates the sparsity of data. Table 8 presents the summary and analysis of the existing Global word co-occurrences based ASTTM models qualitatively based on their respective merits, limitations, the dataset used, and data sources.

4.3.3 Self-aggregation based ASTTM models

The self-aggregation based models are introduced to achieve topic modelling for the short text and automatically aggregate the short text during topic inference at the same time in one iteration. Such models merge short texts into long pseudo-documents to extract the hidden topics and assist in enhancing word co-occurrence information as well as to some extent addressing the problem of data sparsity. Current aggregation models, such as those (Weng et al. 2010; Mehrotra et al. 2013), and (Qiang et al. 2017), aggregate the short text and then apply the topic modelling. The new strategies, such as Pseudo-document-based

Table 8 Comparative analysis of Global word co-occurrences based models

References	Models	Objective	Limitations	Datasets		Platform	Domain	Sources	Evaluation metrics
				Name	Lang				
Cheng et al. (2014)	BTM	Simple to implement and better topic structure extraction	High noise and irrelevant biterm extraction	Question, Tweet2011,[a] Weibo	EN, CHI	C++[b]	Generic	Twitter, Weibo	TC, PMI, AC
Pang et al. (2016)	SBTM	Tackle the irrelevancy problem	Unable to handle complex structures	SemEval	EN	NA	Specific	News headlines	AC, Similarity
Huang et al. (2020)	NBTMWE	Alleviate data sparsity problems, Differentiate a noise topic from meaningful latent topics,	Similar biterms evaluated from embeddings can't be promoted to the same topics	Sina Weibo, Web Snippets	CHI, EN	Java[c]	Generic	Weibo, Web search snippets	TC, PMI, WESim
Li et al. (2018a)	LS-BTM	Latent semantics improve coherence	Topic-irrelevant bi-terms	Collected from Sogou Labs	CHI	NA	Specific	Sogou Labs	AC
Li et al. (2019b)	R-BTM	Overcomes the coherence loss	Noise interference and biterm irrelevancy	Tweets2011[19], StackOverflow from Kaggle.com	EN	Java, Python	Generic	Twitter	PMI, NMI, AC
Zhu et al. (2019b)	IBTM	High coherence in sparse tweets	Less semantic relevance	Sina Weibo BaiduQA	CHI	Java	Generic	Weibo CHI website	PMI, AC, P
Wu and Li (2019)	MTM	Resolves irrelevant biterms	Complexity due to multi-terms and topics	TweetSet StackOverflow[d]	EN	Python	Generic	Twitter, website	TC, PMI, Purity, AC

Table 8 (continued)

References	Models	Objective	Limitations	Datasets		Platform	Domain	Sources	Evaluation metrics
				Name	Lang				
Zuo, Zhao, et al. (2016b)	WNTM	High coherence in imbalanced texts	Cannot express the deep meaning between words	micro-blogs,[e] News corpus[f]	CHI, EN	Java,[g] MPI[i]	Specific	Weibo, News	TC, R, P, F-M
Liqing et al. (2019)	HVCH Fusion	Dealing with data sparsity problems and improving the quality of hot spots found in short texts	Not consider the typical real-time features of short text data	Collect:Tweets, SportsForum, microblogdatabas, microblogPCU	CHI	Python	Specific, Generic	Twitter	Perplexity, R, P, F-M
Wang et al. (2016)	R-WNTM	Solves irrelevant data problem	High sampling rate	Tweets (Zubiaga and Ji 2013), News[i]	EN	Java, C++	Specific, Generic	Twitter, News	TC, F-M
Jiang et al. (2018)	WNTM-W2V	Extracts deep meaning between words, increase the accuracy of the relationship among words	Topic irrelevancy problem	Tweets (Zubiaga and Ji 2013), News	EN	NA	Specific, Generic	Twitter	Ac, R, F-M
Qiang et al. (2018b)	PYPM	assumption of predefined true No. of clusters is solved	PYPM time complexity is linear to the number of active clusters	TweetSet[1] Google News[2]	EN	Python	Generic, Specific	Twitter, GoogleNews	NMI, ARI, AMI

Table 8 (continued)

References	Models	Objective	Limitations	Datasets		Platform	Domain	Sources	Evaluation metrics
				Name	Lang				
Diao et al. (2017)	CWTM	Tackle the sparsity and incomplete description	Difficult to handle complex and noisy data structures	BaiduQA, Web Snippets	CHI,EN	Java	Generic	CHI website, Web Snipet	R, P, F-M
Yang et al. (2018)	COTM	Capture co-occurring structure inherent in such text corpora		NetEase, Sina	CHI	C++,$^{j, k}$		News, Weibo	TC, AC
Hadi and Fard (2020)	AOBTM	Define the optimal No. of topics automatically, Mitigate the sparsity issue	NA	App Reviews (Gao et al. 2018), Tweets2020	EN	C++	Specific	Youtube, Viber,NOAA Radar, Swift-key	PMI, R, P, F-M
Liang et al. (2018)	GLTM	Enhances the semantic relatedness	Unable to handle complex structures	AmazonReview (McAuley and Leskovec 2013), Yahoo Answers (Chang et al. 2008), Web snippet (Phan et al. 2008) Tweet2011	EN	Java	Specific, Generic	Twitter, Amazon, Yahoo, Web snippets	TC, AC, F-M

Table 8 (continued)

References	Models	Objective	Limitations	Datasets		Platform	Domain	Sources	Evaluation metrics
				Name	Lang				
Chen et al. (2019)	KGNMF	A time-efficient algorithm, better in learning topics	Not applied for Long/normal Text	Snippet, Twitter StackOverFlow, TMNtitles, News	EN	Python	Generic	Sina website, News, Google, Twitter	
Chen et al. (2020a)	DP-BMM	Handle the sparsity and topic drift issues of the short text stream	NA	Google News, TweetSet, TweetSet-T	EN	Python	Specific, Generic	Twitter, News	NMI
Chen et al. (2020a)	DP-BMM-FP	Eliminate biterms of antiquated documents efficiently by deleting-eng clusters of antiquated batches							NMI
Singh and Singh (2020)	NSLPCD	Faster to detect topics, effective in run-time performance and quality	can't discover emerging the events of real-time from the feeds of Twitter timely	Collect dataset from Twitter	EN	NA	Specific	Twitter	NMI, R, P, RI, F-M
Wu et al. (2020a)	BG & SLF-k-means	Ameliorate the accuracy of hot topic modelling	Limited dataset	Sinna micro-blogs	EN	Python	Generic	Weibo	Perplexity, NMI, p, R, Purity, AC, F-M

Table 8 (continued)

References	Models	Objective	Limitations	Datasets		Platform	Domain	Sources	Evaluation metrics
				Name	Lang				
Yang and Wang (2021)	AOTM	Mitigate sparsity of data and explores clean user preferences	NA	News, e-commerce Dataset	EN	Python	Specific	Website	TC, Perplexity

TC topic coherence, *P* precision, *R* recall, *F-M* F-Measure, *AC* accuracy, *NA* not mention, *EN* English, *CHI* Chinese

[a] http://trec.nist.gov/data/tweets/

[b] Code of BTM: http://code.google.com/p/btm/

[c] https://github.com/Jenny-HJJ/NBTMWE

[d] https://www.kaggle.com/c/predict-closed-questions-on-stack-overflow/download/train.zip

[e] http://ipv6.nlsde.buaa.edu.cn/zhaojichang/paper/wntm.rar

[f] http://www.sogou.com/labs/dl/ca.html

[g] http://jgibblda.sourceforge.net/

[h] http://code.google.com/p/plda/

[i] http://acube.di.unipi.it/tmn-dataset

[j] https://github.com/dongxiexidian/hdLDA

[k] https://github.com/dongxiexidian/ohdLDA

Topic Modelling (PTM) (Zuo et al. 2016a) and Self-aggregation based Topic Modelling (SATM) (Quan et al. 2015) and etc., are different from the previous models, and they integrate clustering and topic modelling simultaneously in one iteration. The SATM and PTM have been considered the most commonly used models in this category.

This subsection introduces the latest Self-aggregation based models. To this extent, SATM, first proposed by (Quan et al. 2015), considers each short text as a sample from a hidden long pseudo-document and merges them to use Gibbs sampling for topic extraction without relying on metadata or auxiliary information. However, SATM is prone to over-fitting and is also computationally expensive. The time complexity of the SATM model is $O\left(N\bar{I}PK\right)$, where P denotes to the Long pseudo-document set generated by SATM model, K is the number of pre-defined hidden topics, N denotes number of documents in the dataset, $\bar{I}$ is the average length of each document in D. L. Shi et al. (2019a) enhanced SATM to develop a dynamic topic model that improves the efficiency of topic extraction. Blair et al. (2020) also designed aggregated topic models using cosine similarity and Jensen-Shannon divergence for increasing topic coherence. However, it does not seem to consider human perceptions and seems to trade-off between explicit and intrinsic features. The time complexity of Blair et al. (2020) algorithm is $O\left(N_{itr}, DK\bar{I}\right)$ where N_{itr} denotes the number of iterations, K is the number of topics, D is the number of documents, and the average number of words in a document, and $\bar{I}$ is the average number of words in a document.

In order to improve the performance of topic extraction when compared to that of SATM, PTM (Zuo et al. 2016a) presented the concept of the pseudo document to implicitly combine short texts to tackle data sparsity. In addition, Zuo et al. (2021) proposed a novel TM model named Word Embedding-enhanced PTM (WE-PTM) to leverage pre-trained WEs, which overcomes the data sparsity issue. The authors also introduced Sparsity-enhanced PTM (SPTM) by applying Spike and Slab prior for removing unwanted correlations among the pseudo documents. Although efficient, continuous research is needed to further improve its performance. The time complexity of the PTM algorithm is $O\left(N\bar{I}(P+K)\right)$, where K is the number of pre-defined hidden topics, N denotes the number of documents in the dataset, P denotes the Long pseudo-document set generated by the PTM model, $\bar{I}$ is the average length of each document in D. Jiang et al. (2016) developed Biterm Pseudo Document Topic Model (BPDTM) which is extended to BTM. Wandabwa et al. (2021) presented a method for learning the semantic relevance and importance of tweets. It determines a tweeter's degree of interest in a given topic based on the semantic relevance of the user's tweets. Bicalho et al. (2017) proposed Co-Frequency Expansion (CoFE) and Distributed Representation-based Expansion (DREx) to expand the short text into large pseudo-document models. Feng et al. (2020a) presented a User Group-based Topic-Emotion model (UGTE) for topic extraction and Emotion detection, which can alleviate the data sparsity problem by aggregating the short text of the group into long pseudo-documents. Most of the previous work took into account the data sparsity problem; in addition, they did not consider the sensitivity of word order in short texts. To address these issues, Lin et al. (2020a) developed a new topic model for short text named the Pseudo-document-based Topical N-Gram model (PTNG), which tackles the sparsity of the data in the short text as well as is sensitive to word order. Moreover, Lu et al.(2021) proposed a new model, namely Sense Unit based Phrase Topic Model (SenU-PTM), which alleviates the data sparsity problem and enhances the readability of the topics of short-text. The different models of self-aggregation based on ASTTM, along with their respective advantages and disadvantages are summarized in Table 9.

Table 9 Comparative analysis of self-aggregation based ASTTM models

	References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
					Name	Lang				
Self-aggregation based Models	Quan et al. (2015)	SATM	High accuracy	Over-fitting and computationally expensive	NIPS, Yahoo Answers	EN	Yahoo, others	Generic	C++	TC, PMI, Purity, AC
	Shi et al. (2019a)	SADTM	Resolves over-fitting problem	It does not consider human perceptions	Collect data from the Sina microblog	CHI	Twitter	Generic	NA	PMI, P, R, Purity, Entropy
	Zuo, Wu, et al. (2016a)	PTM	High coherence, alleviate data sparsity	Undesired correlations	News,[a]DBLP, Questions[b] (Yan, Guo, Lan, et al. 2013a),Tweets(Zubiaga and Ji 2013)	EN, CHI	Twitter, Websites	Specific	Java	P, R, F-M
	Zuo, Wu, et al. (2016a)	SPTM	Remove the undesired correlations	Still in early research stage		EN, CHI	Twitter, Websites	Specific	Java	P, R, F-M
	Li, Li, et al. (2018b)	LTM	Overcome s of Over-fitting and reduce computationally time	The new raised issue is bursty topic discovery	Snippets, Tweets, NIPS, Newsgroup	EN	Twitter, Websites, others	Generic	Python	TC, PMI, AC, RI Purity, RN
	Gao et al. (2019)	CRFTM	Mitigate the data sparsity	Not applying for tracing evolutions of topic in short text streams	News, StackOverflo		Others	Generic	Java, Python[c]	TC, AC, RT

Table 9 (continued)

References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Blair et al. (2020)	Aggrega-teTM	Increase the topic coherence	Doesn't consider to rank the topics based on	Tweets, Associated Press articles corpus[d]	EN	Twitter, Others	Specific(Amrican Debate)	NA	TC, PMI
Feng et al. (2020a)	UGTE	Alleviate the sparsity problem	Less capacity for modelling emotions and topics at the level of user group	ISEAR[e]	EN	Others	Specific	NA	TC, Kappa, Entropy
Lin et al. (2020a)	PTNG	Sensitive to word order, the sparsity issue in the short texts	Model Complexity is increase after introducing the bigram status of x (Lin et al. 2020a)	Tweets, DBLP, News	EN	Twitter, Newspaper websites	Specific, Generic	Java	TC, AC, R, P, F-M, Perplexity
Lu, Zhang, and Du (2021)	SenU-PTM	Tryinf to alleviate the sparsity and enhance readability problems of topic	NA	Used 2 real-world and publically datasets	EN		Generic	NA	

Table 9 (continued)

	References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
					Name	Lang				
	Zuo et al. (2021)	WE-PTM	Alleviate data sparsity	NA	Tweets, DBLP, News, Questions	EN, CHI	Twitter, Newspaper websites	Generic	Java, Python[f]	TC, P,R, F-M
	Liu et al. (2015)	MB-HDP	Effective burst topic detection	NA		NA	NA		NA	TC, Perplexity
	Xie et al. (2016)	Topicsketch	High similarity in topics	Low performance on complex structures	Two different Twitter datasets	EN	Twitter	Generic	Python, C++	TC, R, P
Other ASTTM models	Zhang et al. (2018a)	PTDAS	Improved coherence and accuracy	Does not consider human perceptions	Collected large-scale corpus from Weibo	CHI	Weibo, Twitter	Generic	Spark, HDFS	R, P, FM
	Liu et al. (2018)	MRTM	Human attentions improve coherence	Slow sampling speed and trade-off between relations	Collected 150,000 tweets from Weibo	CHI	Weibo	Specific	Python	TC, PMI
	Wang et al. (2018)	ASTM	Considering the data sparsity problem	Fixed similarity threshold	Oscars, Snippets(Phan et al. 2008), Title	EN	Twitter, Others	Generic	Java[g]	NPMI, AC

Table 9 (continued)

References	Models	Objective	Limitations	Dataset		Source	Domain	Platform	Evaluation metrics
				Name	Lang				
Wandabwa et al. (2021)	NA	Determines a tweeter's degree of interest in a given topic based on the semantic relevance of the user's tweets	Not address the generate multi-topic profiles from the user automatically	Colleted Twitter dataset, News	EN	Twitter	Generic	NA	AC, FMI-Score, Silhouette Score

TC topic coherence, *P* precision, *R* recall, *F-M* F-measure, *RI* rand index, *AC* accuracy, *RN* run time, *NA*:not mention, *EN* English, *CHI* Chinese

[a] http://acube.di.unipi.it/tmn-dataset/

[b] http://zhidao.baidu.com

[c] https://github.com/nonobody/CRFTM

[d] http://www.cs.columbia.edu/~blei/lda-c/

[e] https://www.kaggle.com/shrivastava/isears-dataset

[f] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[g] Code of ASTM using Java https://github.com/wjmzjx/ASTM

4.3.4 Deep Learning Topic Modelling (DLTM) models

Short text such as social media posts, product reviews, news headlines, etc., is becoming a more popular form of textual data. Unlike long texts from formal documents, messages on social media are generally short. However, automatic extraction of semantic topics from such short textual form is highly desirable in many NLP applications.. Traditional TMs such as LDA and PLSA have some limitations when handling social network data due to the limited word co-occurrence in each tweet or post. DL-based models appear to be viable for extracting valuable knowledge from such short text in complex systems. To this extent, various DLTM techniques for short texts are emerging to achieve flexibility and high performance. Recently, NTM is becoming a key DL tool for dealing with such short text. Driven by the desire to learn more coherent and semantic topics. The NTM approach has attracted much attention as it benefits from both neural networks and probabilistic TMs. The literature on TM has reported several models based on NTM; this section categorizes NTM into Neural Variational Inference (NVI), Variational Autoencoder (VAE), and Graph-based DLTM and then briefly reviews the related works to each category. Table 10 presents the comparative analysis of deep learning topic modelling based models.

4.3.4.1 Neural Variational Inference (NVI) based models Traditional probabilistic TMs are more likely in finding a closed-form solution to model parameters and also approach the intractable posteriors based on approximation methods. Eventually, these models result in inaccurate parameters inference and low efficiency when dealing with large-scale data. Recently, NVI emerged to solve such issues, which provides scalable and powerful deep generative models for modelling latent topics based on neural networks. Surprisingly, most neural variational TM makes the assumption that topics are independent and irrelevant to one another. This assumption, however, is unreasonable in many real-world scenarios. Typically, deep latent variable models have experienced an emergence as a result of recent advances in NVI. Miao et al. (2016) suggested a generic deep NVI approach for generative and conditional text models. The traditional variational techniques yield an analytic approximation for the intractable distributions over latent variables. Whereas an inference network conditioned on the discrete text input provides the variational distribution. This approach was evaluated over two variant text modelling applications: Supervised Question Answering and Generative Document Modelling. The neural variational document model combines a continuous stochastic document representation with a bag of words generative model and scored the lowest reported perplexities on two standard test corpora. The neural answer selection model utilizes a stochastic representation layer within an attention mechanism for semantics extraction.

Text analysis methods such as visualization and TM are widely used. Typically, traditional visualization methods search the visualization space for low-dimensional representations of documents. In contrast, TM aims at detecting topics from text, but for visualization, a post-hoc embedding utilizing dimensionality reduction techniques is required. Some NTM models employ a generative model for jointly discovering topics and visualization, including semantics in the visualization space for a better analysis. The scalability of their inference algorithms is a major barrier to their practical application. Pham and Le (2020) proposed Auto-Encoding Variational Bayes based inference method to jointly visualize and infer topics. Since the proposed model is a black box, it can effectively handle model changes with a little mathematical rederivation effort. Further, Pham and Le (2021)

Table 10 Comparative analysis of deep learning topic modelling based Models

References	Models	Objective	Datasets		Platform	Domain	Sources	Evaluation metrics
			Name	Lang				
Miao et al. (2016)	NASM	To model generative text	QASent and WikiQA, 20NewsGroups and the Reuters RCV1-v2	EN		Generic	TREC QA track, Wikipedia, qwone.com,trec.nist.gov	PPL
Pham and Le (2020)	AEVB	To jointly inferring topics and visualization	REUTERS, 20 NEWSGROUPS, WEB OF SCIENCE, ARXIV	EN	Python	Generic	mendeley, zhang18f.myweb.cs.uwindsor.ca	TC, NPMI
Pham and Le (2021)	HTV	To generate topic hierarchy and document visualization	BBC, Reuters, Newsgroups, Web of Science,	EN	python	Generic	mlg.ucd.ie, ana.cachopo.org, scikit-learn, mendeley	TC, Ac, NPMI, DS, Hierarchical Affinity, Running time
Lu et al. (2017)	RIBS-TM	To address the sparsity issue using word cooccurrence relationship	Online Questions corpus, Online News	CH	NA	Generic	ZhiHu, SogouLab	coherence
Li et al. (2020)	bi-RATM	To detect topics from sequential data	Subset of Wikipedia dataset, News articles (NYTimes)	EN		Generic	Wikipedia NYT	PPL, UCI UMass
Isonuma et al. (2020)	TSNTM	To distribute topics over an infinite tree	20NewsGroups, Amazon product reviews	EN CH	python	Specific	github.com Amazon	PPL, NPMI, TS, HAS
Peng, Xie, et al. (2018b)	NSTC	to enhance sparsity TMs and short text representation	Web Snippet and 20Newsgroups	EN	Python	Generic	qwone.com, jwebpro.sourceforge.net	PPL, AC
Peng et al. (2019)	BSTC-P	To impose hierarchical sparse prior for leveraging the prior information of relevance between sparse coefficients	20 Newsgroups, and Twitter dataset	EN	MATLAB	Generic	qwone.com, Twitter	PMI, AC
Miao et al. (2017)		To detect a notionally unbounded number of topics,	20NewsGroups, MXM song lyrics, Reuters RCV1-v24 news			Generic	qwone.com, trec.nist.gv, labrosa.ee.columbia.edu	PPL, NPMI

Table 10 (continued)

References	Models	Objective	Datasets		Platform	Domain	Sources	Evaluation metrics
			Name	Lang				
Lin et al. (2019)	NSMTM NSMDM	To identify latent topic sparsity	Twitter, New York Time articles (NYT), 20 Newsgroups	EN	NA	Generic	Twitter, qwone.com	PPL, PMI
Feng, Zhang, et al. (2020b)	CRNTM	Tackle feature sparsity problem	20NewsGroups, Snippets	EN	Python	Generic	Sourceforge Google	PPL, TC, AC,
He et al. (2018)	IATM	To model the user interactions in microblog	microblog corpus Li et al., (2016a)	CHI	NA	Generic	Weibo	PPL, TC
Card et al. (2018)	SCHOLAR	To allow for flexible metadata incorporation and rapid exploration of alternative models	Snowball, UIUC Yahoo answers dataset, New York Times articles, newsgroups Dataset	EN	Python	Generic	snowball.tartarus.org, cogcomp.org	NPMI, PPL, sparsity
Dieng et al. (2020)	ETM	To generate interpretable topics from large vocabularies	20Newsgroups, New York Times	EN		Generic	NA	TC, TD
Rezaee and Ferraro (2020)	NA	To handle the discrete variables	APNEWS, IMDB and BNC[a]	EN	Python	Generic		PPL, TSP
Srivastava and Sutton (2017)	ProdLDA	To address the component collapsing problem in AEVB	20 Newsgroups and RCV1 Volume 2	EN	Python	Generic	qwone.com,trec.nist.gov	TC, ELBO
Zeng et al. (2018)	TMN	To encode latent topic representations	Snippets, TagMyNews, Twitter, Weibo	EN, CHI	NA	Generic	Google, Twitter, Weibo	AC, Avg F
Gui et al. (2019)	VTMRL	To guide the learning of a VAE-based topic model using RL	20 Newsgroups and NIPS	EN	NA	Generic	qwone.com	PPL, TC
Bougteb et al. (2019)	NA	To detect eventual topics	20 Newsgroups,[b] Sanders[c]	EN	NA	Generic	Twitter, qwone.com	R, P, F-M
Liu et al. (2019)	NVCTM	To capture the topic correlations	20NewsGroups and Reuters RCV1-v2	EN	NA	Generic	qwone.com,trec.nist.gov	PPL, TC, R, P, F-M

Table 10 (continued)

References	Models	Objective	Datasets		Platform	Domain	Sources	Evaluation metrics
			Name	Lang				
Lin et al. (2020b)	NTM	To generate high quality topics from short texts	StackOverflow, Snippet, TagMyNews	EN	Python	Generic	Kaggle Google	PPL, NPMI, AC, TC
Wu et al. (2020b)	NQTM	To improve the quality and diversity of topics from short texts	StackOverflow, TagMyNews Title, Snippet, Yahoo Answer	EN	NA	Generic	Kaggle Google Yahoo	CV, TU
Xie et al. (2021)	GTNN	To extract the semantics of latent topics	Connected documents	EN	NA	Generic	NA	PMI, Acc, F1
Wang et al. (2021b)	DWGTM	To extract topics from concurrent word co-occurrence and semantic similarity graphs	Trec, Google-News, and YahooAnswer	EN	NA	Generic	cogcomp.csillinois.edu, Google, Yahoo	TC, TSC
Zhu et al. (2018)	Graph-BTM	To generate more coherent topics	20Newsgroups and All News	EN	NA	Specific	qwone.com Kaggle	TC
Yang et al. (2020)	GATON	To address the overfitting PLSI, capturing topical correlations, and high inference complexity	20NewsGroups and Reuters-215782	EN	NA	Generic	qwone.comdaviddlewis.com	TC, P, F1, R
Wang et al. (2019)	ATM	To extract semantic patterns from latent topics and generate word-level semantic representations	Grolier and NYtimes GDELT	EN		Generic	Grolier Multimedia Encyclopedia, NYT, Google	TC
Burkhardt and Kramer (2019b)	DVAE	To deal with trade-off between sparsity and smoothness	20news, NIPS, KOS, Rcv1^{d}	EN	python	Generic	UCI Machine Learning Repository, dailykos	PPL, TC, Topic Redundancy
Chuluunsaikhan et al. (2020)	LSTM	To predict the daily retail price of pork in the South Korean local markets	News articles and retail price data	EN	python	Specific	PigTimes, KAMIS, EKAPEPIA	RMSE, MAE, and MAPE

Table 10 (continued)

References	Models	Objective	Datasets		Platform	Domain	Sources	Evaluation metrics
			Name	Lang				
Bianchi et al. (2021)	NA	To generate more meaningful and coherent topics	20NewsGroups, Wiki20K, Tweets2011, StackOverflow	EN	Python	Generic	qwone.com Wikipedia Google	NPMI, TC, IRBO
Mishra et al. (2021)	NA	To detect hidden topics	Collected dataset related to tourism hospitality and healthcare	EN	python	Specific	Twitter	P, R, F1, Acc
Murakami and Chakraborty (2022)	NTM-PWE	To generate interpretable and coherent corpus-specific topics from short texts	BBC news, 20NewsGroup, SearchSnippets, TrecTweet, Biomedical, GoogleNews, M10, DBLP, PascalFlicker, StackOverflow	EN	NA	Generic	aclanthology.org arxiv.org	TC, TD

TC topic coherence, *P* precision, *R* recall, *F-M* F-measure, *AC* accuracy, *NA* not mention, *EN* English, *CHI* Chinese, *PPL* perplexity, *NPMI* normalized pointwise mutual information, *IRBO* inversed rank-biased overlap, *TSC* topical semantics coherence, *TD* topic diversity, *DS* document specialization, *HAS* hierarchical affinity scores, *TS* topic specialization, *RMSE* root mean squared error; *MAE* mean absolute error, *MAPE* mean absolute percentage error, *TSP* topic switch percent

[a] https://github.com/jhlau/topically-driven-language-model

[b] http://qwone.com/jason/20Newsgroups/

[c] https://github.com/zfz/twitter_corpus

[d] http://trec.nist.gov/data/reuters/reuters.html

proposed NTM approach for Hierarchical Topic detection and Visualization called (HTV) by jointly detecting topic hierarchy and generating document visualization with their topic structure. This helps in the quick detection of documents and significant topics with desirable granularity. To build an unbounded topic tree, they utilized a DRNN for generating topic embeddings. They also used KK layout objective function to regularize the model.

BTM proved its efficiency in addressing sparsity issues utilizes the word co-occurrence relationship. But, BTM and extended models ignore the internal relationship between words. Considering the relationship between words (Lu et al. 2017) suggested a short text TM named RIBS-TM using RNN for relationship learning and IDF for filtering high-frequency words. For document embedding, Li et al. (2020) suggested a bi-Directional Recurrent Attentional Topic Model (bi-RATM). Apart from using the sequential orders of sentences, the bi-RATM also uses the attention mechanism for modeling the relationships between successive sentences. Besides, they presented a bi-Directional Recurrent Attentional Bayesian Process (bi-RABP) for handling the sequences. Based on bi-RABP, the Bi-RATM, fully utilizes the bidirectional sequential information of sentences in a document. Furthermore, an online bi-RATM is suggested for handling large-scale corpora. Isonuma et al. (2020) suggested a tree-structured NTM distribute topics over a tree with an infinite number of branches. this model parameterizes an unbounded ancestral and fraternal topic distribution by employing the doubly-recurrent neural networks. It utilized autoencoding variational Bayes by which the data scalability and performance were enhanced when inducing latent topics and tree structures.

Peng, Xie, et al. (2018b) proposed a model called NSTC by incorporating the neural network and WE to enhance sparsity TMs. They replaced the complex inference process with the back propagation to reduce the computation complexity of TMs, while the WE external semantic information improves short text representation. Based on this work, Min Peng et al. (2019) presented Bayesian hierarchical TM known as Bayesian Sparse Topical Coding with Poisson Distribution (BSTC-P) to impose hierarchical sparse prior for leveraging the prior information of relevance between sparse coefficients. Moreover, Sparse Bayesian learning was proposed in this model to enhance the learning of sparse word, topic, and document representations. BSTC-P benefits from learning the word, topic, and document proportions. In the meantime, a sparsity improved version of BSTC is developed for obtaining the sparsest optimum solution using the Normal-Jeffrey hierarchical prior. For effective learning, the Expectation–Maximization and Variational Inference procedures are used.

Miao et al. (2017) proposed various NTMs based on Gaussian Soft max, Gaussian Stick-Breaking, and Recurrent Stick-Breaking constructions for parameterizing each document's latent multinomial topic distributions. With the assistance of the ground-breaking construction. For detecting latent topic sparsity while maintaining training stability and topic coherence. Lin et al. (2019) suggested two NTM based on the Gaussian sparse max (GSM) construction that provides sparse posterior distributions over topics and allows effective training via stochastic backpropagation. They built an inference network conditioned on the input data and use the Relaxed Wasserstein (RW) divergence to infer the variational distribution. Feng, Zhang, et al. (2020b) suggested a Context Reinforced NTM called (CRNTM). The proposed model infers the topic for each word in a narrow range, assuming that every short text covers a few salient topics. Then, in the embedding space, exploits pre-trained WE by simply treating topics as multivariate Gaussian distributions or Gaussian mixture distributions.

To model the user interactions in microblogs, He et al. (2018) suggested a topic modelling approach called Interaction-Aware Topic Model (IATM). This approach combines both user attention and network embedding. To mine dynamic user behaviours, they built a conversation network by connecting users using repost and reply relationships. They learn interaction ware edge embeddings with social context by modelling dynamic interactions and user attention, then incorporate them into neural variational inference for producing more consistent topics. Card et al. (2018) utilized a stochastic variational inference to introduce a general neural framework based on TMs for flexible metadata incorporation and rapid exploration of alternative models. This model was evaluated over a corpus of articles related to the immigration United States. It achieves high performance while balancing perplexity, coherence, and sparsity in a manageable way. Dieng et al. (2020) proposed an Embedded Topic Model (ETM) that utilizes LDA with WE for generating interpretable topics from large vocabularies containing rare and stop words. ETM models each word with a categorical distribution whose natural parameter is the internal product between the word's embedding and an embedding of its allotted topic. Further, they proposed an efficient Amortized Variational Inference (AVI) algorithm to fit the ETM. Rezaee and Ferraro (2020) used discrete random variables for NTM learning, explicitly modelling the assigned topic of each word based on NVI. To handle discrete variables, it does not rely on stochastic backpropagation. Using NVI, this model combines the expressive power of neural techniques for representing text sequences with the ability of TMs for capturing global, thematic coherence.

4.3.4.2 Variational Autoencoder (VAE) based models VAE enables generative TMs using neural networks. Recently, advances in NVI have resulted in effective text processing. For instance, NTMs are typically built on (VAE) with the goal of minimizing the error of reconstructing original documents based on the learned latent topic vectors. However, reducing reconstruction errors does not always result in high-quality topics. Srivastava and Sutton (2017) suggested an AEVB-based inference model for LDA called Auto encoded Variational Inference for Topic Model (AVITM). This model addresses the issues for AEVB caused by the Dirichlet prior and component collapsing. It matches the traditional methods in accuracy while taking much less time to infer. Indeed, the computational cost of running variational optimization on test data is unnecessary due to the inference network. Since AVITM is a black box, it can be applied easily to any TM. As an example, consider ProdLDA, a new proposed TM approach that replaces the mixture model in LDA with an expert product. By modifying just one line of code from LDA. Even when LDA is trained using collapsed Gibbs sampling, the ProdLDA produces far more interpretable topics.

Zeng et al. (2018) jointly investigate topic inference and short text classification by using Topic Memory Networks (TMN) for encoding latent topic representations indicative of class labels. They focus on extending features by using external knowledge or pre-trained topics. Gui et al. (2019) utilized Reinforcement Learning (RL) as reward signals with topic coherence measures for guiding the learning of a VAE-based TM. This enables the automatic separation of background words from topic words, obviating the need for the pre-processing stage of filtering infrequent and/or top frequent words, which is typically needed for learning traditional TMs. Bougteb et al. (2019) applied a deep autoencoder model with the Kmeans++ algorithm for detecting eventual topics in reconstructed data with less noise. Liu et al. (2019) presented a Centralized Transformation Flow for capturing topic correlations by reshaping topic distributions. Moreover, they proposed the Transformation Flow Lower Bound to enhance the proposed model's performance. Lin et al. (2020b) suggested NTM for short texts based on Auto-Encoding Variational Bayes. This model uses the Clayton Copulas for guiding the estimated topic distributions yield from linear projected samples of re-parameterized posterior distributions. Wu et al. (2020b) designed an NTM to produce a high-quality topic from short texts. This model utilizes a new topic distribution quantization technique for producing peakier distributions to model short texts. They further developed a negative sampling decoder for enhancing the diversity of short text topics.

4.3.4.3 Graph-based DLTM models LDA's failure to capture rich topical correlations between topics, and high inference complexity, to address Probabilistic Latent Semantic Indexing's overfitting problem Graph Neural Networks (GNNs) such as GCN help in learning the document representations efficiently by exploiting the semantic relation graph between documents and words. Despite a few exceptions, most of the previous works in this field do not take into account the underlying topical semantics inherited in document contents and the relation graph, making the representations less effective and difficult to interpret. Few recent studies attempt to apply latent topics to GNNs, where the topics are learned independently from the relation graph modelling. Xie et al. (2021) suggested a Graph Topic Neural Network (GTNN) model for extracting the semantics of latent topics for intelligible document representation learning, considering the word to word, document to word, and document to document relationships in the graph. To extract topics from concurrent word co-occurrence and semantic similarity graphs, Wang et al. (2021b) proposed the Dual Word Graph Topic Model (DWGTM). Where the global word cooccurrence graph is used for training DWGTM to learn word features. Then, it generates text features from word features and feeds them into an encoder network to obtain topic proportions per text; finally, it reforms texts and word co-occurrence graphs using topical distributions and word features, respectively. Furthermore, they used word features for rebuilding a word semantic similarity graph computed by pre-trained WEs in order to extract the semantics of words.

Zhu et al. (2018) suggested an approach for representing bi-terms as graphs known as Graph-BTM. They also designed a Graph Convolutional network along with residual connections for extracting transitive features from bi-terms. Moreover, for addressing LDA's data sparsity and BTM's strong assumption, they sample a fixed number of documents to create a mini-corpus as a training sample. Further, to generate more coherent topics, they proposed an AVI method for Graph-BTM. Yang et al. (2020) suggested an approach to address the overfitting issue of PLSI by applying the AVI with WE as input rather than the Dirichlet prior in LDA. To minimize the number of parameters, the AVI replaces the inference of the latent variable with a function that possesses the shared learnable parameters. The number of the shared parameters is fixed and independent of the scale of the corpus. The number of shared parameters is constant and independent of the corpus scale. They proposed a Graph Attention Topic Network (GATON) for modelling the topic structure of non-independent and identically distributed documents, overcoming the limitations of AVI's application to independent and identically distributed documents.

4.3.4.4 Other deep learning topic modelling based models Traditional TMs often need specific inference procedures for certain tasks. They are also not intended for producing word-level semantic representations. Wang et al. (2019) suggested an NTM model called the Adversarial-neural Topic Model (ATM) using the Generative Adversarial Nets (GANs). It models topics with Dirichlet prior and applies a generator network for extracting the semantic patterns from latent topics. Meanwhile, it utilizes a generator network for extracting semantic patterns from latent topics and models topics with Dirichlet prior. Furthermore, ATM was applied for open-domain event extraction to demonstrate the feasibility of the model for tasks other than TM. To enforce sparseness, LDA-based TMs typically use the Dirichlet distribution as a prior for the topic and word distributions. However, in Dirichlet distributions, there is a trade-off between sparsity and smoothness. Where, the sparsity is significant for low reconstruction error during autoencoder training and the smoothness allows for generalization and leads to a higher loglikelihood of the test data. These properties were encoded in the Dirichlet parameter vector by Burkhardt and Kramer (2019b). This parameter vector can be rewritten as a product of a sparse binary vector and a smoothness vector. This results in a model with both competitive topic coherence and a high log-likelihood. For the reparameterization of the Dirichlet distribution, rejection sampling variational inference allows for efficient training.

Chuluunsaikhan et al. (2020) developed an approach for predicting the daily retail price of pork in the South Korean local markets based on news articles by integrating DL with TM. They initially utilized TM for extracting relevant keywords for expressing price changes. Then, using these keywords, they built a prediction model based on statistical, ML, and DL methods. NTM has shown improvement in overall coherence, and the contextual embeddings have advanced the state of the art of neural models in general. Bianchi, Terragni, and Hovy (2021) integrated the NTM with contextualized representations to generate more meaningful and coherent topics. Mishra et al. (2021) utilized TM to detect hidden topics and identify the narrative and direction of tourism hospitality, and healthcare relevant topics. Furthermore, TM was used to detect inter-cluster similar terms and analyze the flow of information from a group of a similar viewpoint. Finally, a cutting-edge DL classification model was utilized with various epoch sizes of the dataset to predict and classify the people's feelings. Murakami and Chakraborty (2022) proposed a fine-tuning phase with an original corpus for training NTM for generating semantically coherent and corpus-specific topics. They used eight NTM to evaluate the effectiveness of the proposed additional fine-tuning phase and pre-trained WE in generating high-quality interpretable topics by simulation analysis over several datasets.

4.3.5 Other ASTTM models

Certain other techniques that were no way similar to the three categories of models described in the earlier paragraphs were also described. These methods used the benefits of multiple strategies without having similar to them. Liu et al. (2015) designed the Micro Blog Hierarchical Dirichlet Process (MB-HDP) topic model to tackle the problem of sparsity without a fixed number of topics as in the case of LDA. Xie et al. (2016) designed Topicsketch, which is a data sketch-based topic model to perform real-time burst topic detection. This model maps the tweet data as sketches and then extracts the topics from each sketch. Zhang et al. (2018a) developed the Pattern-based Topic Detection and Analysis System (PTDAS) to extract the trending topics from Chinese tweets through interesting cosine patterns. Liu et al. (2018) developed the Multiple Relational Topic Model (MRTM) by establishing document-attribute distribution and a two-step random sampling strategy to exploit both explicit and implicit relations. It improves both the coherence and accuracy of topic extraction but has limitations in terms of slow sampling speed and an unresolved trade-off between explicit and implicit relations. Wang et al. (2018) developed the Attention Segmentation based TM (ASTM) for short texts by integrating a human attention mechanism and word embedding as supplementary information to improve topic coherence. Although significantly efficient, this model depends entirely on a fixed similarity threshold, which might reduce the performance of segmentation. Zhu et al. (2019a) developed the Bayesian topic modelling for hierarchical topic viewpoints discovery from tweets based on the same depth tree formation.

5 Existing datasets

STTM models utilize so many datasets, most datasets are publically available, and the other datasets are collected and used specifically. This section reviews these datasets and provides a comparative analysis to analyze them based on the Number of Documents (ND), Vocabulary Size (VS), Labels/clusters (L), Average of the Document Length (AvgDL), utility, and Language (Lang) where (EN: English and, CHI: Chinese). We also provide the references of the publically available datasets to facilitate their accessibility to researchers. Table 11 shows a comparative analysis of the prominent datasets used by STTM models, whereas Table 12 summarises the datasets used by STTM models based on domain and availability. Here, the dominant datasets are briefly described below.

5.1 Tweetset dataset

This dataset contains 2472 tweets that are highly relevant to 89 queries. The relevance between queries and tweets is labelled manually in the 2011 and 2012 microblog tracks at the Text REtrieval Conference (TREC). The vocabulary size in this dataset is 5098 and the Average of the Document Length (AvgDL) is 8.56. They denoted this dataset as the TweetSet dataset. The other details of the TweetSet dataset are listed in Table 11.

5.2 Tweets dataset

A huge collection of tweets are gathered and labelled by (Zubiaga and Ji 2013). They scrape tweets, including URLs and classify them according to web page categories pointed by the URLs. The web pages categories are identified by the Open Directory Project (ODP). This dataset includes ten various groups and a total of around 360 k labelled tweets. They chose nine topic-related groups and sampled 182,671 tweets in the total under those categories. This Tweets dataset has been utilized in a few studies (Jiang et al. 2018; Wang et al. 2016; Zuo, Wu, et al. 2016a; Lin et al. 2020a; Indra and Pulungan 2019). Table 11 describes the details of this dataset.

5.3 Tweets2011 dataset

Tweets2011 dataset is a standard short text collection published in TREC 2011 microblog track, which includes approximately 16 million tweets sampled in the period from January 23rd to February 8th, 2011. Besides, each tweet contains a timestamp and user id. In this dataset, most of the studies selected the tweets randomly for the experiments. For instance, Liang et al. (2018) chose 3200 tweets for their experiments before performing the pre-processing stage and 30,946 after pre-processing, whereas Cheng et al. (2014) chose randomly part of this dataset, where the total number of tweets after pre-processing is 4230578. Moreover, Li et al. (2019b) obtained 5.42 Million tweets after pre-processing. Table 11 presents the details of these variations of the dataset.

Table 11 Comparative analysis of the prominent datasets used by STTM models

Dataset	ND	V.S	L	AvgDL	Lang	Utility	References
TweetSet	2472	5098	89	8.56	EN	(Yin and Wang 2014; Wu and Li 2019; Qiang et al. 2018b; Chen et al. 2020a; and Mazarura et al. 2020)	https://github.com/qiang2100/STTM, https://github.com/jackyin12/GSDMM
Tweets	182,671	21,480	9	8.5	EN	(Jiang et al. 2018; Wang et al. 2016; Zuo, Wu, et al. 2016a ; Lin et al. 2020a; Zuo et al. 2021)	(Zubiaga and Ji 2013) original paper
	316,924	7305	5			(Liqing et al. 2019)	
	20,000	13,621	1	8.26	EN	(Chen et al. 2020b)	
	34,554	733,861	10		EN	(Zhang et al. 2019)	
Tweets2011	4,230,578	98,857		5.21	EN	(Cheng et al. 2014; Liang et al. 2018)	http://trec.nist.gov/data/tweets/
	5.42 M	112 K	NA	5.05	EN	(Li, A. Zhang, et al. 2019b)	
	200 M	NA	NA	NA	EN	(Sharath et al. 2019)	
	2 M	121,788		NA	EN	(Rashid et al. 2019b)	
	4520	2502	NA	8.59	EN	(Yan et al. 2012; Yan, Guo, Liu, et al. 2013b)	
	49,461	30,421	50	5.92	EN	(Chen and Kao 2017)	
News	29,200	11,007	7	12.4	EN	(Jiang et al. 2018; Wang et al. 2016; Zuo, Wu, et al. 2016a ; Lin et al. 2020a; Zuo et al. 2021)	http://acube.di.unipi.it/tmn-dataset/
	32,600	6347	7	4.9	EN	(Nguyen et al. 2015; Yi et al. 2020; Gao et al. 2019; Chen et al. 2019)	
	10,193	5352	7	6.543	EN	(Wang et al. 2018)	
Google News	11,109	8110	152	6.23	EN	(Yin and Wang 2014; Qiang et al. 2018b; and J. Chen et al. 2020a)	https://github.com/qiang2100/STTM
Web Snipptes	12,340	30,445	8	17,5	EN	(Zhang et al. 2018b; Li, J. Zhang, et al. 2019; Akhtar et al. 2019a; Rashid et al. 2019b; and Li et al. 2021)	https://github.com/jacoxu/STC2 also available with 12,295 samples https://github.com/qiang2100/STTM,the original(Phan et al. 2008)
	12,265	5,581	8	10.72	EN	(Li et al. 2016a; Li et al. 2017; Huang et al. 2020; Yi et al. 2020; and Mai et al. 2021)	
	12,283	4,067	8	15.32	EN	(Chen et al. 2020b; and Chen et al. 2019)	
	1707	2,904	8	7.607		(Wang et al. 2018)	http://acube.di.unipi.it/tmn-dataset/ (Phan et al. 2008)

Table 11 (continued)

Dataset	ND	V.S	L	AvgDL	Lang	Utility	References
BaiduQA/ Question	189,080	26,565	35	3.94	CHI	(Cheng et al. 2014; Yan, Guo, Lan, et al. 2013a; Li et al. 2021)	http://zhidao.baidu.com
	179,042	26,560	35	4.11	CHI	(C. Li et al. 2016a; Zhu et al. 2019b; Phan et al. 2008; and Rashid et al. 2019b)	
	142,690	26,470	35	4.6	CHI	(Zuo, Wu, et al. 2016a ; Zuo et al. 2021)	
	36,219	4,956	34	5.8	CHI	(Yan, Guo, Liu, et al. 2013)	
Sina Weibo	688,738	203,886		11.12	CHI	(Yu and Qiu 2019; Zhu et al. 2019b; Xiao et al. 2019; and Liu et al. 2020b)	http://aminer.org/structinf
	15,561,7473	187,994	NA	5.87	CHI	(Cheng et al. 2014)	
	214,054				CHI	(Zheng et al. 2019)	
	11,176				CHI	(Han et al. 2020)	
	10,000		10		CHI	(Ge et al. 2019)	
	18,846	118,373	8	18.58	CHI	(Yang et al. 2018)	https://pan.baidu.com/s/1boVox3p
NG20	2000	181,754	20	137.85	EN	(Korshunova et al. 2019; Yin and Wang 2016; Nguyen et al. 2015; Li et al. 2019a; and Lacoste-Julien et al. 2009)	http://qwone.com/~jason/20Newsgroups/
	88,120		20		EN	(Abou-Of 2020)	
Yahoo answers	6310	5972	11	NA	EN	(Quan et al. 2015; Li et al. 2015)	https://answers.yahoo.com/
	19,980	15,776		117.4	EN	(Liang et al. 2018; Shahbazi and Byun 2020)	(Chang et al. 2008)
Amazon Reviews	337,559	NA	7	17.5	EN	(Zhang et al. 2018b; Liang et al. 2018)	(McAuley and Leskovec 2013)
	20,000	5574	50	3.39	EN	(Chen et al. 2020b)	https://www.cs.uic.edu/~zchen/downloads/ ICML2014-Chen-Dataset.zip

Table 11 (continued)

Dataset	ND	V.S	L	AvgDL	Lang	Utility	References
StackOverFlow	19,965	17,996	20	4.93	EN	(Li, J. Zhang, et al. 2019; Wu and Li 2019; X. Li, A. Zhang, et al. 2019b; Li et al. 2021; Gao et al. 2019; and Chen et al. 2019)	https://github.com/jacoxu/STC2, https://github.com/qiang2100/STTM
	55,290	12,087	20	5.35	EN	(Wu and Li 2019)	
DBLP	5966	7525	6	6.4	EN	(Lin et al. 2020a; Zuo, Wu, et al. 2016a ; Zuo et al. 2021)	
Reuters	9980		10		EN	(Zhang and Zhang 2020)	
	9094		82		EN	(Akhtar et al. 2019b)	
	5946	18,933			EN	(Farahat et al. 2015)	http://www.cad.zju.edu.cn/home/dengcai/Data/data.html
	9100	12,608	52	59.61	EN	(Yin and Wang 2016)	http://www.daviddlewis.com/resources/testcollections/reuters21578/
	11,476	39,600	116	102.734	EN	(Li et al. 2019a)	https://www.kaggle.com/nltkdata/reuters
SemEval (Mohammad et al. 2016)	1246				EN	(Pang et al. 2019, 2016; Dey et al. 2018; Mohammad et al. 2016; and Ozyurt and Akcayol 2021)	
Ohsumed	56,984	92,135	23	114.884	EN	(Karami et al. 2018; Rashid et al. 2019a; Li et al. 2019a)	http://davis.wpi.edu/xmdv/datasets/ohsumed.zip
Twitter	2520	1,390	4	5.0	EN	(Nguyen et al. 2015; Yi et al. 2020)	http://www.sananalytics.com/lab/index.php
Trec	5952	8392	6	4.94	EN	(Li, J. Zhang, et al. 2019; Li et al. 2021)	https://cogcomp.seas.upenn.edu/Data/QA/QC/
Pascal Flickr	4821	3188	20	4.9		(Everingham et al. 2010)	https://github.com/qiang2100/STTM
Oscars	5966	609		7.609	EN	(Wang et al. 2018)	https://www.kaggle.com/madhurinani/oscars-2017-tweets

5.4 BaiduQA/Questions dataset

The BaiduQA or Questions dataset contains 648,514 questions that were gathered from a well-known Chinese Q&A website. Each question is labelled into one of the 35 categories by its annotator. This dataset was prepared and utilized in (Cheng et al. 2014; Yan et al. 2013a; Li et al. 2021), which contains 189,080 questions for evaluating the proposed methods. In contrast, the dataset after pre-processing in the other studies, such as (Li et al. 2016a; and Zhu et al. 2019b), includes 179,042 questions. Another study presented by (Zuo et al. 2016a) used only 142,690 questions from the dataset for their experiments. The authors pre-processed the dataset, e.g., Chinese word segmentation was employed, and duplicate words were deleted. The detail of this dataset is presented in Table 11.

5.5 News dataset

This dataset is a collection of 29,200 English news articles gathered from RSS feeds of three famous newspaper news websites: reuters.com, usatoday.com, and nyt.com. It consists of seven clusters/categories: business, health, sport, sci&tech, U.S., world, and entertainment. The description of each news in the dataset is retained as standard short text, and its content is one or two simple sentences of this news article. This dataset is used for many STTM models, such as those (Jiang et al. 2018; Wang et al. 2016; Zuo, Wu, et al. 2016a; and (Lin et al. 2020a), whereas the other models (Nguyen et al. 2015; and Yi et al. 2020) used 32,600 English news and the detail is introduced in Table 11.

5.6 Google news dataset

The Google news dataset is one of the labelled collections utilized for evaluating the clustering performance. The news articles are classified automatically into topics/stories/ clusters. The authors took a snapshot of Google News on November 27th, 2013, and crawled the snippets and titles of 11,109 news articles as short text documents and grouped them into 152 clusters. The Google News dataset is split into three sub-datasets: SnippetSet (SSet), TitleSnippetSet (TSSet) and, TitleSet (TSet) as used in (Qiang et al. 2018b; Yin and Wang 2014; and J. Chen et al. 2020a). The TitleSnippetSet includes both snippets and titles, whereas TitleSet and SnippetSet datasets only include the titles and snippets, respectively.

5.7 Web snippets dataset

Web Snippet dataset contains 12,340 web search snippets. Each snippet belongs to one of eight groups. It has been utilized in many research like (Zhang et al. 2018b; Li et al. 2019c), which use all the samples of web search snippets for their experiments, While the other studies such as those (Li et al. 2016a; Li et al. 2017; Huang et al. 2020) use the dataset after pre-processing, which consists of 12,265 web search snippets. Moreover, one study proposed by (Wang et al. 2018) selected only 1707 snippets for evaluating their model. Table 11 shows the details of these datasets.

Table 12 Summary of the datasets used by STTM models based on domain and availability

Domain	Name/category	Lang	References	Availability
Social media	Collected from Twitter	EN	(Korshunova et al. 2019; Belford et al. 2016; Li et al. 2021; Kumar and Vardhan 2019; Ni et al. 2018; Yan et al. 2012; Fang et al. 2017; Fang et al. 2014; Muliawati and Murfi 2017; Prakoso et al. 2018; Pornwattanavichai et al. 2020; Capdevila et al. 2017; Mustakim et al. 2019; Zhang and Zhang 2020; Ali and Balakrishnan 2021; Yu et al. 2019; Abdulwahab et al. 2022; Yu et al. 2017; Indra and Pulungan 2019; Curiskis et al. 2020; He et al. 2020b; Fang et al. 2017; Karami et al. 2018 Wandabwa et al. 2021; Chen et al. 2019; Yang et al. 2019)	Specifically
	Collected from Weibo	CHI	(Ni et al. 2018; Yuan Zuo, Zhao, et al. 2016b; Liqing et al. 2019; Li et al. 2013; Zheng et al. 2019)	Specifically
	TweetSet	EN	(Yin and Wang 2014; Wu and Li 2019; Qiang et al. 2018b; J. Chen et al. 2020b; and Mazarura et al. 2020)	Publically
	Tweets	EN	(Jiang et al. 2018; Wang et al. 2016; Zuo, Wu, et al. 2016a ; Lin et al. 2020a; Zuo et al. 2021; Chen et al. 2020b; Zhang et al. 2019; and Liqing et al. 2019)	Publically
	Tweets2011	EN	(Cheng et al. 2014; Liang et al. 2018; Li et al. 2019b; Sharath et al. 2019; Rashid et al. 2019b; Y. Chen et al. 2020b; Yan et al. 2012; and Chen and Kao 2017; Yan, Guo, Liu, et al. 2013b)	Publically
	Twitter DataSet	EN	(Nguyen et al. 2015; and Yi et al. 2020)	Publically
	Sina Weibo	CHI	(Yu and Qiu 2019; Zhu et al. 2019b; Xiao et al. 2019; Liu et al. 2020b; Cheng et al. 2014; Zheng et al. 2019; Han et al. 2020; Li et al. 2013; Ge et al. 2019; and Yang et al. 2018)	Publically
	Youtube	EN	(Hadi and Fard 2020)	Publically
	Viber	EN	(Hadi and Fard 2020)	Publically
	MicroblogPCU	CHI	(Liqing et al. 2019)	Specifically

Table 12 (continued)

Domain	Name/category	Lang	References	Availability
News	News	EN	(Jiang et al. 2018; Wang et al. 2016; Zuo, Wu, et al. 2016a; Lin et al. 2020a; Zuo et al. 2021; Nguyen et al. 2015; Yi et al. 2020; and Wang et al. 2018)	Publically
	Google News	EN	(Yin and Wang 2014; Qiang et al. 2018b; and Chen et al. 2020a)	Publically
	NG20	EN	(Korshunova et al. 2019; Yin and Wang 2016; Nguyen et al. 2015; Li et al. 2019a; Lacoste-Julien et al. 2009; and Abou-Of 2020)	Publically
	SemEval (Mohammad et al. 2016)	EN	(Pang et al. 2019; Dey et al. 2018; Pang et al. 2016; and Mohammad et al. 2016)	Publically
	NewsTitle(4 dataset)	EN	(He et al. 2020a)	
	SemEval2016	TUR	(Ozyurt and Akcayol 2021)	Specifically
	Reuters	EN	(Akhtar et al. 2019b; Farahat et al. 2015; Zhang and Zhang 2020; Yin and Wang 2016; and Li et al. 2019a)	Publically
	Collected Sogou Labs	CHI	(Chen and Kao 2017; Yan et al. 2012; Liu et al. 2020b; Mai et al. 2021; Li et al. 2018a; and Yang et al. 2019; Yan, Guo, Liu, et al. 2013)	Specifically
Websites	Web Snipptes	EN	(Zhang et al. 2018b; Li et al. 2019c; Akhtar et al. 2019a; Rashid et al. 2019b; Li et al. 2016a, 2017, 2021; Huang et al. 2020; Yi et al. 2020; Mai et al. 2021; Chen et al. 2020b; and Wang et al. 2018)	Publically
	Wikipedia	EN	(Zuo, Zhao, et al. 2016b)	Specifically
	BaiduQA/ Question	CHI	(Cheng et al. 2014; Yan, Guo, Lan, et al. 2013; Li et al. 2021; C. Li et al. 2016a; Zhu et al. 2019b; Phan et al. 2008; Rashid et al. 2019b; Yan, Guo, Liu, et al. 2013; Zuo, Wu, et al. 2016a ; Zuo et al. 2021)	Publically

Table 12 (continued)

Domain	Name/category	Lang	References	Availability
Others	StackOverFlow	EN	(Li et al. 2019c; Wu and Li 2019; Li et al. 2019b; and Li et al. 2021) (Wu and Li 2019)	Publically
	Yahoo answers	EN	(Quan et al. 2015; Li et al. 2015; Liang et al. 2018; and Shahbazi and Byun 2020)	Publically
	DBLP	EN	(Lin et al. 2020a; Zuo, Wu, et al. 2016a ; Zuo et al. 2021; and Shahbazi and Byun 2020)	Publically
	AmazonReviews	EN	(X. Zhang et al. 2018b; Liang et al. 2018; Y. Chen et al. 2020b)	Publically
	ISEAR	EN	(Feng et al. 2020a; Pang et al. 2019)	Specifically
	Ohsumed	EN	(Karami et al. 2018; Rashid et al. 2019a; L. Li et al. 2019a)	Publically
	Trec	EN	(Li et al. 2019c; Li et al. 2021)	Publically
	Swiftkey	EN	(Hadi and Fard 2020)	Publically
	NOAA Radar	EN	(Hadi and Fard 2020)	Publically
	Sports Forum	CHI	(Liqing et al. 2019)	Specifically

5.8 DBLP dataset

The DBLP dataset consists of 55,290 short texts (titles of conference articles) from 6 research areas: NLP, Information Retrieval (IR), DataBase (DB), Data Mining (DM), computer vision, and Machine Learning (ML). The vocabulary size of this dataset is 7,525. Each conference title (short text) is labelled with one of the research areas. Zuo, Wu, et al. (2016a) and Lin et al. (2020a) used this dataset for their experiments to evaluate the proposed models. The detail of the DBLP dataset is provided in Table 11. The rest of the datasets used only one time for assessing the advanced STTM models are presented with their details in Table 12.

6 Tools and open-source library for topic modelling

This section presents the available open-source libraries and software tools for traditional and short text topic modelling. There are many tools and open-sources packages that can be used for both traditional and short text topic modelling.

6.1 Tools and open-source library of STTM

Several open sources and packages are available, especially for short text topic modelling. Table 13 shows the available tools of STTM. The prominent, well-known STTM tools are briefly described in this sub-section.

- **STTM** is an open-source java package developed by Qiang et al. (2020), which integrates most of the representatives of the SoTA short text topic modelling models such as LDA, DMM, LF-LDA, LF-DMM, GPU-PDMM, GPU-DMM, SATM, PTM, WNTM, and BTM. The STTM package is intended to facilitate the extension of new models in this work area as well as the accessibility of comparisons between new models and the SoTA models in the field.
- **jLDADMM** is an open-source Java toolkit proposed for traditional topic modelling DMM (Nigam et al. 2000) and LDA (Blei et al. 2003) utilizing collapsed Gibbs sampling. jLDADMM package has been developed by Nguyen et al. (2018) to provide different ways for topic modelling over short or long texts.
- **LFTM** is an open-source Java package suggested by Nguyen et al. (2015) to provide different ways for extracting latent topics over short or long texts. It includes two models, LF-DMM and LF-LDA.
- **BTM** and OnlineBTM are open-source c++ packages developed by Cheng et al. (2014) based on word co-occurrence that learns hidden topics using word-word co-occurrences patterns (biterms). OnlineBTM includes two models: OBTM and iBTM.
- **CRFTM** is an open-source java and python tool developed by Gao et al. (2019) and used to extract hidden topics from short text data. It alleviates the data sparsity issue by aggregating the short text data into pseudo-documents.
- **Palmetto** is an open-source tool and is publicly available in the project Palmetto. Palmetto was developed for topic modelling by (Röder, Both, and Hinneburg 2015) and used to measure topics' quality based on coherence computations on an external corpus. It carries out the six well-known topic coherences: UMass (Mimno et al. 2011), UCI (Newman et al. 2010), C_V and C_P (Röder et al. 2015), C_A and NPMI (Aletras

and Stevenson 2013). These topic coherences metrics are determined using co-occurrences of words in the English Wikipedia corpus and have been shown to be correlated with human ratings.

6.2 Tools and open-source library for traditional topic modelling

This sub-section presents the software tools and open-source library that can be utilized for the traditional topic modelling. Table 14 shows the available tools for traditional topic modelling.

- *MALLET* is a Java-based package released in 2002 by (McCallum 2002). It is a TM tool used for document clustering, classification, information extraction, topic modelling, statistical NLP, and other ML application to text. MALLET TM contains various models to discover and detect the latent topics from a dataset, including hierarchical LDA and Pachinko Allocation Model (PAM).
- *Stanford TMT* developed by (Ramage et al. 2009) and released for the first time in Sep 2009. It was conducted by the Stanford NLP group. TMT contains several TM models, including latent Dirichlet allocation (PLDA), labelled LDA, and LDA (Blei et al. 2003).
- *Mr. LDA* is a Java topic modelling package in the MapReduce framework developed by (Zhai et al. 2012). Mr. LDA utilizes the Variational Bayesian inference, and it has two advantages compared to Gibbs sampling based models: (1) Topic discovery is guided by informed priors. (2) Discovering the latent topics from the multilingual dataset.
- *JGibbLDA* is an LDA implementation developed in Java language by (Phan and Nguyen 2006) that uses the Gibbs Sampling method for inference and parameter estimation. It is helpful for many areas such as Retrieval of Information (inferring latent topic and analyzing semantics), text Summarization, text Clustering, text Classification, and generally text data mining.
- *JGibbLDA++* is a C/C++ free software of LDA proposed in 2006 by (Phan and Nguyen 2007) utilizing the Gibbs Sampling. It is a fast processing algorithm. GibbsLDA++ is well-suited for analyzing and extracting the hidden topics structures in large text datasets.
- *Gensim* is an open-source topic modelling toolkit developed by (Řehůřek and Sojka 2011) written in python language that can leverage large-scale unstructured texts to discover and detect the latent topics from the datasets by utilizing an efficient model. Gensim has a lot of different models like LSI, LDA, SVD, hierarchical Dirichlet processes (HDPs), TF-IDF, and LSA. It is faster and more scalable than the MALLET topic modelling tool.
- *TopicXP* is an open-source package (Eclipse plugin) written in java language developed by (Savage et al. 2010) that utilizes LDA for extracting latent topics from the natural language utilized in comments and identifiers of source code as well as visualizing the discovered topics for the users.
- *Keyphrase extraction algorithm (KEA)* is an open-source distributed tool written in Java language developed by Medelyan et al. it is utilized for extracting the keyphrase from the whole text of the dataset. KEA can be used for both free and controlled vocabulary indexing in a supervised manner.

Table 13 Tools and Open-source library of STTM

Reference	Tools/open-source	Programming language	Inference/parameter	Source code reference	Build-in models
Qiang et al. (2020) Qiang, Li, Yuan, Liu, et al. (2018a)	STTM	Java	Gibbs sampling	https://github.com/qiang2100/STTM	LDA, DMM, LF-LDA, LF-DMM, GPU-PDMM, GPU-DMM, SATM, PTM, WNTM, and BTM
Nguyen (2018)	jLDADMM	Java	Collapsed Gibbs sampling	https://github.com/datquocnguyen/jLDADMM	LDA, DMM
Nguyen et al. (2015)	LFTM	Java	Gibbs sampling	https://github.com/datquocnguyen/LFTM	LF-LDA, LF-DMM
Zhao et al. (2011)	Twitter-LDA	Java	Gibbs sampling	https://github.com/minghui/Twitter-LDA	Twitter-LDA
Rubin et al. (2012)	DependencyLDA	MATLAB (and C)		https://github.com/timothyrubin/DependencyLDA	Dependency-LDA, Prior-LDA and Flat-LDA
Mai et al. (2021)	TSSE-DMM	Python		https://github.com/PasaLab/TSSE	TSSE-DMM
Li et al. (2021)	LapDMM	C++		https://github.com/li-ximing/LapDMM	LapDMM
Yin and Wang (2014)	GSDMM	Python	Collapsed Gibbs Sampling	https://github.com/jackyin12/GSDMM	GSDMM
Huang et al. (2020)	NBTMWE	Java	Collapse Gibbs sSmpling	https://github.com/Jenny-HJJ/NBTMWE	NBTMWE
Cheng et al. (2014)	BTM	C++	Gibbs sampling	https://github.com/xiaohuiyan/BTM https://github.com/xiaohuiyan/OnlineBTM	BTM, OBTM, IBTM,
Yan et al. (2015)	BurstyBTM	C++	Gibbs sampling	https://github.com/xiaohuiyan/BurstyBTM	BurstyBTM
Gao et al. (2019)	CRFTM	java	Gibbs sampling	https://github.com/nonobody/CRFTM	CRFTM

Table 13 (continued)

Reference	Tools/open-source	Programming language	Inference/parameter	Source code reference	Build-in models
Wang et al. (2018)	ASTM	java		https://github.com/wjmzjx/ASTM https://github.com/wjmzjx/ASTM	ASTM
Li et al. (2016a)	GPUDMM	Java	Gibbs sampling	https://github.com/NobodyWHU/GPUDMM	-
Miao et al. (2016)	NVDM	Python	variational inference	https://github.com/ysmiao/nvdm	-
Srivastava and Sutton (2017)	ProdLDA	Python		https://github.com/akashgit/autoencoding_vi_for_topic_models	-
Zhu et al. (2018)	GraphBTM	Python	Amortized variational inference	https://github.com/valdersoul/GraphBTM	-
Wu et al. (2020b)	NQTM	Python		https://github.com/BobXWu/NQTM	-
Pham and Le (2020)	PLSV-VAE	Python	variational auto-encoder (VAE) inference	https://github.com/dangpnh2/plsv_vae	
Pham and Le (2021)	HTV	Python		https://github.com/dangpnh2/htv	
Röder et al. (2015)	Project Palmetto	Java	Tools for evaluation TM	https://github.com/dice-group/Palmetto	Topic coherence

- *Yahoo_LDA* is an open-source implementation written in C++ language and developed by Ahmed et al. (Ahmed et al. 2012) for their proposed framework architecture. The source code can be found at https://github.com/shravanmn/Yahoo_LDA.
- *R Language* has several libraries and packages for effective topic modelling, such as (1) LSAfun is a standard package in R consisting of a set of functions developed by (Günther et al. 2014).
- *The Structural Topic Model (STM)* is an R package developed by (Roberts et al. 2019) for the structural topic model. The STM provides several characteristics such as assessing uncertainty, ways to explore the topics and visualizing the discovered topics.

7 Quantitative analysis of the literature

This section quantitatively analyzes the literature of STTM models. It shows the percentage of publications based on main categories, publication year, sub-categories, platform, and evaluation metrics. Moreover, it quantitatively analyses the existing datasets based on utility, language, and source, answering the following research questions?

- How many recent research articles were yearly published in each STTM category?
- Which categories of the STTM models are studied the most and the least?
- What does a distribution of the papers look like?
- What are the prominent evaluation metrics in the literature on STTM that are used the most and the least?
- What are the most used existing datasets by STTM?
- What are the prominent sources of existing datasets?
- What are the languages of existing datasets?
- What is the implementation platform used by the existing STTM models?

For answering the aforesaid research questions, the reviewed research articles in this taxonomy were quantitatively analyzed based on their main category. Then each main category was quantitatively analyzed based on the publication year and their sub-categories, usage of programing language in current models, and evaluation metrics. Finally, the utilized existing datasets for evaluating the STTM models were quantitatively analyzed based on language, sources, and their names.

7.1 STTM papers based on (TSTTM and ASTTM) main categories

The total number of surveyed articles, including surveys papers in this taxonomy, is 231 articles, out of which 51.52% (119) research articles are related to TSTTM models main category, and 41.13% (95) research papers are related to ASTTM models, whereas the rest 7.36% (17) articles are related to the STTM surveys, this is clearly depicted in Fig. 7.

Table 14 Tools and open-source library for traditional Topic modelling

References	Tools/open-source	Programming language	Inference/parameter	Source code reference	Build-in models
Ramage et al. (2009)	Stanford TMT	Java	Gibbs sampling	https://nlp.stanford.edu/software/tmt/tmt-0.4/	LDA, Labelled LDA and PLDA
Zhai et al. (2012)	Mr.LDA	Java	Variational Bayesian inference	https://github.com/lintool/Mr.LDA	Mr.LDA
(Phan and Nguyen (2007)	JGibbLDA ++	C ++	Gibbs sampling	http://gibbslda.sourceforge.net/	LDA
Phan and Nguyen (2006)	JGibbLDA	Java	Gibbs sampling	http://jgibblda.sourceforge.net/	LDA
Řehůřek and Sojka (2011)	Gensim	Python	Gibbs sampling	https://radimrehurek.com/gensim	LDA, RP, LSA, TF-IDF, HDPs, LSI, and SVD
Savage et al. (2010)	TopicXP	Java(Eclipse plugin)		http://www.cs.wm.edu/semeru/TopicXP/	LDA
Ahmed et al. (2012)	Yahoo LDA	C ++	Gibbsampling	https://github.com/shravanmn/Yahoo_LDA	LDA
Zhao et al. (2011)	Lda in R	R	Gibbsampling	https://cran.r-project.org/web/packages/lda/	LDA, sLDA, corrLDA
	KEA	Java	Gibbs sampling	http://community.nzdl.org/kea/download.html	
Günther et al. (2014)	LSAfun	R		http://CRAN.R-project.org/ package = LSAfun	LSA
Roberts et al. (2019)	STM	R		https://CRAN.R-project.org/package=stm	LDA, CTM
Dieng et al. (2020)	ETM	Python		https://doi.org/10.1162/tacl_a_00325	NA
Isonuma et al. (2020)	TSNTM	python	Nested Chinese restaurant process (nCRP) with collapsed Gibbs sampling	https://github.com/misonuma/tsntm	NA

7.2 STTM publications based on Publication year

This sub-section presents the quantitative analysis of both TSTTM and ASTTM publications based on the publication year.

7.2.1 TSTTM publications

This paper has surveyed 119 published research articles (51.52%) related to TSTTM publications which have been published from ≤ 2011 to 2022. This part of taxonomy is classified into twelve classes, and each publication year represents a class, starting from 2011 till 2022, except the first class represents the publication earlier than 2012, named as ≤ 2011. Figure 8a illustrates the distribution of research publications of the TSTTM by publication year. It can be observed that from Fig. 8a, 12.61% of the papers of TSTTM models were published in the years before 2012. In recent years, researchers have obviously given greater attention to STTM like Twitter, etc. it can be clearly observed that 24.37% of the papers were published in 2019, which is the most productive year for publications in TSTTM models that's mean it is one-fourth of the publications. Moreover, it is clearly noted that the percentage of publication papers increased continuously from 2016 to 2019, and the rate was 4.2%, 5.04%, 10.08%, 11.76%, and 24.37%, respectively. Whereas the year of 2020 got a 11.76% rate, the rate in 2021 decreases to 4.2%. The rate of 2022 is 0.84%. Here, we can't decide the final rate due to it does not complete.

7.2.2 ASTTM publications

The taxonomy surveyed 95 publication articles (41.13%) related to ASTTM publications which have been published from 2014 to 2021. ASTTM publications are categorised into eight classes, starting from 2014 to 2022. It can be obviously noted that from Fig. 8b, 22.34% of the papers were published in 2020, which is the highest rate among all the classes in ASTTM publications. Moreover, it is clearly observed that a continuous increase in the number of published papers from 2014 to 2020, except in 2017, the number of publications is more significantly decreased. In the 2021 class, the percentage of publications rate is 13.83%, which is considered reasonable. The rate of publication in the year 2022 is 1.06%; however, the year has not been completed until now.

From Fig. 8a and b, we can conclude that the higher rate of publications was in the 2019 class according to the TSTTM publications, whereas the 2020 class is considered the most productive year for publications in ASTTM publications.

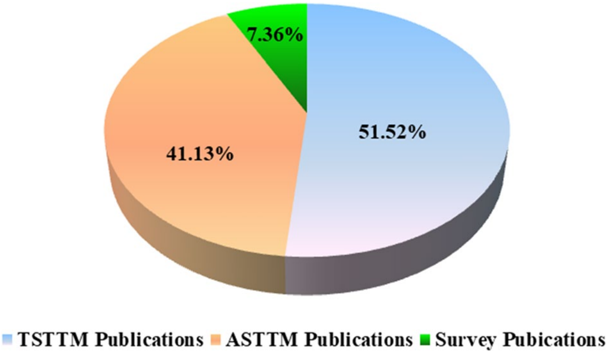
7.3 STTM publications based on sub-categories

This section presents the quantitative analysis of both TSTTM and ASTTM publications based on the sub-Categories.

7.3.1 TSTTM publications

This part of the taxonomy paper has categorized the TSTTM models into eleven sub-categories: probabilistic models, matrix factorization, unsupervised, supervised, dynamic-based, exemplar-based, data source based, word types, application-based, Frequent Pattern

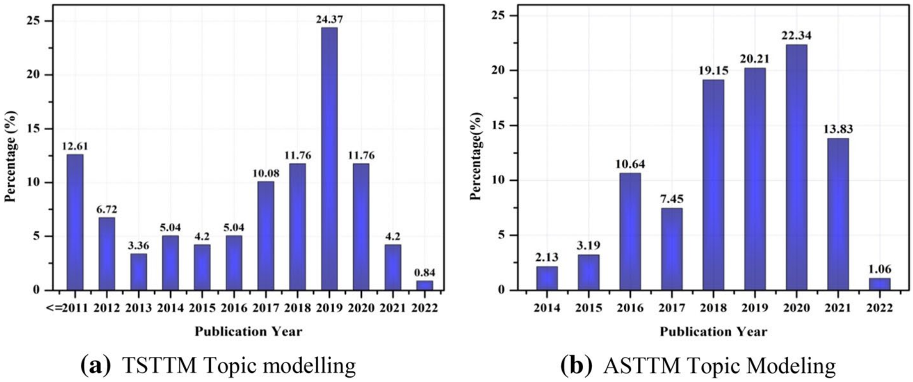
Fig. 7 Distribution of TSTTM, ASTTM and surveys publications



Mining (FPM), and Hybrid. This section presents the quantitative analysis of the publications of TSTTM based on their sub-categories. It can be observed that from Fig. 9a the most prominent category in this paper is the probabilistic models. Over one-third of the research papers are dedicated to this category; it got 33.65% rate of the publications, which is the highest rate among all sub-categories. The second and third prominent groups of this category are clustering and label-based models, respectively; each of them covers over one-tenth of the papers.

7.3.2 ASTTM publications

The second part of this taxonomy has categorized the ASTTM models into five sub-categories: DMM, Global-word co-occurrence, self-aggregation, deep learning-based models, and other ASTTM model. This section analyzes quantitatively the publications of ASTTM based on their sub-categories. We can observe from Fig. 9b that over one-fourth (26.6%) of the papers cover Global word co-occurrences based models, and one-third (34.04%) of the papers cover deep learning-based models. Recently, most researchers have paid attention to deep learning-based models. The rate of the published papers in DMM-based models is 17.02%, which is the third prominent sub-category in ASTTM publications. From both Figs. 9a and b, we can conclude that the higher rate of publications was probabilistic class



(a) TSTTM Topic modelling **(b)** ASTTM Topic Modeling

Fig. 8 Papers distribution based on year of publication in STTM

according to the TSTTM publications. In contrast, the deep learning-based models class is considered the most productive sub-category for publications in ASTTM publications.

7.4 STTM publications based on both Publication years and sub-categories

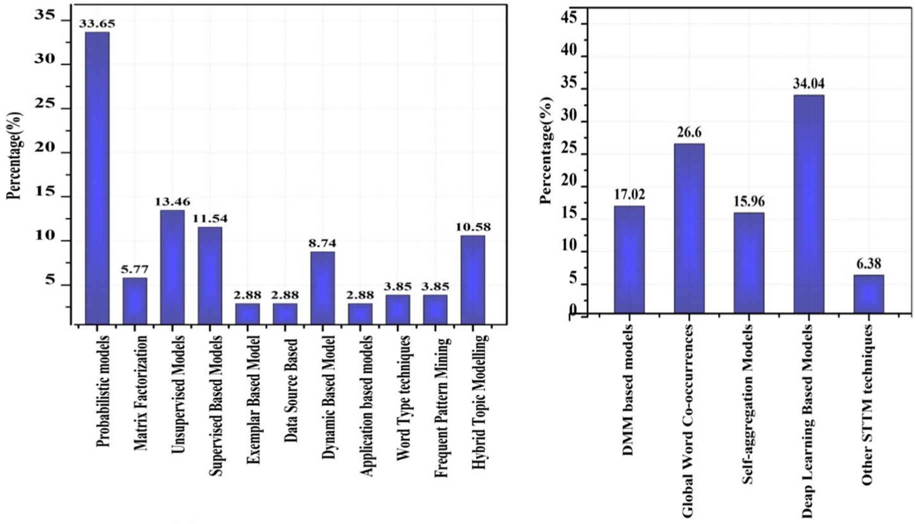
This section quantitatively presents the analysis of the STTM publications based on both the publication years and the sub-categories. Figure 10 shows the number of papers in each sub-category of TSSTM models over the publication year. In recent years, probabilistic, clustering, and supervised based models seem to be worth exploring for researchers over the course of years. In contrast, exemplar, matrix factorization, and frequent pattern mining have gained less attention currently than before. Figure 11 illustrates the number of papers in each sub-category of ASSTM models over publication years. It can be observed that the deep learning-based models, Global word co-occurrences and DMM models have gained more attention, respectively, further than the self-aggregation based model from the side of researchers in recent years.

7.5 STTM models publications based on platform

In this section, we analyze the environment platform used for the implementation of the existing STTM models.

7.5.1 TSTTM publications

This sub-section quantitatively analyzes the TSTTM publications based on the utilized platform. From Fig. 12a, it is observed that most of the publications did not mention the



(a) TSTTM models

(b) ASTTM models

Fig. 9 Papers distribution over selected categories

utilized platform where Not Mention (NM) class has got a 48.21% rate, the Java has got 17.88% rate which is the highest of the mentioned platform. Moreover, Matlab has earned 14.29% rate, which is the second-highest rate. The Python platform has obtained 10.71%, whereas the R language has got 3.57% rate which is the less rate.

7.5.2 ASTTM publications

In this sub-section, we analyze the ASTTM publications based on the utilized platform quantitatively. From Fig. 12b, it is clearly noted that Java has got 17.75% rate, Python has 35.14% rate, and the C++ language has got 5.41%, whereas Spark has used only in one paper and got 1.35% rate. From both Fig. 12a and b, we can conclude that the spark has less attention from the researcher, and the researchers can bridge this gap for future works as the spark will process the data in parallel, and the cost of computational time will reduce.

7.6 Quantitative analysis of the existing datasets

This sub-section provides a quantitative analysis of the existing datasets utilized in the current STTM models based on their utility, language, and source.

7.6.1 Quantitative analysis of the existing datasets based on the utility

This section quantitatively analyzes the prominent datasets utilized to evaluate the existing STTM models. The number of existing and collecting datasets is 51; 20 of them are collected from the Twitter platform, whereas six datasets are collected from Chinese Sogou Labs and used only specifically, as introduced in Table 12. The other 25 rest are the prominent datasets available publicly. Here, we focus only on the 25 prominent datasets used in 114 papers. It is clearly observed from Fig. 13 that the Web Snippets have got a 10.53% using rate, which considers the highest rate, whereas the (Youtube, Viber, Swiftkey, NOAA Radar, and Wikipedia) datasets have got 1.7% which considers the least rate of usage among of them. The second highest datasets are BaiduQA/Question and Sina Weibo, which have gained of 8.77% rate. Generally, if we evaluate the utility of the prominent dataset based on source, then the Twitter dataset is the highest rate, which is 20.05% (sum of the four prominent twitter datasets). Finally, the observation is that less attention is given to real-world data in social media. Therefore, it is better to evaluate the STTM models on the real-world dataset to extract the trending topics and discover the emerging latent topics of discussion from constant background chatter in social media.

7.6.2 Quantitative analysis of the existing datasets based on language

This section quantitatively presents the analysis of the datasets language used in both TSTTM and ASTTM models. The existing datasets languages utilized to evaluate the state-of-the-art STTM models are six languages: English (EN) and Chinese (CHI). Turkish (TUR), Japanese (JPN), Germanize (GER) and Portuguese (POR).

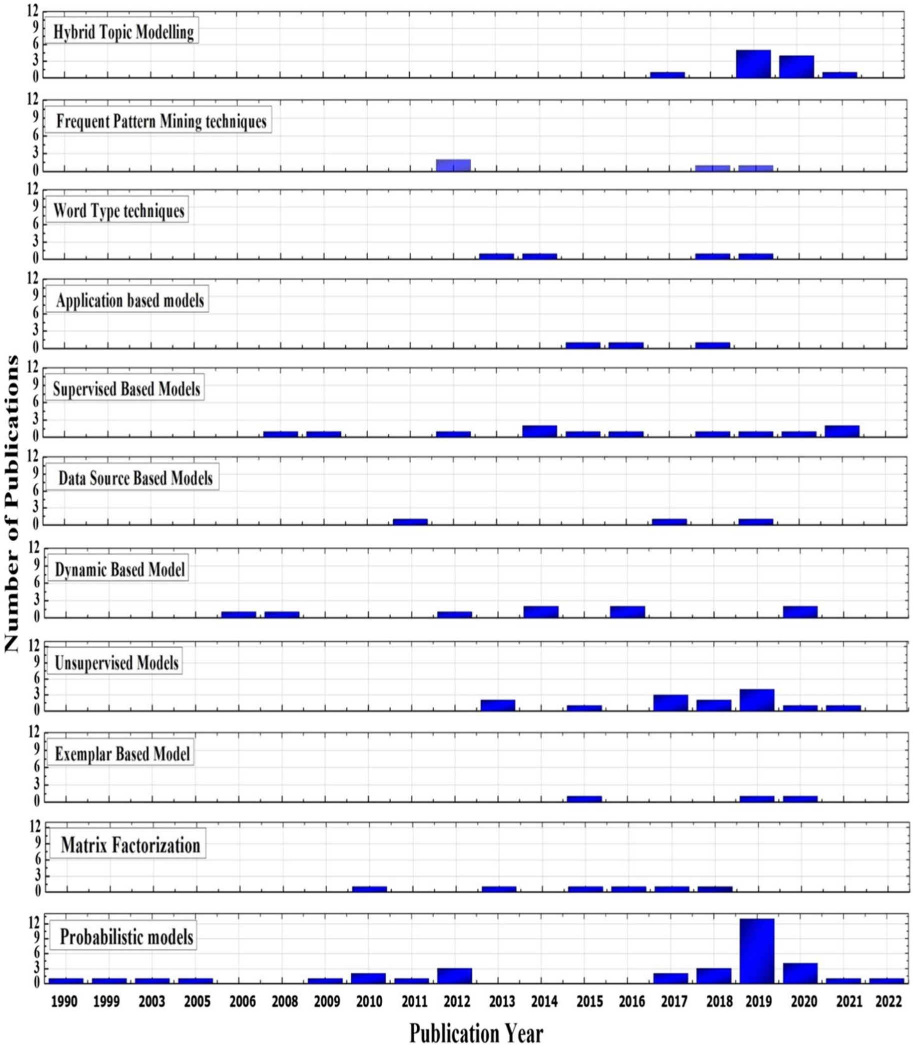


Fig. 10 Distribution of papers in categories of TSTTM models over publication years

Datasets languages utilized on TSTTM models As it is clearly observed from Fig. 14a, the English dataset has got 77.94% which is the highest rate, and the Chinese datasets have gained 13.24% rate, whereas the Germanize has obtained 1.47%, which is the less rate.

Datasets languages utilized on ASTTM models From Fig. 14b, it is noted that the Portuguese (POR) has obtained 1.01%, CHI has got 26.04%, whereas the EN has acquired 72.92%, which is the highest datasets language used in the ASTTM models. From Fig. 14a and b, we can conclude that the English datasets have gotten more attention from researchers. Besides, the researchers have not paid attention to other languages in the existing works. So the researchers can bridge this gap for future works.

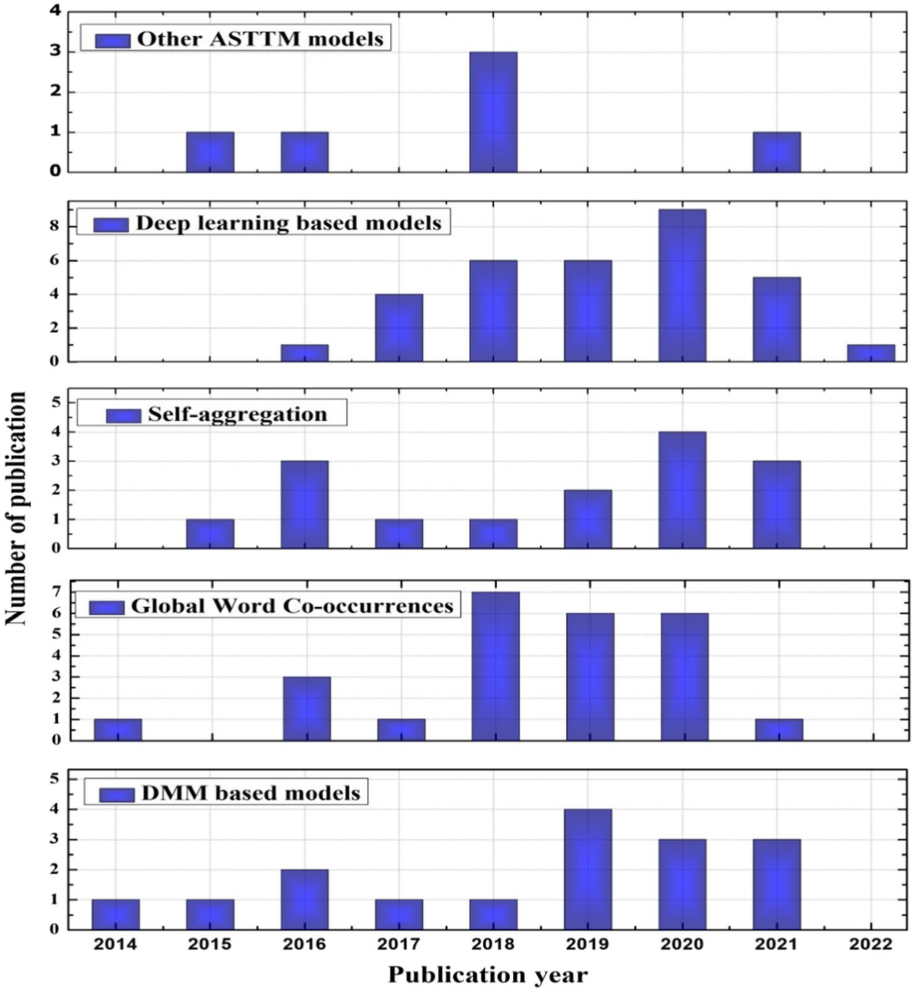
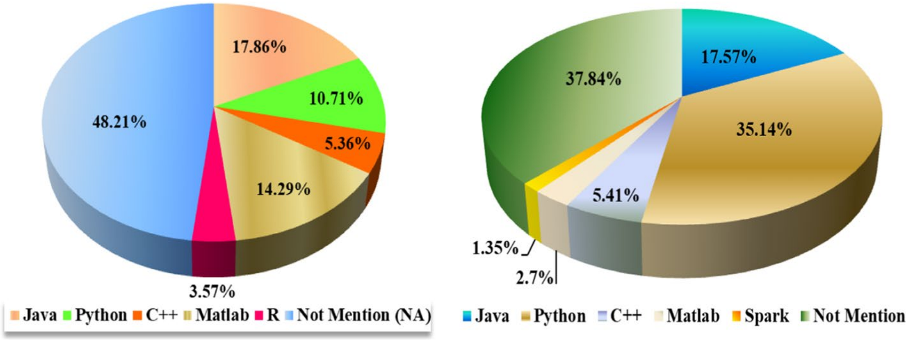


Fig. 11 Distribution of papers in categories of ASTTM models over publication years



(a) TSTTM publications **(b)** ASTTM publications

Fig. 12 Distribution of platforms utilized for STTM models

7.6.3 Quantitative analysis of the existing datasets based on sources

This section shows the quantitative analysis of the sources of the datasets utilized in the both TSTTM and ASTTM models. Generally, the sources of the datasets, which have been used for gathering datasets to evaluate the state-of-the-art STTM models, are as follows: Twitter, Weibo, Sogou Labs, Yahoo, websites, amazon, news, and others. Here, it is observed that from Fig. 15, Twitter social media has used 30.77% for the dataset, which is the highest rate of all the sources. Weibo has got a 12.09% rate, whereas amazon has got a 3.3% rate among datasets on STTM models, which are the least sources used to gather the datasets. We can conclude that the Twitter platform has gotten more attention from researchers. The researchers have not paid attention to using Facebook, Instagram, Tik Tok, and Whatsup social media to gather the datasets in the existing works. So the researchers can bridge this gap for future works.

7.7 Quantitative analysis of STTM evaluation metrics

This section summarizes quantitatively the different metrics utilized for evaluating both TSTTM and ASTTM models, which evaluate the quality of extracted topics as shown in Tables 15 and 16, respectively. The topics must be evaluated to measure their efficiency based on appropriate performance evaluation metrics. The main common evaluation metrics which were considered in the literature are: topic coherence (Fang et al. 2016a; Mimno et al. 2011), perplexity, Point-wise Mutual Information (PMI) (Newman et al. 2010), Normalized PMI (NPMI), purity (Zhao and Karypis 2001), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI) (Yan, Guo, Liu, et al. 2013), topic recall, Accuracy, precision, recall, and F-scores. Also, there are other metrics only once or twice times

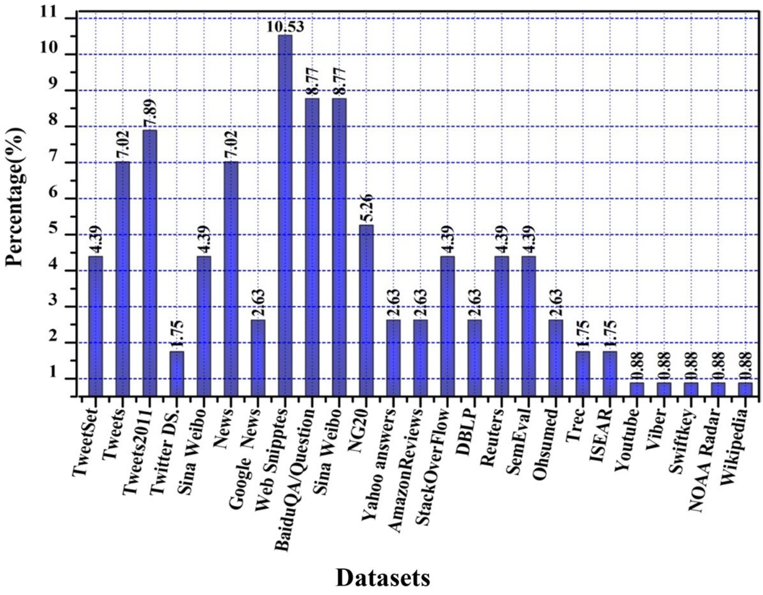
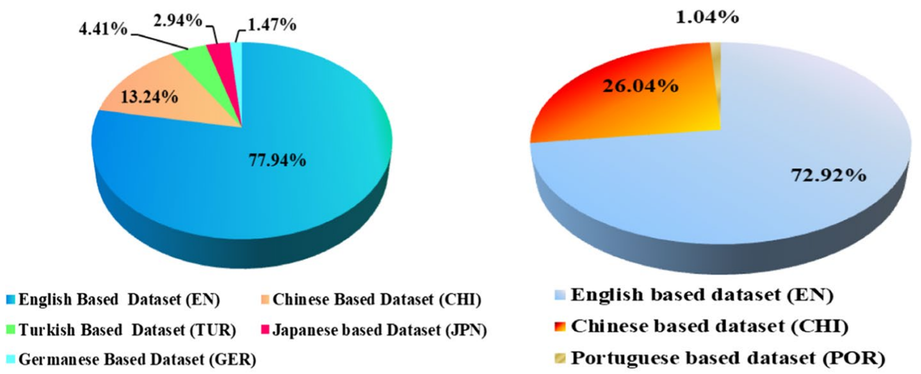


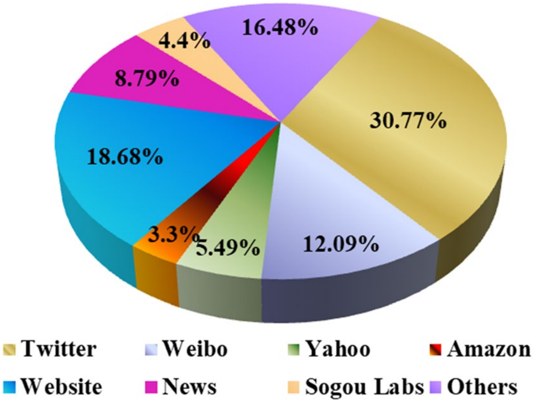
Fig. 13 Distribution of prominent datasets based on the utility of STTM



(a) TSTTM models **(b)** ASTTM models

Fig. 14 Distribution of dataset languages on STTM models

Fig. 15 Distribution of the existing datasets utilized based on Sources



used throughout the studies, such as Rank index (RI), Mean Average Precision(MAP), Topic Mixing Degree (MD), Estimate Error (ERR), Topic Effectiveness(T.E), KL Distance(KLD), Matthews Correlation Coefficient (MCC), Calinsiki-Har-abasz index (CH), log-likelihood indicates (Log-LH), Classification Error Rate (C.E.R), Mean Absolute Error (MAE), Topic Relevance (Topic Rev), Hellinger distance (HD), Mean Average Precision (MAP), Averaged Pearson's correlation coefficients (AP), Discounted Cumulative Gain (DCG), Normalized Discounted Cumulative Gain (NDCG), Similarity (Murshed et al. 2020), Root Mean Squared Error (RMSE), Mean Reciprocal Rank (MRR), Cohen's kappa score (Kappa), Homogeneity (H), Completeness (C) (Rosenberg and Hirschberg 2007), and Word Embedding-based topic coherence measure (WE-based Metrics Similarity) (Fang et al. 2016b) as shown in Tables 15 and 16.

This sub-section presents the quantitative analysis of the evaluation metrics utilized in both TSTTM and ASTTM models.

Evaluation metrics utilized in TSTTM models In terms of coherence, It can be observed that from Fig. 16 (a), the perplexity has been used more than the other coherence metrics, and it has got 9.68% rate, whereas the topic coherence and PMI/NPMI have got 6,45% and 4.84% respectively. Perplexity has been less effectively utilized incomprehension of the semantic essence of the learned topics. Therefore, many researchers such as (Fang et al. 2016a; and Mimno et al. 2011) proposed topic coherence metrics to address this issue. The accuracy has got 19.35% which is the highest evaluation metric used in the TSTTM models, whereas the ARI has got 0.81 which is the least evaluation metric used in this category.

Evaluation metrics utilized in ASTTM models In this part of the taxonomy ASTTM, the researchers have more attention to the utility of topic coherence to evaluate the quality of the extracted topics, so the topic coherence has got 17.9% rate which is the highest used among all the other evaluation metrics. The PMI/NPMI has got 12.35%, which is the second-highest metric in terms of coherence, whereas the accuracy, F-measure, and precision have got 14.81%, 11.73% and 10.49%, respectively. On the other hand, the AMI and ARI have got the least rate. Figure 16 (b) depicts the percentage of using the evaluation metrics in ASTTM models. There is another topic coherence metric based on word embedding, namely (WE-based metrics Sim) proposed by (Fang et al. 2016b), which was used by (Huang et al. 2020). In summary, Beneficial assessment metrics have never been resolved for topic discovery models (Qiang et al. 2020). Topic coherence cannot differentiate between topics. Moreover, only one category of the topic modelling models has gotten the attention of using the recent metrics. Developing new evaluation criteria is a future research work for topic modelling that matches how the models are utilized.

8 Experimental analysis

This section presents details about the evaluation of the prominent topic models for extracting hidden topics from short texts performed over two datasets crawled from the Twitter social media platform: the Real-world pandemic Twitter dataset (RW-Pan–Twitter) and Real-world Cyberbullying Twitter dataset (RW-CB-Twitter). Ten main models from the literature are selected: the traditional models of LDA (Blei et al. 2003) and NMF are chosen based on their extensive usage for short text topic models, especially in Twitter-related topic modelling. An extension of the LDA known as the Twitter-LDA (Zhao et al. 2011) is chosen as it is one of the first topic models that addressed the problem of sparsity by defining an application-specific approach. Besides, CDTM (Wang et al. 2008) and FTM (Rashid et al. 2019b) models are chosen. BTM (Cheng et al. 2014), PTM (Zuo, Wu, et al. 2016a), SATM (Quan et al. 2015), WNTM (Zuo, Zhao, et al. 2016b), and GLTM (Liang et al. 2018) are chosen because they represent common and most effective short text topic models for discovering latent topics from a short text. Hence, the ten models: LDA, NMF, Twitter-LDA, CDTM, FTM, BTM, PTM, SATM, WNTM and GLTM, were implemented over both the real-world datasets: RW-Pand–Twitter and RW-CB-Twitter datasets to evaluate their performance. The implementation of the considered models is written in Python language, and the experiments were conducted using Pycharm IDE. The datasets consist of tweets collected from multiple topics of interest. The Twitter datasets could be collected using the Twitter streaming API using Python language with tweepy package. Following that, the pre-processing step is applied to the dataset to clean up the data using toolkits such as the NLTK python package (Anil Phand and Chakkarwar 2018) that provides stopword and punctuation removal, tokenization, lemmatizing, stemming, identifying n-gram procedures, and lowercase transformation and also other pre-processing and data cleansing steps (Murshed et al. 2021). Finally, the TM methods are applied to extract a set of recurrent themes/topics that are explored throughout the collection of posts and the extent to which each post reflects those themes. We compare the performance of considered

Table 15 Analysis of the evaluation metrics utilized in prominent TSTTM models

Methods/Ref	References	Evaluations metrics												
		Coherence			In terms of clustering				In terms of classification					Other metrics
		Topic Coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	Entropy	Accuracy	Recall	Topic Recall	Precision	F-Measure	
PLSA-ICA	Kumar and Vardhan (2019)								✓				✓	
PLSA	Hofmann (1999)		✓							✓		✓		
LDA	Blei et al. (2003)		✓						✓					
RO-LDA	Chen and Kao (2017)				✓	✓			✓					RI
TSVB-LDA	Fang et al. (2017)	✓												MD,ERR
BR-LDA	Ni et al. (2018)									✓		✓	✓	
CTD	Sharath et al. (2019)					✓				✓		✓	✓	MAP
Logistic LDA	Korshunova et al. (2019)								✓					
TIN-LDA	Zheng et al. (2019)		✓											KLD, TE
TM-LDA	Wang et al. (2012)		✓											Time
FBLDA	Akhtar et al. (2019b)			✓	✓	✓								

Table 15 (continued)

Methods/Ref	References	Evaluations metrics												
		Coherence			In terms of clustering				In terms of classification					Other metrics
		Topic Coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	Entropy	Accuracy	Recall	Topic Recall	Precision	F-Measure	
Twitter-LDA	Zhao et al. (2011)	✓												
TH-LDA	Yu et al. (2019)		✓	✓										
MBA-LDA	Han et al. (2020)									✓		✓	✓	
LSA	Deerwester et al. (1990)									✓		✓		
FLSA	Karami et al. (2018)								✓				✓	ROC, MCC
W2V-LSA	Kim et al. (2020)	Umass		NPMI										KMS
NMF	Yan, Guo, Liu, et al. (2013b)		✓		✓	✓			✓					
Ensemble NMF	Belford et al. (2016)				✓									Stability
N-cut-weighted NMF	Yan et al. (2012)				✓	✓	✓							
RED-NMF	Iskandar (2017)	✓	✓											
NMF-LTM	Chen et al. (2020b)	✓												NDCG@Υ

Table 15 (continued)

Methods/Ref	References	Evaluations metrics												
		Coherence			In terms of clustering				In terms of classification					Other metrics
		Topic Coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	Entropy	Accuracy	Recall	Topic Recall	Precision	F-Measure	
CSS	Farahat et al. (2015)								✓					
Exemplar	Elbagoury et al. (2015)										✓			
MVTD	Fang et al. (2014)				✓			✓	✓					
EFCM	Muliawati and Murfi (2017)								✓		✓			
KEFCM	Prakoso et al. (2018)								✓		✓			
Tweet-SCAN	Capdevila et al. (2017)												✓	Js Distance
FTM	Rashid et al. (2019b)			NPMI		✓		✓	✓				✓	
IFCM	Abou-Of (2020)							✓					✓	
BN-grams and Doc-pivotal	Indra and Pulungan (2019)							✓	✓		✓			
DCT	Liang et al. (2016)		✓							✓		✓		NDCG, MAP

Table 15 (continued)

Methods/Ref	References	Evaluations metrics												Other metrics
		Coherence			In terms of clustering				In terms of classification					
		Topic Coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	Entropy	Accuracy	Recall	Topic Recall	Precision	F-Measure	
GDTM	Ghoorchian and Sahlgren (2020)	✓							✓	✓		✓	✓	
DTM	Yao and Wang(2020)	UCI		NPMI										
SSHLDA	Mao et al. (2012)		✓										✓	
LF-LDA	Zhang et al. (2018c)												✓	
DF-LDA	Li et al. (2015)												✓	
MS-LDA	Yu et al. (2017)									✓		✓	✓	
Bi-Labeled LDA	He et al. (2020b)													DCG, Hit No
WLTM, XETM	Pang et al. (2019)								✓					HD,AP
Bilingual LDA	Wang and Iwaihara (2015)													Cos. Sim
Wiki-LDA	Pu et al. (2016)												✓	
ED-LDA	Feng (2018)		✓											

Table 15 (continued)

Methods/Ref	References	Evaluations metrics												
		Coherence			In terms of clustering				In terms of classification					Other metrics
		Topic Coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	Entropy	Accuracy	Recall	Topic Recall	Precision	F-Measure	
SOW	Koike et al. (2013)											✓		
Twitter-TTM	Sasaki et al. (2014)		✓											
SOW-LSTM(T-PAN)	Dey et al. (2018)								✓				✓	
FPM	Guo et al. (2012)													C.E.R
ET-EPM	Peng et al. (2018a)									✓		✓	✓	
ET-HUPM	Choi and Park (2019)									✓	✓	✓	✓	Topic Relevance
T-LDA	Huang et al. (2017)		✓							✓		✓	✓	
TF-IDF and LDA	Ge et al. (2019)									✓		✓	✓	
deep learning and LDA	Zhang et al. (2019)									✓		✓	✓	
FKLSA	Rashid et al. (2019a)								✓	✓		✓	✓	Specificity, CH, Log-LH

Table 15 (continued)

Methods/Ref	References	Evaluations metrics												
		Coherence			In terms of clustering				In terms of classification					Other metrics
		Topic Coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	Entropy	Accuracy	Recall	Topic Recall	Precision	F-Measure	
LDA and matrix factorization-based neural network	Pornwattanavichai et al. (2020)													MAE, Coverage
LSTM-HC	Zhang and Zhang (2020)				✓				✓	✓			✓	
TRNMF	Yi et al. (2020)	✓		NPMI					✓	✓		✓	✓	
SS-LDA	Ozyurt and Akcayol (2021)									✓		✓	✓	
Hybrid Model	Shahbazi and Byun (2021)													
SL-LDA,AL-LDA	(Wang et al. 2021a)									✓		✓	✓	

Table 16 Comparative analysis of the evaluation metrics utilized in ASTTM models

Methods	References	Evaluations metrics												
		Coherence			In terms of clustering					In terms of classification				Other metrics
		Topic coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	AMI	Entropy	Accuracy	Recall	Precision	F-Measure	
GSDMM	Yin and Wang (2014)				✓		✓	✓						H, C
FGSDMM	Yin and Wang (2016)				✓									H,C
PDMM	Li et al. (2017)	✓								✓				
GPU-DMM	Li et al. (2017)	✓								✓				
GPU-PDMM	Li et al. (2017)	✓								✓				
Improved GPU-DMM	Zhang et al. (2018b)	✓												
GPM	Mazarura et al. (2020)	✓		✓						✓			✓	
LF-DMM	Nguyen et al. (2015)	✓		✓	✓	✓							✓	
ULW-DMM	Yu and Qiu (2019)													
Lap-DMM	Li, J. Zhang, et al. (2019b)	✓			✓					✓				
X-DMM	Li et al. (2019a)		✓		✓									
WS-DMM	Xiao et al. (2019)											✓		
CME-DMM	Liu et al. (2020b)	✓		✓							✓	✓	✓	
TATM	He et al. (2020a)	✓		✓		✓				✓				

Table 16 (continued)

Methods	References	Evaluations metrics												
		Coherence			In terms of clustering					In terms of classification				Other metrics
		Topic coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	AMI	Entropy	Accuracy	Recall	Precision	F-Measure	
COTM	Yang et al. (2018)	✓								✓				
BTM	Cheng et al. (2014)	✓		✓						✓				
SBTM	Pang et al. (2016)									✓				Sim
NBTMWE	Huang et al. (2020)	✓		✓										WESim
LS-BTM	Li et al. (2018a)									✓				
R-BTM	Li, A. Zhang, et al. (2019)			✓	✓					✓				
IBTM	Zhu et al. (2019b)			✓						✓		✓		
NSLPCD	Singh and Singh (2020)				✓						✓	✓	✓	RI
MTM	Wu and Li (2019)	✓		✓		✓				✓				
WNTM	Zuo, Zhao, et al. (2016b)	✓									✓	✓	✓	
HVCHfusion	Liqing et al. (2019)		✓								✓	✓	✓	
R-WNTM	Wang et al. (2016)	✓											✓	

Table 16 (continued)

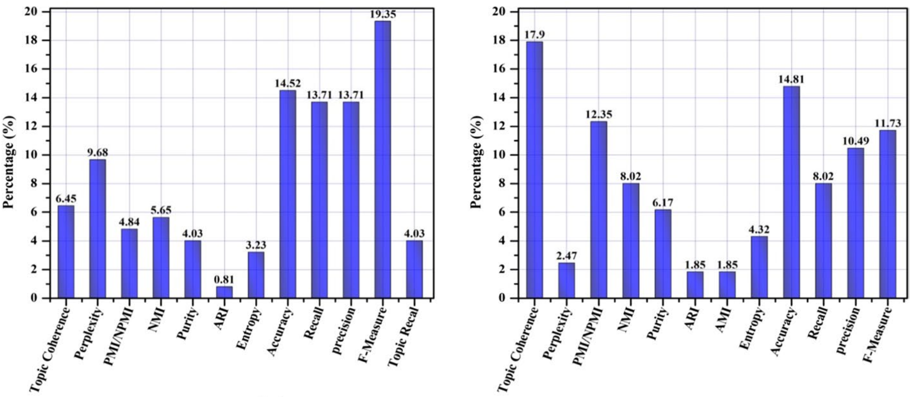
Methods	References	Evaluations metrics												
		Coherence			In terms of clustering					In terms of classification				Other metrics
		Topic coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	AMI	Entropy	Accuracy	Recall	Precision	F-Measure	
WNTM-W2V	Jiang et al. (2018)									✓	✓		✓	
AOBTM	Hadi and Fard (2020)			✓							✓	✓	✓	Dis Score
CWTM	Diao et al. (2017)										✓	✓	✓	
UGTM	Akhtar and Beg (2019a)			✓	✓	✓			✓					
PYPM	Qiang et al. (2018b)				✓		✓	✓						C, H
GLTM	Liang et al. (2018)	✓								✓			✓	
DP-BMM	Chen et al. (2020a)				✓									
DP-BMM-FP	Chen et al. (2020a)				✓									
CSTM	Li, Wang, et al. (2018c)			NPMI		✓			✓	✓				
FTM	Rashid et al. (2019b)			NPMI		✓			✓	✓			✓	
BG & SLF–Kmeans	Wu et al. (2020a)		✓		✓	✓				✓	✓	✓	✓	
COTM	Yang et al. (2018)	✓								✓				

Table 16 (continued)

Methods	References	Evaluations metrics												
		Coherence			In terms of clustering					In terms of classification				Other metrics
		Topic coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	AMI	Entropy	Accuracy	Recall	Precision	F-Measure	
SATM	Quan et al. (2015)	✓		✓		✓				✓				
SADTM	Shi et al. (2019a)			✓		✓			✓		✓	✓		
Aggregate TM	Blair et al. (2020)	✓		✓										
PTM	Zuo, Wu, et al. (2016a)	UCI									✓	✓	✓	
UGTE	Feng et al. (2020a)	✓							✓					Kappa
SPTM	Zuo, Wu, et al. (2016a)	UCI									✓	✓	✓	
BPDTM	Jiang et al. (2016)	✓			✓	✓	✓	✓	✓	✓				
MB-HDP	Liu et al. (2015)	✓	✓											
PTNG	Lin et al. (2020a)	UCI	✓							✓	✓	✓	✓	
Topicsketch	Xie et al. (2016)	✓									✓	✓		
PTDAS	Zhang et al. (2018a)										✓	✓	✓	
MRTM	Liu et al. (2018)	✓		✓										

Table 16 (continued)

Methods	References	Evaluations metrics												
		Coherence			In terms of clustering					In terms of classification				Other metrics
		Topic coherence	Perplexity	PMI/NPMI	NMI	Purity	ARI	AMI	Entropy	Accuracy	Recall	Precision	F-Measure	
ASTM	Wang et al. (2018c)			NPMI						✓				
GSDMM& sentiment Analysis	Garcia and Berton (2021)											✓	✓	
Lab-DMMT OLapDMM	Li et al. (2021)	✓		✓					✓					
WE-PTM	Zuo et al. (2021)	✓								✓		✓	✓	
AOTM	Yang and Wang (2021)	✓	✓											



(a) TSTTM models **(b)** ASTTM models

Fig. 16 Distribution of using evaluation metrics in both TSTTM and ASTTM models

topic discovery models in terms of purity, NMI, ARI, accuracy and topic coherence with k different number of topics such as $k = \{\#ground\, truth\, topics, 20, 40, 60,\ and\, 80\}$, determined according to the labelled datasets. The number of iterations is fixed as 1000 for all the experiments except that when analyzing the influence of the number of iterations on the performance of the selected topic modeling. In this case, we fixed the *Number_of_Iteration*$=\{5, 10, 20, 40, 60, 80, 100, 200, and\, 400\}$. All setup parameters were set for the experiments based on the used setup parameters in the original papers of the selected models. The evaluation metrics were selected to display high performance of the topic discovery of each model, and the implemented algorithms were run 20 times for each evaluation metric to obtain an average value.

8.1 Parameters settings

In this sub-section, we provide the setting of the common parameters used in the considered models. The number of topics for all experiments is set as $k = \{\#ground\, truth\, topics, 20, 40, 60, and\, 80\}$. The number of iterations is fixed to 1000 for all models, whereas we set the number of iterations as the following *Number_of_Iterations* $= \{5, 10, 20, 40, 60, 80, 100, 200, and\, 400\}$, only for checking the influence of the number of iterations on the performance of the considered ten topic modelling as illustrated in Sect. 8.4.4. All the parameters of the considered models for experiments are chosen as recommended by the authors in the original papers. In GLTM, BTM, and SATM, the value of α is fixed as 50/K, and we set α = 0.1 for both PTM and WNTM, whereas we set α = 0.05 for all the following models: LDA, TwitterLDA and CDTM. We set β = 0.1 for SATM and WNTM whereas, we set β = 0.01 for the other models such as BTM, GLTM, LDA, TwitterLDA, WNTM, CDTM, and PTM models. In according to the hyper-parameter λ, we set λ = 0.5 and λ = 0.1 for GLTM and PTM models, respectively. The number of pseudo documents in PTM and SATM is fixed to 1000 and 300, respectively. The sliding window is fixed as 10 in the WNTM model.

8.2 Datasets

In this section, we have collected two real-world Twitter datasets: The Real-World Pandemic Twitter dataset named (RW-Pand-Twitter) and the Real-World Cyberbullying Twitter dataset named (RW-CB-Twitter). The descriptions of both datasets are provided below. Table 17 shows the statistics of the collected datasets.

8.2.1 Real-world Pandemic Twitter (RW-Pand-Twitter) dataset

This dataset consists of tweets collected from multiple topics of interest. The Twitter dataset on seven major topics of pandemics was extracted using the Twitter streaming API using Python language with tweepy package. These seven selected topics of interest include Cholera, Coronavirus, Dengue fever, Malaria, Chikungunya, Ebola virus disease and Swine flu. A total of 971,132 tweets were collected using the streaming API and were filtered to remove the re-tweets, duplications, and non-English Tweets. After removing such tweets, a total of 330,159 tweets were obtained in the seven specified topics. From these tweets, 42,000 tweets were selected and fed to the system to evaluate the described topic models.

8.2.2 Real-world Cyberbullying Twitter (RW-CB-Twitter) dataset

The RW-CB-Twitter dataset is collected from the Twitter social media platform utilizing Twitter API streaming with the use of some Cyberbullying keywords like a threat, terrorist, attack, kill, hate, black, Islam, racism, Islamic, and ban as provided in (Zhang et al. 2018d) whereas the other keywords such as whale, fuck, pussy, fucking, moron, ugly, LGBTQ, poser, idiot, bitch, whore, nigger, etc., as suggested in the psychology literature (Nand et al. 2016); Squicciarini et al. 2015; Cortis and Handschuh 2015; Cheng et al. 2019). The RW-CB-Twitter dataset is extended to the dataset utilized in (Murshed et al. 2022) and labelled into five classes: sexism, racism, aggressive (harassment), insult, and not-bullying. The gathered dataset consists of 435,764 tweets, which includes several outliers. So, only the English tweets are needed, and the other language tweets are filtered out. Similarly, re-tweets have been deleted from the dataset because they are not informative. Finally, after removing the irrelevant tweets, we selected only 20,000 tweets from the rest of the dataset for the experiments in this research.

8.3 Evaluation metrics

This sub-section briefly describes the utilized evaluation metrics for evaluating the efficiency of STTM models. The evaluation in this study was conducted in terms of (1) evaluation utilizing clustering measures such as Normalized Mutual Information(NMI) (Singh and Singh 2020), Purity (Zhao and Karypis 2001), Adjusted Random Index (ARI), (2)

Table 17 The statistics of the datasets

# Tweets	# Labels (topics)	Sources of data
42,000	7	Twitter
20,000	5	Twitter

Evaluation utilizing classification accuracy, and finally, (3) Evaluation utilizing Topic Coherence (TC) (Röder et al. 2015) using PMI. These measures are described briefly below.

8.3.1 Clustering evaluations

Supposing $D = \{d_1, d_2, \dots, d_k\}$ is the set of derived clusters with the number of clusters denoted as k, each d_k is a tweet in cluster k, and $S = \{s_1, s_2, \dots, s_m\}$ is the set of labelled clusters (ground-truth) already determined in the dataset with the number of labelled clusters denoted as m. We have adopted three measures to assess the quality of clusters of set D.

8.3.1.1 Purity metric Purity (Zhao and Karypis 2001) is the measure used to evaluate the degree of the clustered tweets similar to the labelled datasets. It is based on the accuracy of the clustered tweets and is computed as the number of correctly clustered tweets among the total number of labelled tweets in the dataset. Purity is expressed as given in Eq. (7).

$$Purity(D, S) = \frac{1}{N} \sum_k \max_m |d_k \cap s_m| \tag{7}$$

Here, N denotes the number of labelled tweets in the dataset. It must be noted that the purity values lie between 0 and 1. The low clustering quality results in a purity value of zero, while a high perfect clustering quality results in a purity value of 1.

8.3.1.2 NMI metrics NMI (Yan, Guo, Liu, et al. 2013) computes the mutual information shared between D and S, which is normalized by the mean entropy of clusters represented as $H(D)$ and entropy of classes represented as $H(S)$. Similar to the purity value, the NMI ranges from 0 to 1, and the larger values of NMI imply a higher level of accuracy in clustering. It can be computed as given in Eq. (8).

$$NMI(D, S) = \frac{I(D, S)}{[H(D) + H(S)]/2} \tag{8}$$

Here $I(D, S)$ is the mutual information between D and S which can be statistically computed as

$$I(D, S) = \sum_k \sum_m P(d_k \cap s_m) log \frac{P(d_k \cap s_m)}{P(d_k)P(s_m)} \tag{9}$$

Here $P(d_k)$ indicates the probability of a tweet possibly present in the cluster d_k, $P(s_m)$ the probability of a tweet possibly present in s_m and $P(d_k \cap s_m)$, the probability of a tweet possibly present in both the clusters d_k and s_m. The mutual information can be rewritten based on the number of tweets N in the original labelled dataset as in Eq. (10).

$$I(D, S) = \sum_k \sum_m \frac{|d_k \cap s_m|}{N} log \frac{N|d_k \cap s_m|}{|d_k||s_m|} \tag{10}$$

Here N denotes the number of tweets in S, $|d_k|$, $|s_m|$ indicate the number of tweets in d_k and s_m, respectively and $|d_k \cap s_m|$, the number of tweets occurring in both the clusters d_k and s_m.

Similarly, the entropy of classes $H(S)$ and entropy of clusters $H(D)$ are computed as follows:

$$H(C) = -\sum_{m} P(s_m) logP\left(s_m\right) = -\sum_{m} \frac{\left|s_m\right|}{N} log \frac{\left|s_m\right|}{N} \tag{11}$$

$$H(D) = -\sum_{m} P(d_k) logP\left(d_k\right) = -\sum_{m} \frac{\left|d_k\right|}{N} log \frac{\left|d_k\right|}{N} \tag{12}$$

8.3.1.3 ARI metric ARI is the Adjusted Random Index that measures the correctness of a decision on clustering two tweets based on similarity. Clustering is considered as a pair-wise decision. For measuring ARI, if two tweets are located in the same class and in the same cluster or both in different classes and clusters, then the decision is considered correct. Rand Index (RI) is the percentage of correct decisions. The ARI is the corrected for the chance version of RI. The highest value of 1 denotes perfect clustering and the more similarity between labels and clustering results. It is computed as expressed in Eq. (13).

$$ARI(D,S) = \frac{\sum_{k,m}\binom{\left|d_k \cap s_m\right|}{2} - \left[\sum_k \binom{\left|d_k\right|}{2} \sum_m \binom{\left|s_m\right|}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_k \binom{\left|d_k\right|}{2} + \sum_m \binom{\left|s_m\right|}{2}\right] - \left[\sum_k \binom{\left|d_k\right|}{2} \sum_m \binom{\left|s_m\right|}{2}\right] / \binom{N}{2}} \tag{13}$$

8.3.2 Topic coherence metric

Topic coherence is a measure to assess the quality of topic models. For each K topic of tweets generated, the topic coherence is applied to the top N words. In the experiment, we selected 10 top words with high probabilities $(w_1, \ldots, w_N)$ as a sliding window. Topic coherence is computed using the PMI metric and following to (Li et al. 2017) (Murshed et al. 2020). It measures the semantic score of a single topic by measuring the degree of semantic similarity between high scoring words in the topic and is computed as follows.

$$Coherence(K) = \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} PMI(w_i, w_j) \tag{14}$$

Here w_i, w_j indicates to the top words pairs of the topic. This score can be computed by means of PMI. PMI is the Point-wise Mutual Information value between the topic words in a cluster.

$$PMI\left(w_i, w_j\right) = \log \frac{P\left(w_i, w_j\right)}{P\left(w_i\right)P\left(w_j\right)} \tag{15}$$

Here $P(.)$ denotes the probability or likelihood of the topic words in the clusters.

8.3.3 Classification accuracy metric

The topic modelling performance can be evaluated using text classification. We leverage the extracted topics determined by STTM models as tweet features. Each tweet can be described by document-topic distribution. We adopt the document-topic distribution by $p\left(z_i|d\right)$ for all the considered models. The accuracy is chosen as the metric for classification purposes. We randomly split both datasets: RW-CB-Twitter and RW-Pand-Twitter datasets, into training and test sub-datasets and used the Support Vector Machine (SVM) classifier for classification. K fold Cross-Validation (CV) has been used to compute the classification accuracy, where K is fixed to 5.

8.4 Experimental results

This section describes the experimental results obtained from the selected ten state-of-the-art models such as LDA, TwitterLDA, NMF, FTM, CDTM, SATM, BTM, PTM, WNTM, and GLTM over two Twitter datasets: Real-world pandemic Twitter dataset (RW-Pand-Twitter) and Real-world Cyberbullying Twitter dataset (RW-CB-Twitter). We compare the performance of considered models in terms of Purity, NMI, ARI, accuracy and topic coherence with k different number of topics such as k = $\{\#\text{ ground truth topics}, 20, 40, 60,\ \text{and } 80\}$.

8.4.1 Classification accuracy findings

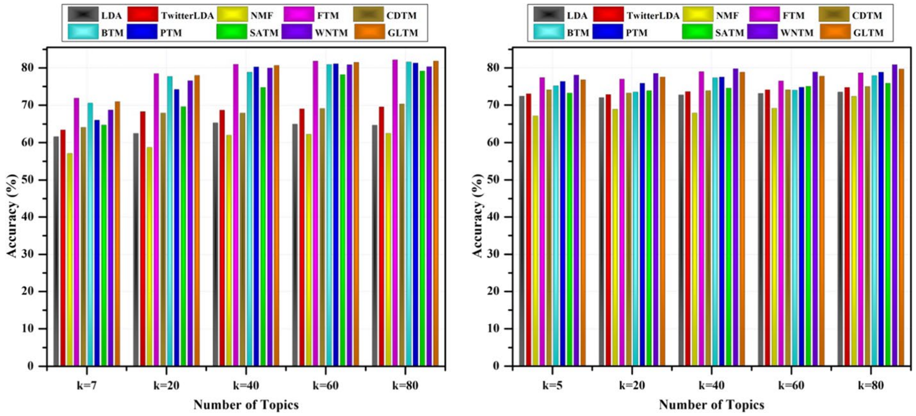
The classification accuracy is used to assess the performance of the considered STTM models. We leverage the extracted topics determined by STTM models as tweet features. Each tweet can be described by document-topic distribution. We adopt the document-topic distribution by $p\left(z_i|d\right)$ for all the considered models. The classification accuracy over both RW-Pand-Twitter and RW-CB-Twitter datasets utilizing the considered existing models LDA, TwitterLDA, NMF, CDTM, and FTM, SATM, BTM, and PTM, WNTM, and GLTM are shown in Fig. 17a and b. Firstly, in the pandemic Twitter (RW-Pand-Twitter) dataset, it can be clearly observed that from Fig. 17a, the FTM has a higher value of accuracy compared to all other models with a different number of topics $k = \{7, 20, 40, 60, and\, 80\}$, while the GLTM has the second high accuracy with all various number of topics. In according to the BTM, it has got the third higher accuracy with $k = \{7, 20, and\, 80\}$ But when $k = 40$ and $k = 60$, the PTM has obtained a higher accuracy than BTM. While SATM is securing the poorly accuracy among all the ASTTM models. NMF has obtained poor accuracy among all the considered models in both ASTTM and TSTTM models. In according to the Real-world cyberbullying (RW-CB-Twitter) dataset, the WNTM model has achieved the best model among all the models, followed by GLTM, as shown in Fig. 17b. The FTM has obtained the third-best model among the ASTTM and TSTTM models, and the PTM model has gained better accuracy than BTM. In conclusion, the models using Global word co-occurrence such as GLTM, WNTM, and BTM performed better performance than the models using self-aggregation such as SATM and PTM as well as better than TwitterLDA, LDA and NMF over the RW-CB-Twitter dataset, indicating their higher levels of efficiency, whereas the FTM has achieved the best model among ASTM and TSTTM models over RW-Pand-Twitter dataset, it gives proposition term weighting and fuzzy clustering which enhances the topic models. The WNTM, GLTM, and BTM have shown their robustness on both datasets. We noted that the self-aggregation models' results, especially the SATM model have not high accuracy compared with Global word co-occurrence such as

GLTM, WNTM, and BTM. Therefore, the effectiveness of self-aggregation-based models is affected by creating long pseudo-documents.

8.4.2 Short text clustering NMI and purity findings

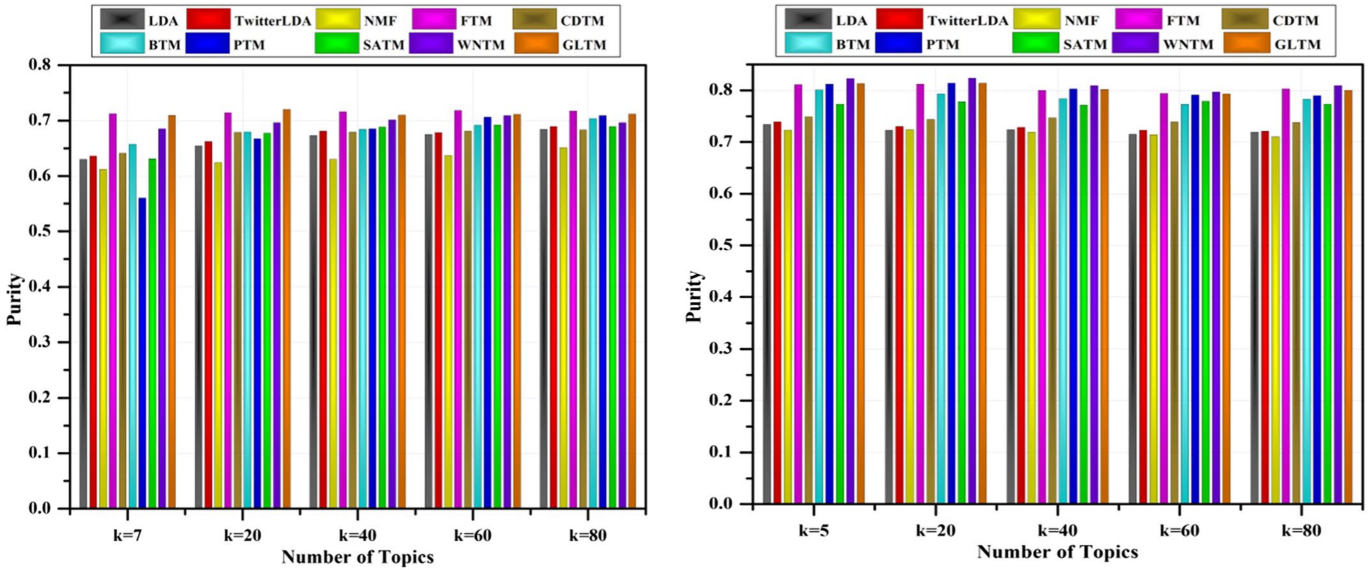
This sub-section illustrates the obtained results of the considered models over both Twitter datasets in terms of purity and NMI metrics.

8.4.2.1 Purity Findings Figure 18a and b show the experimental findings of the considered different TSTTM and ASTTM models in terms of purity to discover the latent topics over both real-world Twitter datasets: RW-Pand-Twitter and RW-CB-Twitter datasets. It can be observed that from Fig. 18a, the FTM model has achieved a better value of purity when compared to other models with a different number of topics over RW-Pand-Twitter. The GLTM has obtained the second-best model among all the TSTTM and ASTTM models, it is considered the best model compared with other Global word co-occurrence based models such as WNTM and BTM and self-Aggregation-based models. The PTM model performs better than BTM with the number of topics $k = \{40, 60, \mathit{and}\, 80\}$ and BTM performs better performance than PTM with $k = \{7, \mathit{and}\, 20\}$ and NMF has got the least purity among all other models for all k different topics over RW-Pand-Twitter as displayed in Fig. 18a. In according to the Cyberbullying twitter dataset(RW-CB-Twitter), the WNTM performs the best performance in terms of purity among all the considered models. FTM and GLTM have achieved second-best and third-best purity compared with other models with all the different k topics. PTM performs better purity than BTM and traditional models LDA, TwitterLDA, CDTM, and NMF. As shown in Fig. 18b, it is evident that the WNTM, GLTM, FTM, and PTM models have higher values of purity when compared with the other traditional NMF, CDTM, LDA and TwitterLDA, indicating their higher levels of efficiency. Finally, we state that the FTM model appears to obtain the best performance in terms of purity over the RW-Pand-Twitter dataset, which is consistent with findings of classification accuracy, positioning itself as the most prominent TSTTM among the considered models with the various k topics. We conclude that the WNTM gains the best performance, followed by GLTM and



(a) RW-Pand-Twitter Dataset **(b)** RW-CB-Twitter Dataset

Fig. 17 Performance evaluation in terms of accuracy on both RW-Pand-Twitter and RW-CB-Twitter datasets with $\mathrm{k} = \{5 \text{ or } 7, 20, 40, 60, \text{ and } 80\}$ topics



(a) RW-Pandemic Dataset

(b) RW-CB-Twitter Dataset

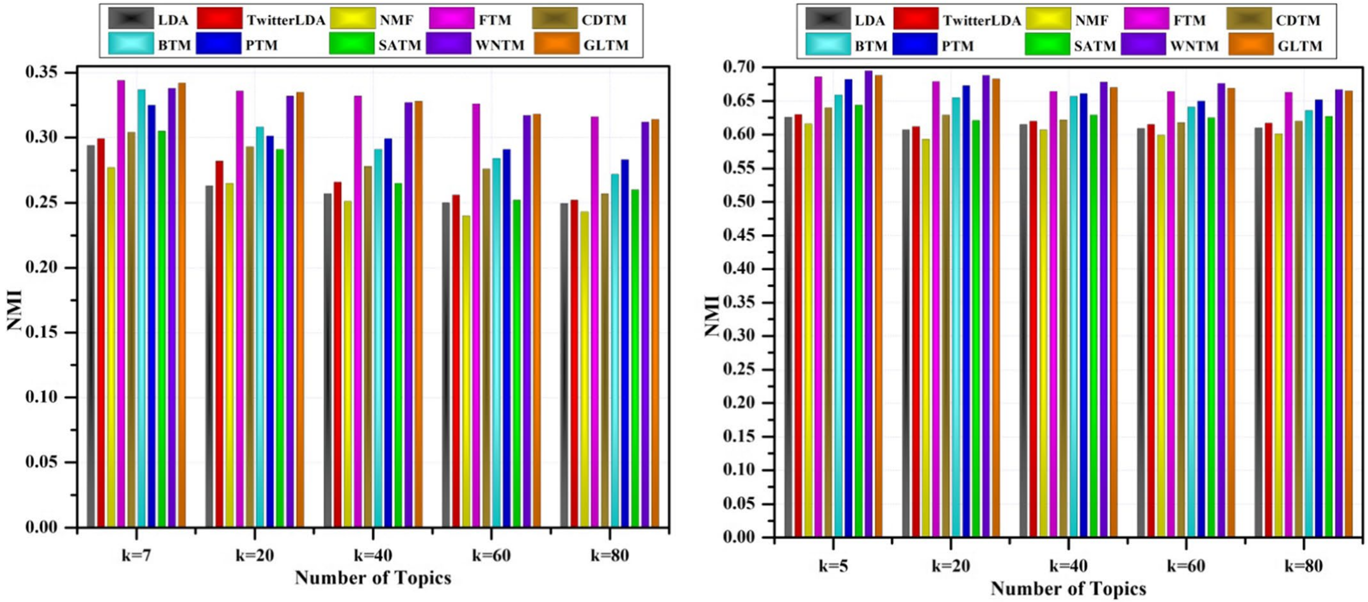
Fig. 18 Performance evaluation in terms of purity on both RW-Pand-Twitter and RW-CB-Twitter datasets with k = {5 or 7, 20, 40, 60, and 80} topics

FTM in terms of clustering evaluation(purity) over the RW-CB-Twitter dataset. In general, the self-aggregation models' results, especially the SATM model, have less purity compared with Global word co-occurrence models. Therefore, the effectiveness of self-aggregation-based models is affected by creating long pseudo-documents. The Global word co-occurrence based models such as WNTM and GLTM show their robustness over both datasets.

8.4.2.2 NMI findings The performance results of the considered models are assessed using the NMI measure. The NMI results of the selected models over both Real-world Twitter datasets: RW-Pand-Twitter and RW-CB-Twitter datasets are shown in Figs. 19a, b. Firstly, the FTM model performs better NMI values than other models for all k different topics $k = \{7, 20, 40, 60, and 80\}$ over the RW-Pand-Twitter dataset. The GLTM and WNTM have achieved the second-best and the third-best performance in terms of NMI compared to other topic modelling models on the RW-Pand-Twitter dataset. The PTM performs better than the BTM model for k different topics except when $k = \{7 and 20\}$, the BTM is better as depicted in Fig. 19 (a). Secondly, according to the performance results of the considered models over the RW-CB-Twitter dataset, it can be observed that from Fig. 19 (b), the WNTM has got the best results in comparison with NMI values of the other considered methods such as GLTM, FTM, BTM, PTM, SATM, CDTM, TwitterLDA, LDA, and NMF. GLTM gains the second-best results compared to all the other models on the RW-CB-Twitter dataset. Here, the PTM outperforms the BTM models with all the k various number of topics $k = \{7, 20, 40, 60, and 80\}$. Finally, we can conclude that the FTM outperforms all the considered ASTTM and TSTTM models over the RW-Pand-Twitter dataset, and the WNTM has gained superior results compared to all the considered other models over the RW-CB-Twitter dataset. In contrast, the SATM has gained a poor NMI value compared to ASTTM models, and NMF has got the least NMI among both the TSTTM and ASTTM models with different k topics over both the RW-Pand-Twitter and RW-CB-Twitter datasets, as depicted in Figs. 19 (a) and 19 (b).

8.4.3 Topic coherence findings

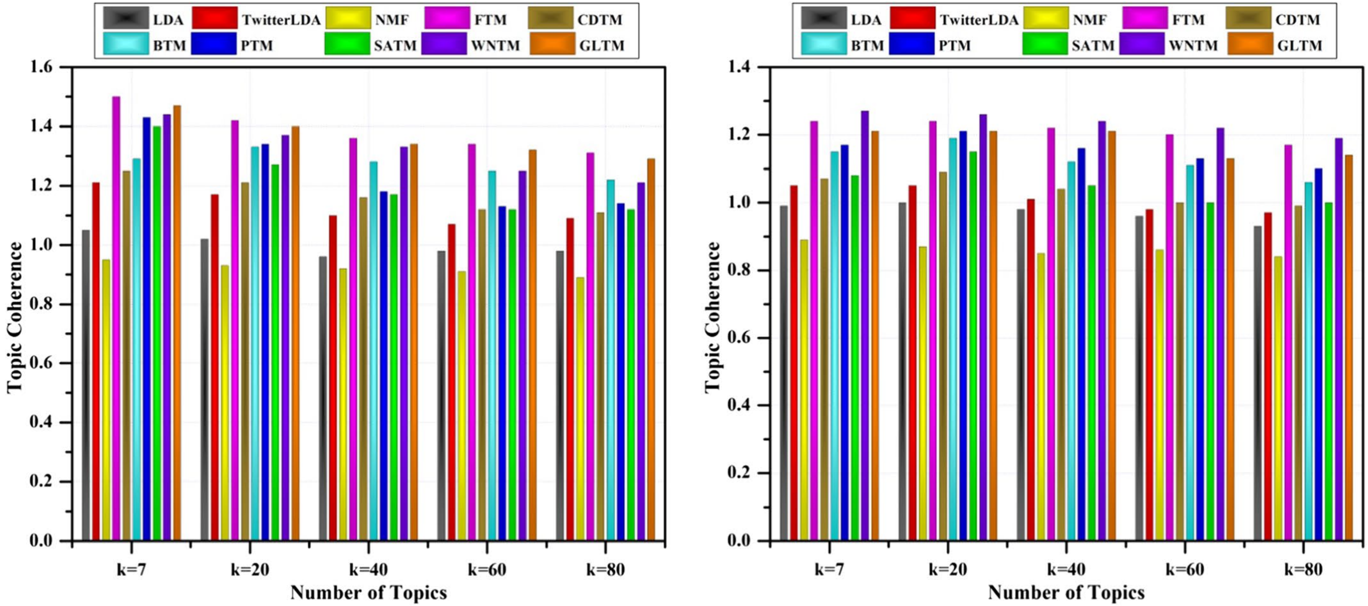
This section compares the quality of latent topics discovered by the considered models using topic coherence. The number of topmost words is fixed to $\mathrm{N} = 10$. We set the number of topics for all the selected STTM models as $\mathrm{k} = \{\#\text{ground truth topics}, 20, 40, 60 \text{ and } 80\}$. Figure 20a and b show the topic coherence of the selected TSTTM and ASTTM models on two datasets: RW-Pand-Twitter and RW-CB-Twitter datasets with a different number of topics. First, on the RW-Pand-Twitter dataset, FTM achieves the best topic coherence with all the k different number of topics $\mathrm{k} = \{7, 20, 40, 60, \text{ and } 80\}$ comparing to other models, which ensures the consistency of results from the purity, NMI and accuracy metrics. The GLTM is the second-best model and superior to WNTM, PTM, BTM, LDA, TwitterLDA, CDTM, and NMF. The third-best model among all the models is WNTM. The BTM outperforms PTM in terms of topic coherence with all k topics except when $\mathrm{k} = \{7 \text{ and } 20\}$ the PTM achieves better results than BTM. In contrast, NMF has the worst coherence in all k topics compared to other STTM models, as depicted in Fig. 20a. These results ensure the consistency of results from the purity, NMI and accuracy metrics. Second, on the real-world cyberbullying Twitter(RW-CB-Twitter) dataset, the WNTM achieves the best topic coherence with all different numbers of topics, and the FTM and GLTM have



(a) RW-Pandemic Dataset

(b) RW-CB-Twitter Dataset

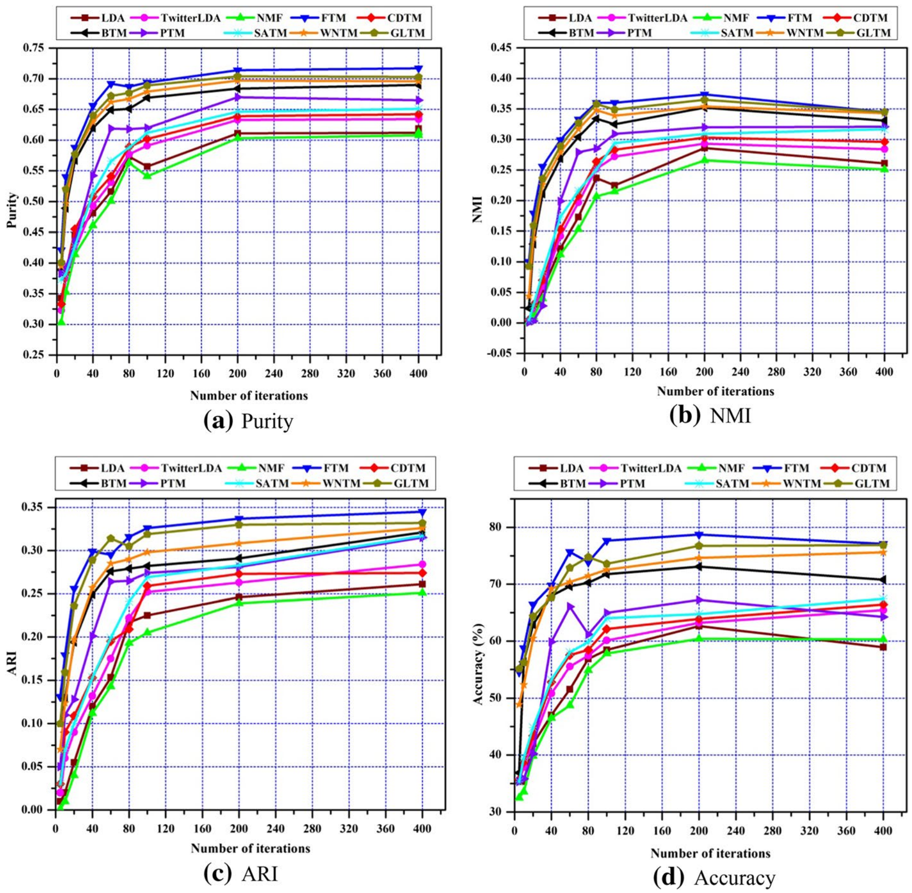
Fig. 19 Performance evaluation in terms of NMI on both RW-Pand-Twitter and RW-CB-Twitter datasets with k = {5 or 7, 20, 40, 60, and 80} topics



(a) RW-Pandemic Dataset

(b) RW-CB-Twitter Dataset

Fig. 20 Performance evaluation in terms of topic coherence on both RW-Pand-Twitter and RW-CB-Twitter datasets with k = {5 or 7, 20, 40, 60, and 80} topics



(a) Purity

(b) NMI

(c) ARI

(d) Accuracy

Fig. 21 Purity, NMI, ARI, and Accuracy values with various numbers of iterations on the RW-Pand-Twitter dataset

achieved the second-best and third-best models compared to all models, respectively, as shown in Fig. 20b. PTM has gained the value of topic coherence better than BTM. This also seems to suggest that the FTM (TSTTM) has significantly better performance when compared with all TSTTM and ASTTM on the RW-Pand-Twitter dataset, and also the modern specific ASTTM models such as WNTM, GLTM, PTM, BTM and SATM have significantly better performance when compared to other TSTTM models such as TwitterLDA, LDA, CDTM, and NMF. Besides, The WNTM has significantly better performance when compared with all TSTTM and ASTTM on the RW-CB-Twitter dataset.

8.4.4 The number of iterations influence on the performance of topic models

In this sub-section, we select the real-world pandemic Twitter (RW-Pand-Twitter) dataset and discuss the effect of iterations number utilizing the Purity, NMI, ARI, and accuracy metrics on the performance of the selected STTM models such as LDA, Twitter-LDA, NMF, FTM, CDTM, BTM, PTM, SATM, WNTM, and GLTM. Figures 21a–d show the effect of the iteration number on the performance of the topic models in terms of Purity, NMI, ARI, and accuracy respectively. From the figures, we observe that when the number of iterations is 5, the values of all metrics are very less while the values of all metrics increase gradually when the number of iterations at 20, 40, 60, 80 etc. That means when the number of iterations increases, the performance of the topic models increases and even reaches convergence, as shown in Fig. 21. On the selected dataset (RW-Pand-Twitter) dataset, FTM has the maximum values of purity, NMI, ARI, and accuracy compared to all the other models. It can be observed that the FTM, GLTM, and WNTM models are superior to all other models in terms of all metrics and secure the first-best, second-best, and third-best value in terms of all metrics, respectively, followed by BTM, PTM, SATM, CDTM TwitterLDA, LDA and NMF. We can notice that Global word co-occurrences based models can get stable and converge at 400 iterations to be near the optimal solutions. Self-aggregation models have the slowest convergence rate and the poorest iterative performance compared with Global word co-occurrences based models.

In conclusion, the FTM model-based TSTTM is a superior model among all TSTTM and ASTTM models over the Real-world pandemic Twitter (RW-Pand-Twitter) dataset in terms of all metrics such as accuracy, NMI, purity, and topic coherence because the FTM model provides fuzzy-clustering and terms weighting which enhances the topic model. Then, the GLTM and WNTM, and BTM models, which are based on simple assumptions, are superior to SATM self-aggregation based models. SATM always outperform NMF, CDTM, LDA, and its improvement (Twitter-LDA). The NMF model achieves poor performance in terms of all metrics when compared to other models. In according to the RW-CB-Twitter dataset, WNTM outperforms all the TSTTM and ASTTM models, and GLTM and FTM have got the second-best and third-best models among all the models. It can be observed here in this dataset that the Global word co-occurrence models such as WNTM and GLTM have got high values of accuracies, NMI, purity and topic coherence compared with the other models. PTM outperforms BTM in terms of all metrics over the RW-CB-Twitter dataset. SATM models have poor values compared with other ASTTM models. The NMF model achieves poor performance in terms of all metrics when compared to other ASTTM and TSTTM models. The Global word co-occurrence models show their robustness on both datasets and indicate their higher levels of efficiency. The self-aggregation models' findings, especially the SATM model, have less values of purity, NMI, accuracy, and topic coherence compared with Global word co-occurrence models. Therefore, the effectiveness of self-aggregation-based models is affected by creating long pseudo-documents. The structure of SATM and PTM are very complex, and there are numerous hidden variables that request to be sampled, resulting in significant time consumption.

9 Discussions

This section discusses the overall observations noted from the qualitative and quantitative analysis, as well as the comparative analysis.

Qualitative analysis The qualitative analysis highlighted some observations related to existing TSTTM and ASTTM models. According to the existing TSTTM models, it can be observed that from Tables 2, 3, and 4, some of the works tried to increase the accuracy, such as (Wang et al. 2012; Fang et al. 2017; Kumar and Vardhan 2019; Sharath et al. 2019; Valdez et al. 2018; Belford et al. 2016; Yan et al. 2012; Muliawati and Murfi 2017; Capdevila et al. 2017; and Li et al. 2015), enhance the coherence such as (Zhao et al. 2011; Zheng et al. 2019; Han et al. 2020; Kim et al. 2020; Farahat et al. 2015; Lacoste-Julien et al. 2009; and He et al. 2019b), alleviate the data sparsity problem as in (Ozyurt and Akcayol 2021; Akhtar et al. 2019b; Iskandar 2017; and Pang et al. 2019), extract the topics from the unlabeled data (Korshunova et al. 2019; Ozyurt and Akcayol 2021). Other existing works resolved the lack of semantic information and local word co-occurrence (Akhtar et al. 2019a; Chen and Kao 2017; Chen et al. 2020b). However, the existing works addressed some of the above challenges. There are still issues in this sub-category, such as time complexity and sparsity, whereas the others are incapable of handling large scale of short text data.

On the other hand, it can be observed that from Tables 7, 8, 9, and 10, the efficiency of the some of ASTTM models is better than the TSTTM, and also there are TSTTM model is better than ASTTM as concluded from the experimental analysis that the FTM models under TSTTM category is superior than BTM, PTM, GLTM, and WNTM over RW-Pand-Twitter dataset. The ASTTM models attempt to resolve the common issues in the short text, such as improving classification accuracy and increasing topic coherence. It can be noted that the accuracy increases by mixture components in EM as in the DMM category. This strategy suggested to sample one document or short text by one topic, and it is suited for short text to increase the accuracy issue as in (Yin and Wang 2014; Yu and Qiu 2019; Yu and Qiu 2019; Li et al. 2019c; Liu et al. 2020b; and Garcia and Berton 2021). Moreover, due to the nature of the short text, there is not enough word Co-occurrence information; hence, some models such as BTM (Cheng et al. 2014; Pang et al. 2016; and Huang et al. 2020) attempt to utilize the wealthy Global word co-occurrence pattern to infer hidden topics. In contrast, the other models tried to alleviate the sparse of a short text by aggregating the short text into long pseudo-documents to discover latent topics, e.g. PTM (Zuo, Wu, et al. 2016a), SATM (Quan et al. 2015), WE-PTM (Zuo et al. 2021), and SenU-PTM(Lu et al. 2021). It can be concluded that the short text is still suffering from the data sparsity problem, time complexity, visualizing the topics, and data quality problems (the noise of data) in social media. Since the tweets data are informal in nature, it contains slang, typos, Elongated (repeated Characters), transposition, Concatenated words, and complex spelling mistakes, such as unorthodox use of acronyms. Therefore, researchers are advised to work and address these issues for future works.

Quantitative Analysis The quantitative analysis highlighted some observations regarding the publications related to STTM. According to the publications of TSTTM based on the timeline, the higher rate of publications was in 2019, but in the case of ASTTM, 2020 was the most productive year. On the other side, TSTTM probabilistic based have scored a higher rate of publications, whereas Deep Learning Topic modelling (DLTM) based models were the most productive sub-category among the ASTTM categories, followed by Global word co-occurrences which is the second productive sub-category. Contrary, the self-aggregation based had been given less attention by the researchers. Coming to the social media data source utilized by STTM, Twitter has given more attention from researchers. Whereas other social media data sources such as Facebook, Instagram, Tik Tok, and WhatsApp have not gotten attention from the researchers. This gives an indication to the research to work on such platforms (Facebook, Instagram, Tik Tok, and WhatsApp). To the best of our knowledge, these platforms have gotten less attention due to the limited access to such social media data sources. Twitter facilitates access to data through the Twitter API Streaming. Therefore, researchers can create new datasets from these social media platforms and make them publicly available for the researchers to draw their attention to work on these platforms. Furthermore, real-world data (streaming data) in social media has been given less attention. Therefore, researchers are advisable to evaluate the STTM models on the real-world dataset to extract the trending topics and discover the emerging latent topics of discussion from constant background chatter in social media. Finally, we observed from both Fig. 12a and b that Spark has given less attention, where researchers can consider the spark streaming for further work, as it has the capability to process huge datasets and process the data parallelly with less computational time. On the other side, Fig.14a and b highlight that the English language has been given more attention than other languages such as Chinese, Arabic, Hindi, etc. Therefore, researchers are advised to work on such other languages.

Comparative Analysis The comparative analysis highlighted some observations from the experimental results obtained for the selected TSTTM and ASTTM models. This subsection briefly discusses these observations as follows:

Topic coherence, purity, NMI, and accuracy evaluation metrics: The conclusion of the Comparative Analysis is that the FTM model-based TSTTM is a superior model among all TSTTM and ASTTM models over the Real-world pandemic Twitter (RW-Pand-Twitter) dataset in terms of all metrics accuracy, NMI, purity, and topic coherence because the FTM model provides fuzzy-clustering and terms weighting which enhances the topic model, indicating its high level of efficiency. Then, the GLTM, WNTM, and BTM models, which are based on simple assumptions, are superior to SATM self-aggregation based models. SATM always outperform NMF, CDTM, LDA, and its improvement (Twitter-LDA). The NMF model achieves poor performance in terms of all metrics when compared to other models. In according to the RW-CB-Twitter dataset, WNTM outperforms all the TSTTM and ASTTM models, and GLTM and FTM have got the second-best and third-best models among all the models. It can be observed that in this dataset, the Global word co-occurrence models such as WNTM and GLTM have got high values of accuracies, NMI, purity and topic coherence compared with the other models. PTM outperforms BTM in terms of all metrics over the RW-CB-Twitter dataset. SATM models have poor values compared with other ASTTM models. The NMF model achieves poor performance in terms of all metrics when compared to other ASTTM and TSTTM models. The Global word co-occurrence based models show their robustness on both datasets and indicate their higher levels of efficiency. The self-aggregation models' findings, especially the SATM model, have less values of purity, NMI, accuracy, and topic coherence compared with Global word co-occurrence models. Therefore, the effectiveness of self-aggregation-based models is affected by creating long pseudo-documents. The structure of SATM and PTM are very complex, and there are numerous hidden variables that request to be sampled resulting in significant time consumption.

The number of iterations it was noted that when the number of iterations is 5, the values of all metrics are significantly less, while the values of all metrics increase gradually when the number of iterations at 20, 40, 80, 120, etc. However, When the number of iterations increases, the performance of the topic modelling increases, as shown in Fig. 21. For purity, NMI, ARI and Accuracy, FTM has the maximum values in all models when the number of iterations at 200. The FTM model was found to be superior to all other models in terms of all metrics over the RW-Pand-Twitter dataset, whereas the GLTM and WNTM have achieved the second-best and the third-best performance compared with all other models over. In according to RW-CB-Twitter Dataset, the WNTM model was the best model in comparison with all other models over the RW-CB-Twitter dataset in terms of the majority of the considered evaluation metrics. The BTM and PTM models have the tradeoff values in terms of all metrics over both datasets, followed by SATM, CDTM, TwitterLDA, LDA, and NMF.

10 Challenges and future directions

From the above deep discussions, it can be observed that issues such as data sparsity, data noise, evaluation metrics, visualization, and deep learning were given less attention by the research community. This section briefly highlights the aforesaid challenges and open research issues in STTM that can help in further research enhancement to STTM.

Data sparsity LDA, NMF, PLSA, and LSA are some of the key unsupervised techniques that extract topics from a set of documents texts solely using the content of documents. Many extensions were developed for supporting short texts. But many issues are still left unaddressed. The models which are entirely based on short text and tweet contents are still suffering from the data sparsity problem (Likhitha et al. 2019). The co-occurrence density of the matrix of words in a collection of tweets can be as low as 0.274% on average (Nugroho et al. 2017), which results in low overlapping terms with less semantic relationships.

Data quality problem Augmentation of short texts in TSTTM and ASTTM models with auxiliary content is mostly from external sources might seem efficient. However, they always result in noise interference and sometimes also cause unrelated term selection. Since the tweets are informal in nature, slang, typos, Elongated (repeated Characters), transposition, Concatenated words, complex spelling mistakes, such as unorthodox use of acronyms, word boundary errors, manifold forms of abbreviations of the same words and grammar are challenging for the learning models and matching with the content of external resources. Further, these external sources can also cause stability issues due to increased usage of resources for handling external source data. Many models use the hashtags feature in tweets to extract and ameliorate topics learned for topic discovery (Alash and Al-Sultany 2020). Our earlier work (Murshed et al. 2021) has addressed the data quality problem of the Twitter social media datasets. However, since not all tweets have hashtags, the sparsity problem continues. Further, some models require user information to track the topics. But, in real-time, a Twitter user can post tweets on many topics, where the user mentions are not included in most of these tweets. In a dynamic twitter environment, user-related information is prone to privacy issues.

Visibility Most models in STTM fail to consider time as an important feature in extracting topics. Even methods with temporal features use time only as a timing window for topic extraction. However, it is important to identify the time factor as a quality factor to be improved as it derives the static tweet content for dynamic processing. Advanced models such as DMM, BTM, PTM, WNTM and SATM have been developed in recent years and are not being widely used owing to limited visibility of lack of exposure. However, they seem to provide better performance when compared to the traditional models. In order to enhance their usage for STTM, the limitations discussed in Sect. 4.3 must be addressed. It is also important to process complex data so that they can be used to support big data sets (large-scale datasets).

Visualization Topics are displayed by the most frequent terms/words for each topic (Chuang et al. 2012; Qiang et al. 2020). How to view a document or short text utilizing topic models (TM) is also a challenging issue. Topic modelling is a process of automatically discovering the hidden thematic structure from the short text, such as posts or tweets and facilitates building new ways to browse and summarize the large archive of text as topics. This structure can help to define the most important sections of the documents by linking to the topic labels. Visualizing a considerable number of topics (for example, more than 100 topics) in a compressed way is a challenging issue (Sievert and Shirley 2014).

Streaming data Concept drift and recurring concepts are two characteristics of streaming data. Most available models do not adequately handle such scenarios. Another future research direction, training TMs can be achieved through active learning (Burkhardt and Kramer 2019a). Specifically, when dealing with text data streams, labelling all incoming new documents by hand is a hard task. Active learning allows for aggressively selecting documents that differ from recently viewed documents or where the algorithm has the least confidence during labelling and infers labels for the remainder. Extending Semi-supervised may also help in better training with less labelled training data (Burkhardt et al. 2020; Burkhardt et al. 2018).

Deep learning Although the literature reported that several existing TMs were trained using neural networks (Burkhardt and Kramer 2019b), the research on multi-label classification using NTM is still limited (Panda et al. 2019). This necessitates the use of advanced DL methods such as recurrent networks, convolutions, and various prior distributions. For instance, dropping the assumption of a mixture model, that all documents are mixtures of topics, does not increase the complexity of NTM (Srivastava and Sutton 2017). This allows each document to be represented by various combinations of topics, which makes the model more expressive. Furthermore, neural networks can be more easily extended by using word vectors or other layer types, which can be pre-trained to extract semantics and syntactic attributes of words(Burkhardt and Kramer 2019a).

Evaluation metrics Purity and NMI, and recall metrics are the three most popularly utilized to assess the quality of extracted topics. These metrics are dependent on the statistical accuracy of clustering outcomes. Due to the great data sparsity of correlation among words/terms, a numerical examination of the coherency among terms/words in the representation of the topic may not yield a consistent outcome for estimation purposes. Hence, it is necessary to use more suitable and efficient metrics. Therefore, PMI, Topic coherence and measures have been used in the experiments section. Beneficial assessment metrics have never been resolved for topic discovery models (Qiang et al. 2020). Topic coherence cannot differentiate between topics. Moreover, only one part of the topic modelling models is evaluated by current metrics. Developing more such new evaluation criteria is a future research work for topic modelling.

Optimization for STTM From the comprehensive review and extensive analysis, the limitations of existing STTM models are regarded as critical for accurately representing current knowledge on the topic. While there are numerous significant researches for STTM, modelling refinement and optimization are still required for optimizing the accuracy and generated outputs. Meta-heuristic optimization algorithms can be a good tool for achieving high accuracy and optimized results.

Comparative analysis Comparing short text topic modeling methods in terms of how many sub-tasks are there in short text topic modeling, the methods designed to address the sub-tasks, and the general processing framework of the method for each sub-task.

11 Conclusion

This article presented a comprehensive survey and comparative analysis along with an extended structured taxonomy for the most recent, ever-growing efficient STTM models in social media. It mainly focused on most aspects of TSTTM models such as probabilistic models, matrix factorization, unsupervised and supervised models, Exemplar based, dynamic-based categories, data source based, word types, application-based, Frequent

Pattern Mining (FPM) techniques). Moreover, it provided ASTTM models such as DMM-based, Global word co-occurrences-based, self-aggregation-based, and deep learning topic modelling models. The taxonomy provided a qualitative analysis of existing STTM models based on their performance and respective strengths and weaknesses. The utilized datasets by STTM were reviewed and analyzed quantitatively. The useful software tools and open-sources libraries for STTM were reviewed and summarized.

Furthermore, a quantitative analysis of the literature was performed to highlight the research trends and future directions. Moreover, a comparative analysis of the topic quality and performance of representative STTM models is presented. The performance evaluation is conducted on two real-world Twitter datasets: RW-Pand-Twitter and RW-CB-Twitter datasets in terms of several metrics such as topic coherence, purity, NMI, and accuracy. Finally, the open challenges and future research directions in this promising field are discussed. To the best of our knowledge, the findings of this study will serve as a catalyst to develop new and efficient models for topic discovery of short texts overcoming all the limitations of the existing STTM models. The final suggestion offered by this study is the development of the STTM method incorporating all vital features without increasing the complexity, which would be a viable solution for the future.

Acknowledgements The first author is grateful to the Department of Studies in Computer Science and the chairman of the Department Dr. D. S. Guru, Professor at the University of Mysore in India for all the facilities offered for this research. Besides, the first author wishes to acknowledge the Department of Computer science, University of Amran, Amran, yemen, for their support and all the facilities provided for this research work. A Special thanks and gratitude to my brother Dr. Samir Abdullah Hizam [ENT consultant] for his constant support financially and morally. The help of Dr. Nabil Abdullah Al-bahili is highly appreciated.

References

Abdel-Hafez A, Yue Xu (2013) A survey of user modelling in social media websites. Comput Inf Sci 6(4):59–71. https://doi.org/10.5539/cis.v6n4p59

Abdulwahab HM, Ajitha S, Saif MAN (2022) Feature selection techniques in the context of big data: taxonomy and analysis. Appl Intell. https://doi.org/10.1007/s10489-021-03118-3

Abou-Of MA (2020) A fuzzy, incremental and semantic trending topic detection in social feeds. In: 2020 11th international conference on information and communication systems (ICICS). IEEE, pp 118–24

Ahmed A, Aly M, Gonzalez J, Narayanamurthy S, Smola AJ (2012) Scalable inference in latent variable models. In: Proceedings of the fifth ACM international conference on Web search and data mining—WSDM '12. ACM Press, New York, pp 123–32

Aiello LM, Petkos G, Martin C, Corney D, Papadopoulos S, Skraba R, Goker A, Kompatsiaris I, Jaimes A (2013) Sensing trending topics in Twitter. IEEE Trans Multimed 15(6):1268–1282. https://doi.org/10.1109/TMM.2013.2265080

Akhtar N (2017) Hierarchical summarization of news Tweets with Twitter-LDA. In: Applications of soft computing for the web. Springer, Singapore, pp 83–98

Akhtar N, Sufyan Beg MM (2019a) User graph topic model. J Intell Fuzzy Syst 36(3):2229–2240. https://doi.org/10.3233/JIFS-169934

Akhtar N, Sufyan Beg MM, Javed H (2019b) Topic modelling with fuzzy document representation. In: Singh M, Gupta PK, Tyagi V, Flusser J, Ören T, Kashyap R (eds) Advances in computing and data sciences. ICACDS 2019b. Communications in Computer and Information Science, vol 1046. Springer, Singapore, pp 577–87

Al-Sultany GA, Aleqabie HJ (2019) Enriching tweets for topic modeling via linking to the wikipedia. Int J Eng Technol 8(15):144–150

Alash HM, Al-Sultany GA (2020) improve topic modeling algorithms based on twitter hashtags. J Phys 1660:012100. https://doi.org/10.1088/1742-6596/1660/1/012100

Albalawi R, Yeap TH, Benyoucef M (2020) Using topic modeling methods for short-text data: a comparative analysis. Front Artif Intell 3:1–14. https://doi.org/10.3389/frai.2020.00042

Aletras N, Stevenson M (2013) Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th international conference on computational semantics, IWCS 2013—Long Papers, pp 13–22

Alghamdi R, Alfalqi K (2015) A survey of topic modeling in text mining. Int J Adv Comput Sci Appl 6(1):147–153. https://doi.org/10.14569/IJACSA.2015.060121

Ali IMS, Balakrishnan M (2021) Population and global search improved squirrel search algorithm for feature selection in big data classification. Int J Intell Eng Syst 14(4):177–189. https://doi.org/10.22266/ijies2021.0831.17

Anil Phand S, Chakkarwar VA (2018) Enhanced sentiment classification using geo location tweets. In: Proceedings of the 2nd international conference on inventive communication and computational technologies, ICICCT 2018. IEEE, pp 881–86

Belford M, Mac Namee B, Greene D (2016) Ensemble topic modeling via matrix factorization. In: 24th Irish conference on artificial intelligence and cognitive science (AICS'16), vol 1751, Dublin, Ireland, 20–21 September 2016, CEUR Workshop Proceedings, pp 21–32

Bhadury A, Chen J, Zhu J, Liu S (2016). Scaling up dynamic topic models. In: Proceedings of the 25th international conference on world wide web. Republic and Canton of Geneva, International World Wide Web Conferences Steering Committee, Switzerland, pp 381–90

Bhattacharya P, Zafar MB, Ganguly N, Ghosh S, Gummadi KP (2014) Inferring user interests in the twitter social network. In: Proceedings of the 8th ACM conference on recommender systems. ACM Press, New York, pp 357–360

Bianchi F, Terragni S, Hovy D (2021) Pre-training is a hot topic: contextualized document embeddings improve topic coherence. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, vol 2: Short Papers. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 759–66

Bicalho P, Pita M, Pedrosa G, Lacerda A, Pappa GL (2017) A general framework to expand short text for topic modeling. Inf Sci 393:66–81. https://doi.org/10.1016/j.ins.2017.02.007

Blair SJ, Bi Y, Mulvenna MD (2020) Aggregated topic models for increasing social media topic coherence. Appl Intell 50(1):138–156. https://doi.org/10.1007/s10489-019-01438-z

Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd international conference on `11`Machine learning—ICML '06, vol 148. ACM Press, New York, pp 113–2

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Bougteb, Y, Ouhbi B, Frikh B, Zemmouri EM (2019) Deep learning based topics detection. In: 2019 Third international conference on intelligent computing in data sciences (ICDS). IEEE, pp 1–7 (2019)

Burkhardt S, Kramer S (2019a) A survey of multi-label topic models. ACM SIGKDD Explor Newsl 21(2):61–79. https://doi.org/10.1145/3373464.3373474

Burkhardt S, Kramer S (2019b) Decoupling sparsity and smoothness in the Dirichlet variational autoencoder topic model. J Mach Learn Res 20:1–27

Burkhardt S, Siekiera J, Kramer S (2018) Semi-supervised bayesian active learning for text classification. In: Bayesian deep learning workshop at NeurIPS (NeurIPS)

Burkhardt S, Siekiera J, Glodde J, Andrade-Navarro MA, Kramer S (2020) Towards identifying drug side effects from social media using active learning and crowd sourcing. In: Pacific symposium on biocomputing. World Scientific, pp 319–330

Cao B, Liu X, Liu J, Tang M (2017) Domain-aware mashup service clustering based on lda topic model from multiple data sources. Inf Softw Technol 90:40–54. https://doi.org/10.1016/j.infsof.2017.05.001

Capdevila J, Cerquides J, Nin J, Torres J (2017) Tweet-SCAN: an event discovery technique for geo-located tweets. Pattern Recogn Lett 93:58–68. https://doi.org/10.1016/j.patrec.2016.08.010

Card D, Tan C, Smith NA (2018) Neural models for documents with metadata. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol 1: Long Papers. Association for computational linguistics, Stroudsburg, PA, USA, pp 2031–2040

Casalino G, Castiello C, Del Buono N, Mencar C (2018) A framework for intelligent twitter data analysis with non-negative matrix factorization. Int J Web Inf Syst 14(3):334–356. https://doi.org/10.1108/IJWIS-11-2017-0081

Chan WN (2020) Development of a real-time social big data analytics system using topic modeling. Int J Comput Sci Inf Secur 18(4):27–31

Chang MW, Ratinov L, Roth D, Srikumar V (2008) Importance of semantic representation: dataless classification. In: Proceedings of the national conference on artificial intelligence, vol 2, pp 830–35

Chen GB, Kao H-Y (2017) Word co-occurrence augmented topic model in short text. Intell Data Anal 21(S1):S55-70. https://doi.org/10.3233/IDA-170872

Chen Y, Zhang H, Liu R, Ye Z, Lin J (2019) Experimental explorations on short text topic mining between LDA and NMF based schemes. Knowl-Based Syst 163:1–13. https://doi.org/10.1016/j.knosys.2018.08.011

Chen J, Gong Z, Liu W (2020a) A Dirichlet process biterm-based mixture model for short text stream clustering. Appl Intell 50(5):1609–1619. https://doi.org/10.1007/s10489-019-01606-1

Chen Y, Junjie Wu, Lin J, Liu R, Zhang H, Ye Z (2020b) Affinity regularized non-negative matrix factorization for lifelong topic modeling. IEEE Trans Knowl Data Eng 32(7):1249–1262. https://doi.org/10.1109/TKDE.2019.2904687

Cheng X, Yan X, Lan Y, Guo J (2014) BTM: topic modeling over short texts. IEEE Trans Knowl Data Eng 26(12):2928–2941. https://doi.org/10.1109/TKDE.2014.2313872

Cheng L, Li J, Silva Y, Hall D, Liu H (2019) PI-bully: personalized cyberbullying detection with peer influence. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence. vol 2019-Augus. International Joint Conferences on Artificial Intelligence Organization, California, pp 5829–35

Choi H-J, Park CH (2019) Emerging topic detection in twitter stream based on high utility pattern mining. Expert Syst Appl 115:27–36. https://doi.org/10.1016/j.eswa.2018.07.051

Chuang J, Manning CD, Heer J (2012) Termite: visualization techniques for assessing textual topic models. In: Proceedings of the international working conference on advanced visual interfaces, ACM. ACM Press, pp 74–77

Chuluunsaikhan T, Ryu G-A, Yoo K-H, Rah H, Nasridinov A (2020) Incorporating deep learning and news topic modeling for forecasting pork prices: the case of South Korea. Agriculture 10(11):513. https://doi.org/10.3390/agriculture10110513

Cortis K, Handschuh S (2015) Analysis of cyberbullying tweets in trending world events. In: Proceedings of the 15th international conference on knowledge technologies and data-driven business, vols 21–22-Octo. ACM, New York, NY, USA, pp 1–8

Cotelo JM, Cruz FL, Troyano JA (2014) Dynamic topic-related tweet retrieval. J Am Soc Inf Sci 65(3):513–523. https://doi.org/10.1002/asi.22991

Curiskis SA, Drake B, Osborn TR, Kennedy PJ (2020) An evaluation of document clustering and topic modelling in two online social networks: twitter and reddit. Inf Process Manag 57(2):102034. https://doi.org/10.1016/j.ipm.2019.04.002

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407

Dey K, Shrivastava R, Kaushik S (2018) Topical stance detection for twitter: a two-phase lstm model using attention. In: European conference on information retrieval, LNCS 10772, pp 529–536

Diao Y, Du Y, Xiao P, Liu J (2017) A CWTM model of topic extraction for short text. In: China conference on knowledge graph and semantic computing (CCKS 2017), communications in computer and information science (CCIS 784). Springer, Singapore, pp 80–91

Dieng AB, Ruiz FJR, Blei DM (2020) Topic modeling in embedding spaces. Trans Assoc Comput Linguist 8:439–453. https://doi.org/10.1162/tacl_a_00325

Doan T-N, Hoang T-A (2021) Benchmarking neural topic models: an empirical study. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 4363–68

Dutta L, Maji G, Sen S (2020) A study on spatiotemporal topical analysis of twitter data. In: JKM, Bhattacharya D (eds) Emerging technology in modelling and graphics, vol 937, Advances in intelligent systems and computing. Springer, Singapore, pp 699–711

Earle PS, Bowden DC, Guy M (2011) Twitter earthquake detection: earthquake monitoring in a social world. Ann Geophys 54(6):708–715. https://doi.org/10.4401/ag-5364

Ediger D, Jiang K, Riedy J, Bader DA, Corley C (2010) Massive social network analysis: mining twitter for social good. In: 2010 39th international conference on parallel processing. IEEE, pp 583–593

Elbagoury A, Ibrahim R, Farahat AK, Kamel MS, Karray F (2015) Exemplar-based topic detection in twitter streams. In: Proceedings of the 9th international conference on web and social media (ICWSM), pp 610–613.

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The Pascal Visual Object Classes (VOC) challenge. Int J Comput vis 88(2):303–338. https://doi.org/10.1007/s11263-009-0275-4

Fang Y, Zhang H, Ye Y, Li X (2014) Detecting hot topics from twitter: a multiview approach. J Inf Sci 40(5):578–593. https://doi.org/10.1177/0165551514541614

Fang A, Macdonald C, Ounis I, Habel P (2016a) Examining the coherence of the top ranked tweet topics. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA. ACM, pp 825–828

Fang A, Macdonald C, Ounis I, Habel P (2016b) Using word embedding to evaluate the coherence of topics from twitter data. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA. ACM, pp 1057–1060

Fang A, Macdonald C, Ounis I, Habel P, Yang X (2017) Exploring time-sensitive variational bayesian inference LDA for social media data. In: European conference on information retrieval, Lecture Notes in Computer Science. Springer, Cham, pp 252–265

Farahat AK, Elgohary A, Ghodsi A, Kamel MS (2015) Greedy column subset selection for large-scale data sets. Knowl Inf Syst 45(1):1–34. https://doi.org/10.1007/s10115-014-0801-8

Feng L (2018) Topic Modeling of environmental data on social networks based on ED-LDA. Int J Environ Monit Anal 6(3):77–83. https://doi.org/10.11648/j.ijema.20180603.12

Feng J, Rao Y, Haoran Xie Fu, Wang L, Li Q (2020a) User group based emotion detection and topic discovery over short text. World Wide Web 23(3):1553–1587. https://doi.org/10.1007/s11280-019-00760-3

Feng J, Zhang Z, Ding C, Rao Y, Xie H (2020b) Context reinforced neural topic modeling over short texts. ArXiv Preprint arXiv:abs/2008.04545

Gao C, Zeng J, Lyu MR, King I (2018) Online app review analysis for identifying emerging issues. In: Proceedings of the 40th international conference on software engineering, Ser. ICSE 18. Association for Computing Machinery, New York, NY, USA, pp 48–58. https://doi.org/10.1145/3180155.3180218.

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. https://doi.org/10.1007/s10115-018-1314-7

Garcia K, Berton L (2021) Topic detection and sentiment analysis in twitter content related to COVID-19 from Brazil and the USA. Appl Soft Comput 101:107057. https://doi.org/10.1016/j.asoc.2020.107057

Ge B, Zheng W, Yang GM, Lu Y, Zheng HJ (2019) Microblog topic mining based on a combined TF-IDF and LDA topic model. In: Automatic Control, Mechatronics and Industrial Engineering: Proceedings of the international conference on automatic control, mechatronics and industrial engineering (ACMIE 2018). CRC Press, Suzhou, China, pp 291–296

Ghoorchian K, Sahlgren M (2020) GDTM: graph-based dynamic topic models. Prog Artif Intell 9(3):195–207. https://doi.org/10.1007/s13748-020-00206-2

Gui L, Leng J, Pergola G, Zhou Y, Xu R, He Y (2019) Neural Topic Model with Reinforcement Learning. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 3476–3481

Günther F, Dudschig C, Kaup B (2014) LSAfun—an R package for computations based on latent semantic analysis. Behav Res Methods 47(4):930–944. https://doi.org/10.3758/s13428-014-0529-0

Guo J, Zhang P, Tan J, Guo L (2012) Mining hot topics from twitter streams. Procedia Comput Sci 9:2008–2011. https://doi.org/10.1016/j.procs.2012.04.224

Gupta P, Chaudhary Y, Schütze H (2019) Multi-view and multi-source transfers in neural topic modeling with pretrained topic and word embeddings. ArXiv Preprint arXiv:abs/1909.06563

Ha C, Tran V-D, Van LN, Than K (2019) Eliminating overfitting of probabilistic topic models on short and noisy text: the role of dropout. Int J Approx Reason 112:85–104. https://doi.org/10.1016/j.ijar.2019.05.010

Hadi MA, Fard FH (2020) AOBTM: adaptive online biterm topic modeling for version sensitive short-texts analysis. In: 2020 IEEE international conference on software maintenance and evolution (ICSME). IEEE, pp 593–604

Han W, Tian Z, Huang Z, Li S, Jia Y (2020) Topic representation model based on microblogging behavior analysis. World Wide Web 23(6):3083–3097. https://doi.org/10.1007/s11280-020-00822-x

Hasan M, Orgun MA, Schwitter R (2018) A survey on real-time event detection from the twitter data stream. J Inf Sci 44(4):443–463. https://doi.org/10.1177/0165551517698564

He R, Zhang X, Jin D, Wang L, Dang J, Li X (2018) Interaction-aware topic model for microblog conversations through network embedding and user attention. In: Proceedings of the 27th international conference on computational linguistics. Santa Fe, New Mexico, USA, pp 1398–1409

He J, Li L, Wang Y, Xindong Wu (2020a) Targeted aspects oriented topic modeling for short texts. Appl Intell 50(8):2384–2399. https://doi.org/10.1007/s10489-020-01672-w

He J, Liu H, Zheng Y, Tang S, He W, Xiaoyong Du (2020b) Bi-labeled LDA: inferring interest tags for non-famous users in social network. Data Sci Eng 5(1):27–47. https://doi.org/10.1007/s41019-019-00113-0

Hennig L (2009) Topic-based multi-document summarization with probabilistic latent semantic analysis. In: Proceedings of the international conference recent advances in natural language processing (RANLP-2009), pp 144–149

Hidayatullah AF, Aditya SK, Gardini ST (2019) Topic modeling of weather and climate condition on twitter using Latent Dirichlet Allocation (LDA). IOP Conf Ser 482(1):012033. https://doi.org/10.1088/1757-899X/482/1/012033

Hoffman MD, Blei DM, Bach F (2010) Online learning for latent Dirichlet allocation. In: Proceedings ofthe 23rd international conference on neural information processing systems, ser. NIPS10. Red Hook., vol 1. Curran Associates Inc., NY, USA, p 856864

Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, vol 99, pp 50–57

Hong L, Dom B, Gurumurthy S, Tsioutsiouliklis K (2011) A time-dependent topic model for multiple text streams. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM Press, New York, New York, USA, pp 832–840

Hua T, Chang-Tien Lu, Choo J, Reddy CK (2020) Probabilistic topic modeling for comparative analysis of document collections. ACM Trans Knowl Discov Data 14(2):1–27. https://doi.org/10.1145/3369873

Huang J, Peng M, Li P, Zhiwei Hu, Chao Xu (2020) Improving biterm topic model with word embeddings. World Wide Web 23(6):3099–3124. https://doi.org/10.1007/s11280-020-00823-w

Huang R, Guan Yu, Wang Z, Zhang J, Shi L (2013) Dirichlet process mixture model for document clustering with feature partition. IEEE Trans Knowl Data Eng 25(8):1748–1759. https://doi.org/10.1109/TKDE.2012.27

Huang L, Ma J, Chen C (2017) Topic detection from microblogs using T-LDA and perplexity. In: 2017 24th asia-pacific software engineering conference workshops (APSECW). IEEE, pp 71–77

Ibrahim R, Elbagoury A, Kamel MS, Karray F (2018) Tools and approaches for topic detection from twitter streams: survey. Knowl Inf Syst 54(3):511–539. https://doi.org/10.1007/s10115-017-1081-x

Indra EW, Pulungan R (2019) Trending topics detection of indonesian tweets using BN-grams and Doc-P. J King Saud Univ Comput Inf Sci 31(2):266–274. https://doi.org/10.1016/j.jksuci.2018.01.005

Iskandar AA (2017) Topic extraction method using RED-NMF Algorithm for detecting outbreak of some disease on twitter. In: AIP conference proceedings, vol 1825. AIP Publishing LLC, p 020010

Isonuma M, Mori J, Bollegala D, Sakata I (2020) Tree-structured neural topic model. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 800–806

Jelisavčić V, Furlan B, Protić J, Milutinović V (2012) Topic models and advanced algorithms for profiling of knowledge in scientific papers. In: MIPRO 2012—35th international convention on information and communication technology, electronics and microelectronics—proceedings, pp 1030–1035

Jiang L, Lu H, Xu M, Wang C (2016) Biterm pseudo document topic model for short text. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI). IEEE, pp 865–872

Jiang M, Liu R, Wang F (2018) Word network topic model based on Word2Vector. In: 2018 IEEE fourth international conference on big data computing service and applications (BigDataService). IEEE, pp 241–247

Karami A, Gangopadhyay A, Zhou B, Kharrazi H (2018) Fuzzy approach topic discovery in health and medical corpora. Int J Fuzzy Syst 20(4):1334–1345. https://doi.org/10.1007/s40815-017-0327-9

Kaur K, Bansal D (2019) Techniques to extract topical experts in twitter: a survey. In: Information and communication technology for intelligent systems (ICTIS 106), Smart innovation, systems and technologies. Springer, Singapore, pp 391–399

Kherwa P, Bansal P (2020) Topic modeling: a comprehensive review. EAI Endors Trans Scalable Inf Syst 7(24):159623. https://doi.org/10.4108/eai.13-7-2018.159623

Kim HD, Park DH, Yue Lu, Zhai CX (2012) Enriching text representation with frequent pattern mining for probabilistic topic modeling. Proc Am Soc Inf Sci Technol 49(1):1–10. https://doi.org/10.1002/meet.14504901209

Kim S, Park H, Lee J (2020) Word2vec-Based Latent Semantic Analysis (W2V-LSA) for topic modeling: a study on blockchain technology trend analysis. Expert Syst Appl 152:113401. https://doi.org/10.1016/j.eswa.2020.113401

Koike D, Takahashi Y, Utsuro T, Yoshioka M, Kando N (2013) Time series topic modeling and bursty topic detection of correlated news and twitter. In: International joint conference on natural language processing, pp 917–921

Korshunova I, Xiong H, Fedoryszak M, Theis L (2019) Discriminative topic modeling with logistic LDA. In: Advances in neural information processing systems, pp 6770–6780

Kraft T, Wang DX, Delawder J, Dou W, Yu L, Ribarsky W (2013) Less after-the-fact: investigative visual analysis of events from streaming twitter. In: 2013 IEEE symposium on large-scale data analysis and visualization (LDAV). IEEE, pp 95–103

Kumar P, Vardhan M (2019) Aspect-based sentiment analysis of tweets using Independent Component Analysis (ICA) and Probabilistic Latent Semantic Analysis (PLSA). In: Advances in data and information sciences, Lecture notes in networks and systems, vol 39. Springer, Singapore, pp 3–13

Lacoste-Julien S, Sha F, Jordan MI (2009). DiscLDA: discriminative learning for dimensionality reduction and classification. In: Advances in neural information processing systems, pp 897–904

Lahoti P, Garimella K, Gionis A (2018) Joint non-negative matrix factorization for learning ideological leaning on twitter. In: Proceedings of the eleventh ACM international conference on web search and data mining. ACM Press, New York, USA, pp 351–59

Lee DD, Seung HSS (2001). Algorithms for non-negative matrix factorizationn. In: Advances in neural information processing systems, pp 556–562

Lewis DD, Yang Y, Rose TG, Li F (2004) RCV1: a new benchmark collection for text categorization research. J Mach Learn Res 5:361–397

Li X, Lei L (2021) A bibliometric analysis of topic modelling studies (2000–2017). J Inf Sci 47(2):161–175. https://doi.org/10.1177/0165551519877049

Li G, Meng K, Xie J (2013) An improved topic detection method for Chinese microblog based on incremental clustering. J Softw 8(9):2313–2320. https://doi.org/10.4304/jsw.8.9.2313-2320

Li X, Ouyang J, Zhou X (2015) Supervised topic models for multi-label classification. Neurocomputing 149:811–819. https://doi.org/10.1016/j.neucom.2014.07.053

Li C, Wang H, Zhang Z, Sun A, Ma Z (2019a) Topic modeling for short texts with auxiliary word embeddings. In: SIGIR 2016a—Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. ACM Press, New York, USA, pp 165–74

Li W, Feng Y, Li D, Zhengtao Yu (2016b) Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm. Autom Control Comput Sci 50(4):271–277. https://doi.org/10.3103/S0146411616040040

Li C, Duan Yu, Wang H, Zhang Z, Sun A, Ma Z (2017) Enhancing topic modeling for short texts with auxiliary word embeddings. ACM Trans Inf Syst 36(2):1–30. https://doi.org/10.1145/3091108

Li L, Sun Y, Wang C (2018a) Semantic augmented topic model over short text. In: 2018 5th IEEE international conference on cloud computing and intelligence systems (CCIS). IEEE, pp 652–56

Li X, Li C, Chi J, Ouyang J (2018b) Short text topic modeling by exploring original documents. Knowl Inf Syst 56(2):443–462. https://doi.org/10.1007/s10115-017-1099-0

Li X, Wang Y, Zhang A, Li C, Chi J, Ouyang J (2018c) Filtering out the noise in short text topic modeling. Inf Sci 456:83–96. https://doi.org/10.1016/j.ins.2018.04.071

Li L, Guo L, He Z, Jing Y, Wang XS (2019a) X-DMM: fast and scalable model based text clustering. In: Proceedings of the AAAI conference on artificial intelligence (AAAI-19). vol 33, pp 4197–4204

Li X, Zhang A, Li C, Guo L, Wang W, Ouyang J (2019b) Relational biterm topic model: short-text topic modeling using word embeddings. Comput J 62(3):359–372. https://doi.org/10.1093/comjnl/bxy037

Li X, Zhang J, Ouyang J (2019c) Dirichlet multinomial mixture with variational manifold regularization: topic modeling over short texts. In: Proceedings of the AAAI conference on artificial intelligence. vol 33, pp 7884–91

Li S, Zhang Yu, Pan R (2020) Bi-directional recurrent attentional topic model. ACM Trans Knowl Discov Data 14(6):1–30. https://doi.org/10.1145/3412371

Li X, Wang Y, Ouyang J, Wang M (2021) Topic extraction from extremely short texts with variational manifold regularization. Mach Learn 110(5):1029–1066. https://doi.org/10.1007/s10994-021-05962-3

Liang S, Yilmaz E, Kanoulas E (2016) Dynamic clustering of streaming short documents. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, vols 13–17. ACM, New York, NY, USA, pp 995–1004

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. https://doi.org/10.1109/ACCESS.2018.2863260

Likhitha S, Harish SB, Keerthi Kumar HM (2019) A detailed survey on topic modeling for document and short text data. Int J Comput Appl 178(39):1–9. https://doi.org/10.5120/ijca2019919265

Lim KH, Karunasekera S, Harwood A (2017) ClusTop: a clustering-based topic modelling algorithm for twitter using word networks. In: 2017 IEEE international conference on big data (Big Data). IEEE, pp 2009–18

Lin T, Hu Z, Guo X (2019) Sparsemax and relaxed wasserstein for topic sparsity. In: Proceedings of the twelfth ACM international conference on web search and data mining—WSDM '19, pp 141–149

Lin H, Zuo Y, Liu G, Li H, Junjie Wu, Zhiang Wu (2020a) A pseudo-document-based topical N-grams model for short texts. World Wide Web 23(6):3001–3023. https://doi.org/10.1007/s11280-020-00814-x

Lin L, Jiang H, Rao Y (2020b) Copula guided neural topic modelling for short texts. In: Proceedings of the 43rd international acm sigir conference on research and development in information retrieval. New York, NY, USA. ACM, pp 1773–1776

Liqing Q, Wei J, Haiyan L, Xin F (2019) Microblog hot topics detection based on VSM and HMBTM model fusion. IEEE Access 7:120273–120281. https://doi.org/10.1109/ACCESS.2019.2932458

Liu L, Huang H, Gao Y, Zhang Y, Wei X (2019) Neural variational correlated topic modeling. In: The world wide web conference. New York, NY, USA. ACM, pp 1142–52

Liu SP, Yin J, Ouyang J, Huang Y, Yang XY (2015) Topic mining from microblogs based on MB-HDP model. Chin J Comput 38(7):1408–1419. https://doi.org/10.11897/SP.J.1016.2015.01408

Liu Z, Liu C, Xia B, Li T (2018) Multiple relational topic modeling for noisy short texts. Int J Softw Eng Knowl Eng 28(11–12):1559–1574. https://doi.org/10.1142/S021819401840017X

Liu X, Jianming Fu, Chen Y (2020a) Event Evolution Model for Cybersecurity Event Mining in Tweet Streams. Inf Sci 524:254–276. https://doi.org/10.1016/j.ins.2020.03.048

Liu Z, Qin T, Chen K-J, Li Y (2020b) Collaboratively modeling and embedding of latent topics for short texts. IEEE Access 8:99141–99153. https://doi.org/10.1109/ACCESS.2020.2997973

López-Ramírez P, Molina-Villegas A, Siordia OS (2019) Geographical aggregation of microblog posts for LDA topic modeling. J Intell Fuzzy Syst 36(5):4901–4908. https://doi.org/10.3233/JIFS-179037

Lu HY, Xie LY, Kang N, Wang CJ, Xie JY (2017) Don't forget the quantifiable relationship between words: using recurrent neural network for short text topic discovery. In: Proceedings of the thirty-first AAAI conference on artificial intelligence, AAAI 2017. vol 31, pp 1192–98

Lu H-Y, Zhang Yi, Yuntao Du (2021) SenU-PTM: a novel phrase-based topic model for short-text topic discovery by exploiting word embeddings. Data Technol Appl 55(5):643–660. https://doi.org/10.1108/DTA-02-2021-0039

Magerman T, Van Looy B, Song X (2010) Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. Scientometrics 82(2):289–306. https://doi.org/10.1007/s11192-009-0046-6

Mai C, Qiu X, Luo K, Chen M, Zhao B, Huang Y (2021) TSSE-DMM: topic modeling for short texts based on topic subdivision and semantic enhancement. In: Advances in knowledge discovery and data mining. PAKDD 2021. Lecture Notes in Computer Science, vol 12713. Springer, Cham, pp 640–651

Malleson N, Birkin M (2012) Estimating individual behaviour from massive social data for an urban agent-based model. In: Modeling social phenomena in spatial context, pp 23–29

Mao X-L, Ming Z-Y, Chua T-S, Li S, Yan H, Li X (2012) SSHLDA: a semi-supervised hierarchical topic model. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pp 800–809

Mazarura J, de Waal A, de Villiers P (2020) A gamma-poisson mixture topic model for short text. Math Probl Eng 2020:1–17. https://doi.org/10.1155/2020/4728095

McAuley J, Leskovec J (2013) Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of the 7th ACM conference on recommender systems, pp 165–172. https://doi.org/10.1145/2507157.2507163

Mcauliffe JD, Blei DM (2008) Supervised topic models. In: Advances in neural information processing systems, vol 20, pp 121–128https://doi.org/10.1109/MWSCAS.2011.6026348

McCallum AK (2002) MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu

Mehrotra R, Sanner S, Buntine W, Xie L (2013) Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval—SIGIR '13, pp 889–892. https://doi.org/10.1145/2484028.2484166

Miao Y, Yu L, Blunsom P (2016) Neural variational inference for text processing. In: Proceedings of the 33rd international conference on machine learning, PMLR, vol 48, pp 1727–1736

Miao Y, Grefenstette E, Blunsom P (2017) Discovering discrete latent topics with neural variational inference. In: 34th international conference on machine learning, ICML 2017 PMLR, vol 70, pp 2410–2419

Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. In: Proceedings of the 2011 conference on empirical methods in natural language processing, EMNLP 2011, pp 262–272

Mishra RK, Urolagin S, Jothi JAA, Neogi AS, Nawaz N (2021) Deep learning-based sentiment analysis and topic modeling on tourism during covid-19 pandemic. Front Comput Sci 3:775368. https://doi.org/10.3389/fcomp.2021.775368

Mohammad SM, Kiritchenko S, Sobhani P, Zhu X, Cherry C (2016) SemEval-2016 Task 6: detecting stance in tweets. In: SemEval 2016—10th international workshop on semantic evaluation, proceedings, pp 31–41. https://doi.org/10.18653/v1/s16-1003

Mottaghinia Z, Feizi-Derakhshi M-R, Farzinvash L, Salehpour P (2020) A review of approaches for topic detection in twitter. J Exp Theor Artif Intell. https://doi.org/10.1080/0952813X.2020.1785019

Muliawati T, Murfi H (2017) Eigenspace-based fuzzy c-means for sensing trending topics in twitter. In: AIP Conference Proceedings, vol 1862, p 030140

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. https://doi.org/10.3390/s22030852

Murfi H (2017) Accuracy of separable nonnegative matrix factorization for topic extraction. In: Proceedings of the 3rd international conference on communication and information processing. ACM Press, New York, New York, USA, pp 226–30

Murshed BAH, Al-ariki HDE, Mallappa S (2020) Semantic analysis techniques using twitter datasets on big data : comparative analysis study. Comput Syst Sci Eng 35(6):495–512. https://doi.org/10.32604/csse.2020.35.495

Murshed BAH, Mallappa S, Ghaleb OAM, Al-ariki HDE (2021) Efficient twitter data cleansing model for data analysis of the pandemic tweets. In: Studies in systems, decision and control, vol 348. Springer International Publishing, pp 93–114. https://doi.org/10.1007/978-3-030-67716-9_7

Murshed BAH, Abawajy J, Mallappa S, Saif MAN, Al-ariki HDE (2022) DEA-RNN: a hybrid deep learning approach for cyberbullying detection in twitter social media platform. IEEE Access 10:25857–25871. https://doi.org/10.1109/ACCESS.2022.3153675

Mustakim NG, Reza I, Novita R, Kharisma OB, Vebrianto R, Sanjaya S, Hasbullah TA, Sari WP, Novita Y, Rahim R (2019) DBSCAN algorithm: twitter text clustering of trend topic Pilkada Pekanbaru. J Phys 1363(2019):012001. https://doi.org/10.1088/1742-6596/1363/1/012001

Nand P, Perera R, Kasture A (2016) How bullying is this message ? A psychometric thermometer for bullying. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical Papers. The COLING 2016 Organizing Committee, pp 695–706

Newman D, Lau JH, Grieser K, Baldwin T (2010) Automatic evaluation of topic coherence. In: Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics, pp 100–108

Nguyen DQ (2018) JLDADMM: a java package for the LDA and DMM topic models. ArXiv Preprint arXiv:abs/1808.03835 (Dmm):1–5

Nguyen DQ, Billingsley R, Lan Du, Johnson M (2015) Improving topic models with latent feature word representations. Trans Assoc Comput Linguist 3:299–313. https://doi.org/10.1162/tacl_a_00140

Ni N, Guo C, Zeng Z (2018) Public opinion clustering for hot event based on BR-LDA model. In: International conference on intelligent information processing, IFIP advances in information and communication technology. Springer, Cham, pp 3–11

Nigam K, Mccallum AK, Thrun S, Mitchell T (2000) Text classification from labeled and unlabeled documents using EM. Mach Learn 39(2):103–134. https://doi.org/10.1023/a:1007692713085

Nikolenko SI, Koltcov S, Koltsova O (2017) Topic modelling for qualitative studies. J Inf Sci 43(1):88–102. https://doi.org/10.1177/0165551515617393

Niyogi M, Pal AK (2019) Discovering conversational topics and emotions associated with demonetization tweets in India. Comput Intell 1:215–226. https://doi.org/10.1007/978-981-13-1132-1_17

Nugroho R, Paris C, Nepal S, Yang J, Zhao W (2020) A survey of recent methods on deriving topics from twitter: algorithm to evaluation. Knowl Inf Syst 62(7):2485–2519. https://doi.org/10.1007/s10115-019-01429-z

Nugroho R, Zhao W, Yang J, Paris C, Nepal S (2017) Using time-sensitive interactions to improve topic derivation in twitter. World Wide Web 20:61–87. https://doi.org/10.1007/s11280-016-0417-x

Nur'aini K, Najahaty I, Hidayati L, Murfi H, Nurrohmah S (2015) Combination of singular value decomposition and K-means clustering methods for topic detection on twitter. In: 2015 international conference on advanced computer science and information systems (ICACSIS). IEEE, pp 123–128

Oh O, Kwon KH, Rao HR (2010) An exploration of social media in extreme events: rumor theory and twitter during the HAITI earthquake 2010. In: ICIS 2010 proceedings—thirty first international conference on information systems, vol 231, pp 7332–7336

Ostrowski DA (2015) Using latent Dirichlet allocation for topic modelling in twitter. In: Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015). IEEE, pp 493–497

Ozyurt B, Ali Akcayol M (2021) A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA. Expert Syst Appl 168:114231. https://doi.org/10.1016/j.eswa.2020.114231

Panda R, Pensia A, Mehta N, Zhou M, Rai P (2019) Deep topic models for multi-label learning. In: The 22nd international conference on artificial intelligence and statistics . PMLR, vol 89, pp 2849–2857

Pang J, Li X, Xie H, Rao Y (2016) SBTM: topic modeling over short texts. In: International conference on database systems for advanced applications(DASFAA), Lecture Notes in Computer Science (LNCS 9645). Springer, Berlin, pp 43–56

Pang J, Rao Y, Xie H, Xizhao Wang Fu, Wang L, Wong T-L, Li Q (2019) Fast supervised topic models for short text emotion detection. IEEE Trans Cybern. https://doi.org/10.1109/tcyb.2019.2940520

Peng M, Ouyang S, Zhu J, Huang J, Wang H, Yong J (2018a) Emerging topic detection from microblog streams based on emerging pattern mining. In: 2018 IEEE 22nd international conference on computer supported cooperative work in design (CSCWD). IEEE, pp 259–264

Peng M, Xie Q, Zhang Y, Wang H, Zhang X, Huang J, Tian G (2018b) Neural sparse topical coding. In: ACL 2018—56th annual meeting of the association for computational linguistics, proceedings of the conference (Long Papers), vol 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 2332–2340

Peng M, Xie Q, Wang H, Zhang Y, Tian G (2019) Bayesian sparse topical coding. IEEE Trans Knowl Data Eng 31(6):1080–1093. https://doi.org/10.1109/TKDE.2018.2847707

Pham D, Le T (2020) Auto-encoding variational bayes for inferring topics and visualization. In: Proceedings of the 28th international conference on computational linguistics. International Committee on Computational Linguistics, Stroudsburg, PA, USA, pp 5223–5234

Pham D, Le TMV (2021) Neural topic models for hierarchical topic detection and visualization. In: Oliver N, Pérez-Cruz F, Kramer S, Read J, Lozano JA (eds) Machine learning and knowledge discovery in databases. Research Track. ECML PKDD 2021. Lecture Notes in Computer Science, vol. 12977. Springer International Publishing, Cham, pp 35–51

Phan X-H, Nguyen C-T (2006) Jgibblda: a java implementation of latent dirichlet allocation (Lda) using gibbs sampling for parameter estimation and inference. http://jgibblda.sourceforge.net

Phan X-H, Nguyen C-T (2007) GibbsLDA++: A C/C++ implementation of latent dirichlet allocation (LDA. http://gibbslda.sourceforge.net/

Phan X-H, Nguyen LM, Horiguchi S (2008) Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th international conference on World Wide Web. ACM, pp 91–100

Pornwattanavichai A, Sakolnagara PB, Jirachanchaisiri P, Kitsupapaisan J, Maneeroj S (2020) Enhanced tweet hybrid recommender system using unsupervised topic modeling and matrix factorization-based neural network. In: Supervised and unsupervised learning for data science. Springer, Cham, pp 121–143

Prakoso Y, Murfi H, Wibowo A (2018) Kernelized eigenspace based fuzzy C-means for sensing trending topics on twitter. In: Proceedings of the 2018 international conference on data science and information technology. ACM Press, New York, USA, pp 6–10

Pu X, Chatti MA, Thüs H, Schroeder U (2016) Wiki-LDA: a mixed-method approach for effective interest mining on twitter data. In: Proceedings of the 8th international conference on computer supported education, vol 1 (Csedu). SCITEPRESS, pp 426–433

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. PAKDD 2017. Lecture Notes in Computer Science, vol 10235. Springer, Cham, pp 363–74

Qiang J, Li Y, Yuan Y, Liu W, Wu X (2018a) STTM: a tool for short text topic modeling, pp 1–7

Qiang J, Li Y, Yuan Y, Xindong Wu (2018b) Short text clustering based on pitman-yor process mixture model. Appl Intell 48(7):1802–1812. https://doi.org/10.1007/s10489-017-1055-4

Qiang J, Qian Z, Li Y, Yuan Y, Xindong Wu (2020) Short text topic modeling techniques, applications, and performance: a survey. IEEE Trans Knowl Data Eng 14(8):1–19. https://doi.org/10.1109/TKDE.2020.2992485

Qomariyah S, Iriawan N, Fithriasari K (2019) Topic modeling twitter data using latent dirichlet allocation and latent semantic analysis. In: AIP conference proceedings, vol 2194, p 020093

Quan X, Kit C, Ge Y, Pan SJ (2015) Short and sparse text topic modeling via self-aggregation. In: Proceedings of the twenty-fourth international joint conference on artificial intelligence (IJCAI 2015), pp 2270–2276

Quercia D, Askham H, Crowcroft J (2012) TweetLDA: supervised topic classification and link prediction in twitter. In: Proceedings of the 4rd annual ACM web science conference. ACM Press, New York, New York, USA, pp 247–250

Rahimi M, Zahedi M, Mashayekhi H (2022) A probabilistic topic model based on short distance co-occurrences. Expert Syst Appl 193:116518. https://doi.org/10.1016/j.eswa.2022.116518

Ramage D, Rosen E, Chuang J, Manning CD, Mcfarland DA (2009) Topic modeling for the social sciences. In: NIPS 2009 workshop on applications for topic models: text and beyond, vol 5, pp 1–4

Rashid J, Shah SMA, Irtaza A (2019a) A novel fuzzy K-Means Latent Semantic Analysis (FKLSA) approach for topic modeling over Medical and Health Text Corpora. J Intell Fuzzy Syst 37(5):6573–6588. https://doi.org/10.3233/JIFS-182776

Rashid J, Shah SMA, Irtaza A (2019b) Fuzzy topic modeling approach for text mining over short text. Inf Process Manag 56(6):102060. https://doi.org/10.1016/j.ipm.2019.102060

Řehůřek R, Sojka P (2011) Gensim—statistical semantics in python. Retrieved from Genism.Org

Ren F, Ye Wu (2013) Predicting user-topic opinions in twitter with social and topical context. IEEE Trans Affect Comput 4(4):412–424. https://doi.org/10.1109/T-AFFC.2013.22

Rezaee M, Ferraro F (2020) A discrete variational recurrent topic model without the reparametrization trick. Adv Neural Inf Process Syst 33:13831–13843

Roberts ME, Stewart BM, Tingley D (2019) Stm : an R package for structural topic models. J Stat Softw 91(2):1–40. https://doi.org/10.18637/jss.v091.i02

Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on web search and data mining. ACM, New York, NY, USA, pp 399–408

Rosenberg A, Hirschberg J (2007) V-measure: a conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp 410–420

Rubin TN, Chambers A, Smyth P, Steyvers M (2012) Statistical topic models for multi-label document classification. Mach Learn 88(1–2):157–208. https://doi.org/10.1007/s10994-011-5272-5

Ruths D, Pfeffer J (2014) Social media for large studies of behavior. Science 346(6213):1063–1064. https://doi.org/10.1126/science.346.6213.1063

Saha A, Sindhwani V (2012) Learning evolving and emerging topics in social media: a dynamic NMF approach with temporal regularization. In: Proceedings of the fifth ACM international conference on Web search and data mining—WSDM '12. ACM Press, New York, New York, USA, p 693

Sasaki K, Yoshikawa T, Furuhashi T (2014) Twitter-TTM : an efficient online topic modeling for twitter considering dynamics of user interests and topic trends. In: 2014 joint 7th international conference on soft computing and intelligent systems (SCIS) and 15th international symposium on advanced intelligent systems (ISIS). IEEE, pp 440–445

Savage T, Dit B, Gethers M, Poshyvanyk D (2010) TopicXP: exploring topics in source code using latent dirichlet allocation. In: 2010 IEEE international conference on software maintenance. IEEE, pp 1–6

Shahbazi Z, Byun Y-C (2020) Topic modeling in short-text using non-negative matrix factorization based on deep reinforcement learning. J Intell Fuzzy Syst 39(1):753–770. https://doi.org/10.3233/JIFS-191690

Shahbazi Z, Byun Y-C (2021) Topic prediction and knowledge discovery based on integrated topic modeling and deep neural networks approaches. J Intell Fuzzy Syst. https://doi.org/10.3233/JIFS-202545

Sharath KBR, Kuochen W, Shi-Min S (2019) Corpus-based topic derivation and timestamp-based popular hashtag prediction in twitter. J Inf Sci Eng 35(3):675–696. https://doi.org/10.6688/JISE.201905_35(3).0011

Shi T, Kang K, Choo J, Reddy CK (2018) Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In: Proceedings of the 2018 world wide web conference on world wide web—WWW '18. ACM Press, New York, New York, USA, pp 1105–1114

Shi L, Junping Du, Liang M, Kou F (2019a) Dynamic topic modeling via self-aggregation for short text streams. Peer-to-Peer Netw Appl 12(5):1403–1417. https://doi.org/10.1007/s12083-018-0692-7

Shi X, Xue B, Tsou M-H, Ye X, Spitzberg B, Gawron JM, Corliss H, Lee J, Jin R (2019b) Detecting events from the social media through exemplar-enhanced supervised learning. Int J Digital Earth 12(9):1083–1097. https://doi.org/10.1080/17538947.2018.1502369

Shirolkar AA, Deshmukh RJ (2019) Finding topic experts in the twitter dataset using LDA algorithm. Int J Appl Evol Comput 10(2):19–26. https://doi.org/10.4018/IJAEC.2019040103

Sievert C, Shirley K (2014) LDAvis: a method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 63–70

Singh J, Singh AK (2020) NSLPCD: topic based tweets clustering using node significance based label propagation community detection algorithm. Ann Math Artif Intell. https://doi.org/10.1007/s10472-020-09709-z

Sitorus AP, Murfi H, Nurrohmah S, Akbar A (2017) Sensing trending topics in twitter for greater Jakarta area. Int J Electr Comput Eng 7(1):330–336. https://doi.org/10.11591/ijece.v7i1.pp330-336

Slutsky A, Hu X, An Y (2014) Hash-based stream LDA: topic modeling in social streams. In: Pacific-Asia conference on knowledge discovery and data mining (PAKDD 2014), LNAI 8443. Springer, Cham, pp 151–162

Squicciarini A, Rajtmajer S, Liu Y, Griffin C (2015) Identification and characterization of cyberbullying dynamics in an online social network. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015. ACM, New York, NY, USA, pp 280–285

Srivastava A, Sutton C (2017) Autoencoding variational inference for topic models. In: 5th international conference on learning representations, ICLR 2017—conference track proceedings, pp 1–12

Stieglitz S, Mirbabaie M, Ross B, Neuberger C (2018) Social media analytics—challenges in topic discovery, data collection, and data preparation. Int J Inf Manag 39:156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Tajbakhsh MS, Bagherzadeh J (2019) Semantic knowledge LDA with topic vector for recommending hashtags: twitter use case. Intell Data Anal 23(3):609–622. https://doi.org/10.3233/IDA-183998

Trupthi M, Pabboju S, Narsimha G (2018) Possibilistic fuzzy C-means topic modelling for twitter sentiment analysis. Int J Intell Eng Syst 11(3):100–108. https://doi.org/10.22266/IJIES2018.0630.11

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. In: Proceedings of the international AAAI conference on web and social media (ICWSM), vol 4, pp 178–185

Vaca CK, Mantrach A, Jaimes A, Saerens M (2014) A time-based collective factorization for topic discovery and monitoring in news. In: Proceedings of the 23rd international conference on World wide web—WWW ’14. ACM Press, New York, New York, USA, pp 527–538

Valdez D, Pickett AC, Goodson P (2018) Topic modeling: latent semantic analysis for the social sciences. Soc Sci Q 99(5):1665–1679. https://doi.org/10.1111/ssqu.12528

Vargas-Calderón V, Camargo JE (2019) Characterization of citizens using Word2vec and latent topic analysis in a large set of tweets. Cities 92:187–196. https://doi.org/10.1016/j.cities.2019.03.019

Vayansky I, Kumar SAP (2020) A review of topic modeling methods. Inf Syst 94:101582. https://doi.org/10.1016/j.is.2020.101582

Wandabwa HM, Asif Naeem M, Mirza F, Pears R (2021) Topical affinity in short text microblogs. Inf Syst 96:101662. https://doi.org/10.1016/j.is.2020.101662

Wang Z, Iwaihara M (2015) Cross-lingual tweet recommendation based on user interest using bilingual LDA related work. In: Proceedings of 7th forum on data engineering and information management (DEIM), pp 1–8

Wang C, Blei D, Heckerman D (2008) Continuous time dynamic topic models. In: Proceedings of the 24th conference on uncertainty in artificial intelligence, UAI 2008, pp 579–586

Wang Y, Agichtein E, Benzi M (2012) TM-LDA: efficient online modeling of latent topic transitions in social media. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM Press, New York, New York, USA, pp 123–131

Wang F, Liu R, Zuo Y, Zhang H, Zhang H, Wu J (2016) Robust word-network topic model for short texts. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI). IEEE, pp 852–856

Wang J, Chen L, Qin L, Wu X (2018) ASTM: an attentional segmentation based topic model for short texts. In: 2018 IEEE international conference on data mining (ICDM). IEEE, pp 577–586

Wang R, Zhou D, He Y (2019) ATM: adversarial-neural topic model. Inf Process Manag 56(6):102098. https://doi.org/10.1016/j.ipm.2019.102098

Wang W, Guo B, Shen Y, Yang H, Chen Y, Suo X (2021a) Robust supervised topic models under label noise. Mach Learn 110(5):907–931. https://doi.org/10.1007/s10994-021-05967-y

Wang Y, Li X, Zhou X, Ouyang J (2021b) Extracting topics with simultaneous word co-occurrence and semantic correlation graphs: neural topic modeling for short texts. In: Findings of the association for computational linguistics: EMNLP 2021b. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 18–27

Weng J, Lim EP, Jiang J, He Q (2010) Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on web search and data mining, pp 261–70https://doi.org/10.1145/1718487.1718520

Wilcox KT, Jacobucci R, Zhang Z, Ammerman BA, Wilcox KT (2021) Supervised latent dirichlet allocation with covariates: a bayesian structural and measurement model of text and covariates. https://doi.org/10.31234/osf.io/62tc3

Wu X, Li C (2019) Short text topic modeling with flexible word patterns. In: 2019 International joint conference on neural networks (IJCNN), vols 2019-July. IEEE, pp 1–7

Wu D, Zhang M, Shen C, Huang Z, Mingxing Gu (2020a) BTM and GloVe similarity linear fusion-based short text clustering algorithm for microblog hot topic discovery. IEEE Access 8:32215–32225. https://doi.org/10.1109/ACCESS.2020.2973430

Wu X, Li C, Zhu Y, Miao Y (2020b) Short text topic modeling with topic distribution quantization and negative sampling decoder. In: Proceedings of the 2020b conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 1772–1782

Xia L, Luo D, Zhang C, Wu Z (2019) A survey of topic models in text classification. In: 2019 2nd international conference on artificial intelligence and big data, ICAIBD, IEEE. IEEE, pp 244–250

Xiao Ya, Fan Z, Tan C, Qian Xu, Zhu W, Cheng F (2019) Sense-based topic word embedding model for item recommendation. IEEE Access 7:44748–44760. https://doi.org/10.1109/ACCESS.2019.2909578

Xie W, Zhu F, Jiang J, Lim E-P, Wang Ke (2016) TopicSketch: real-time bursty topic detection from twitter. IEEE Trans Knowl Data Eng 28(8):2216–2229. https://doi.org/10.1109/TKDE.2016.2556661

Xie Q, Huang J, Du P, Peng M, Nie J-Y (2021) Graph topic neural network for document representation. In: Proceedings of the web conference 2021. ACM, New York, NY, USA, pp 3055–3065

Xu Y, Xu H, Zhu L, Hao H, Deng J, Sun X, Bai X (2018) Topic discovery for streaming short texts with CTM. In: 2018 international joint conference on neural networks (IJCNN), pp. 1–7, IEEE.

Yan X, Guo J, Liu S, Cheng X-Q, Wang Y (2012) Clustering short text using ncut-weighted non-negative matrix factorization. In: Proceedings of the 21st ACM international conference on Information and knowledge managementACM Press, New York, New York, USA, pp 2259–2262

Yan X, Guo J, Lan Y, Cheng X (2013a) A bitem topic model for short texts. In: International world wide web conference committee (IW3C2), pp 1445–1455

Yan X, Guo J, Liu S, Cheng X, Wang Y (2013b) Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In: Proceedings of the 2013b SIAM international conference on data mining. Society for Industrial and Applied Mathematics, Philadelphia, PA, pp 749–757

Yan X, Guo J, Lan Y, Xu J, Cheng X (2015) A probabilistic model for bursty topic discovery in microblogs. In: Twenty-ninth AAAI of the national conference on artificial intelligence, pp 353–359

Yang Y, Wang F (2021) Author topic model for co-occurring normal documents and short texts to explore individual user preferences. Inf Sci 570:185–199. https://doi.org/10.1016/j.ins.2021.04.060

Yang C, Zhou M, Ye S, Xu X (2013) An improved hot topic detection method for microblog based on CURE algorithm. Comput Simul 30(11):383–387

Yang Y, Wang F, Zhang J, Jin Xu, Philip SYu (2018) A topic model for co-occurring normal documents and short texts. World Wide Web 21(2):487–513. https://doi.org/10.1007/s11280-017-0467-8

Yang S, Huang G, Cai B (2019) Discovering topic representative terms for short text clustering. IEEE Access 7:92037–92047. https://doi.org/10.1109/ACCESS.2019.2927345

Yang L, Wu F, Gu J, Wang C, Cao X, Jin D, Guo Y (2020) Graph attention topic modeling network. In: Proceedings of the web conference 2020. ACM, New York, NY, USA, pp 144–154

Yao F, Wang Y (2020) Tracking urban geo-topics based on dynamic topic model. Comput Environ Urban Syst 79:101419. https://doi.org/10.1016/j.compenvurbsys.2019.101419

Yeh J-Y, Ke H-R, Yang W-P, Heng Meng I (2005) Text summarization using a trainable summarizer and latent semantic analysis. Inf Process Manag 41(1):75–95. https://doi.org/10.1016/j.ipm.2004.04.003

Yi F, Jiang Bo, Jianjun Wu (2020) Topic modeling for short texts via word embedding and document correlation. IEEE Access 8:30692–30705. https://doi.org/10.1109/ACCESS.2020.2973207

Yin J, Wang J (2014) A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining—KDD '14. ACM Press, New York, New York, USA, pp 233–242

Yin J, Wang J (2016) A text clustering algorithm using an online clustering scheme for initialization. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, vols. 13–17-Augu. ACM, New York, NY, USA, pp 1995–2004

Yirdaw ED, Ejigu D (2012) Topic-based amharic text summarization with probabilistic latent semantic analysis. In: Proceedings of the international conference on management of emergent digital ecosystems—MEDES '12. ACM Press, New York, New York, USA, pp 8–15

Yu J, Qiu L (2019) ULW-DMM: an effective topic modeling method for microblog short text. IEEE Access 7:884–893. https://doi.org/10.1109/ACCESS.2018.2885987

Yu G, Huang R, Wang Z (2010) Document clustering via dirichlet process mixture model with feature selection. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10. ACM Press, New York, New York, USA, p 763

Yu D, Wu Y, Sun J, Ni Z, Li Y, Wu Q, Chen X (2017) Mining hidden interests from twitter based on word similarity and social relationship for OLAP. Int J Softw Eng Knowl Eng 27(09–10):1567–1578. https://doi.org/10.1142/S0218194017400113

Yu D, Dengwei Xu, Wang D, Ni Z (2019) Hierarchical topic modeling of twitter data for online analytical processing. IEEE Access 7:12373–12385. https://doi.org/10.1109/ACCESS.2019.2891902

Zeng J, Li J, Song Y, Gao C, Lyu MR, King I (2018) Topic memory networks for short text classification. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 3120–3131

Zhai K, Boyd-Graber J, Asadi N, Alkhouja ML (2012) Mr. LDA: a flexible large scale topic modeling package using variational inference in MapReduce. In: Proceedings of the 21st international conference on world wide web. ACM Press, New York, New York, USA, pp 879–888

Zhang Y, Eick CF (2019) Tracking events in twitter by combining an LDA-based approach and a density-contour clustering approach. Int J Seman Comput 13(01):87–110. https://doi.org/10.1142/S1793351X19400051

Zhang X, Zhang Li (2020) Topics extraction in incremental short texts based on LSTM. Soc Netw Anal Min 10(1):83. https://doi.org/10.1007/s13278-020-00699-8

Zhang J, Tang J, Zhong Y, Mo Y, Li J, Song G, Hall W, Sun J (2017) StructInf: mining structural influence from social streams. In: 31st AAAI conference on artificial intelligence, AAAI 2017, vol 1, pp 73–79

Zhang Lu, Zhiang Wu, Zhan Bu, Jiang Ye, Cao J (2018a) A pattern-based topic detection and analysis system on chinese tweets. J Comput Sci 28:369–381. https://doi.org/10.1016/j.jocs.2017.08.016

Zhang X, Feng R, Liang W (2018b) Short text topic model with word embeddings and context information. In: International conference on computing and information technology (IC2IT 2018b), AISC 769, Advances in Intelligent Systems and Computing. Springer, Cham, pp 55–64

Zhang Y, Wang Z, Yu Y, Chen B, Ma J, Shi L (2018c) LF-LDA: a supervised topic model for multi-label documents classification. Int J Data Warehousing Mining 14(2):18–36. https://doi.org/10.4018/IJDWM.2018040102

Zhang Z, Robinson D, Tepper J (2018d) Detecting hate speech on twitter using a convolution-GRU based deep neural network. In: GangemiAnna A, Gentile AL, Nuzzolese AG, Rudolph S, Maleshkova M, Paulheim H, Pan IZ, Alam M (eds) The European semantic web conference. ESWC 2018d. Lecture Notes in Computer Science, vol 10843. Springer, Cham, pp 745–760

Zhang C, Shaozhen Lu, Zhang C, Xiao X, Wang Q, Chen G (2019) A novel hot topic detection framework with integration of image and short text information from twitter. IEEE Access 7:9225–9231. https://doi.org/10.1109/ACCESS.2018.2886366

Zhao Y, Karypis G (2001) Criterion functions for document clustering: experiments and analysis

Zhao WX, Jiang J, Weng J, He J, Lim E-P, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: European conference on information retrieval. Springer, Berlin, pp 338–349

Zhao H, Phung D, Huynh V, Jin Y, Du L, Buntine W (2021) Topic modelling meets deep neural networks: a survey arXiv:abs/2103.00498

Zheng W, Ge B, Wang C (2019) Building a TIN-LDA model for mining microblog users' interest. IEEE Access 7:21795–21806. https://doi.org/10.1109/ACCESS.2019.2897910

Zhu Q, Feng Z, Li X (2018) GraphBTM: graph enhanced autoencoded variational inference for biterm topic model. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 4663–4672

Zhu L, He Y, Zhou D (2019a) Hierarchical viewpoint discovery from tweets using bayesian modelling. Expert Syst Appl 116:430–438. https://doi.org/10.1016/j.eswa.2018.09.028

Zhu L, Hua Xu, Yunfeng Xu, Xiao Yi, Li J, Deng J, Sun X, Bai X (2019b) A joint model of extended LDA and IBTM over streaming Chinese short texts. Intell Data Anal 23(3):681–699. https://doi.org/10.3233/IDA-183836

Zubiaga A, Ji H (2013) Harnessing web page directories for large-scale classification of tweets. In: WWW 2013 companion—proceedings of the 22nd international conference on world wide web. https://doi.org/10.1145/2487788.2487904, pp 225–226

Zuo Y, Wu J, Zhang H, Lin H, Xu K, Xiong H (2016a) Topic modeling of short texts: a pseudo-document view. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2016), pp 2105–2114

Zuo Y, Zhao J, Ke Xu (2016b) Word network topic model: a simple but general solution for short and imbalanced texts. Knowl Inf Syst 48(2):379–398. https://doi.org/10.1007/s10115-015-0882-z

Zuo Y, Li C, Lin H, Junjie Wu (2021) Topic modeling of short texts: a pseudo-document view with word embedding enhancement. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2021.3073195

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

Authors and Affiliations

Belal Abdullah Hezam Murshed[1,2] · **Suresha Mallappa**[1] · **Jemal Abawajy**[3] · **Mufeed Ahmed Naji Saif**[4] · **Hasib Daowd Esmail Al-ariki**[5,6] · **Hudhaifa Mohammed Abdulwahab**[7]

Suresha Mallappa
sureshasuvi@gmail.com

Jemal Abawajy
jemal.abawajy@deakin.edu.au

Mufeed Ahmed Naji Saif
mufeed.a.nsaif@gmail.com

Hasib Daowd Esmail Al-ariki
hasibalariki@gmail.com

Hudhaifa Mohammed Abdulwahab
hudhaifa.alhimyari@gmail.com

[1] Department of Studies in Computer Science, Mysore University, Mysore 570006, Karnataka, India

[2] Department of Computer Science, College of Engineering and IT, Amran University, Amran, Yemen

[3] School of Information Technology, Faculty of Science, Engineering and Built Environment, Deakin University, Geelong, VIC 3220, Australia

[4] Department of Computer Applications, Sri Jayachamarajendra College of Engineering, VTU, Mysore, Karnataka, India

[5] Department of Computer Networks and Distributed Systems, Al Saeed Faculty for Engineering and IT, Taiz University, Taiz, Yemen

[6] Department of Computer Networks Engineering and Technologies, Sana'a Community College, Sana'a, Yemen

[7] Department of Computer Application, Ramaiah Institute of Technology, VTU, Bengaluru, India