J. Inst. Eng. India Ser. B (October 2022) 103(5):1431–1441
https://doi.org/10.1007/s40031-022-00776-w

ORIGINAL CONTRIBUTION

# A Novelty Analysis about an Impact of Tweets and Twitter Bios on Topic Quality Discovery using the Topic Modeling

**Rathinasamy Muthusami**[1] · **Kandhasamy Saritha**[2]

Received: 16 August 2021 / Accepted: 3 July 2022 / Published online: 20 July 2022

**Abstract** Tweets and users are two key sources that characterize Twitter. Users have bios to describe their background and personal interests, as with tweet messages, furthermore analyzing the content of these tweets and the user's bio is the inspiration for this research. The topics discovered from tweets and bios are fairly conceivable alone, and the research challenge is how to measure the topics' quality once coupled. In this research, an attempt has been made in the novelty analysis of tweets and user's account bios by implementing topic models, i.e., Latent Dirichlet Allocation and Correlated Topic Model with the number of topics 10, 20, and 30, wherein hashtag and user pooling schemes have been applied to review tweets and bios that are topically equivalent, to receive documents that are not only convenient but also topically coherent, where better topics have been revealed. These tweets and bios are gathered from a Twitter user in a certain timeline using the hashtag #App. A set of dendrograms for bios and tweets text has been created to analyze the topics that have been rendered by topic modeling in order to build dendrograms and compare them. The entanglement value was determined after a visual comparison of the dendrograms. Between dendrograms, the cophenetic correlation coefficient has also been estimated. The findings showed that both the user bio and the tweet text had an impact on topic quality discovery. On the basis of numerous measurements conducted in this study, it has also been discovered that the LDA topic model outperforms the CTM topic model.

**Keywords** Twitter bios · Short text mining · Topic discovery · Clustering techniques

## Introduction

In recent years, microblogging has grown in popularity as a means of communication and knowledge discovery, with Twitter being one of the most well-known microblogging sites, with over 500 million active users [1–7]. Tweets, retweets, account users, bios, geo locations, user language, link, screen name address, individual friend counts, and entity strings are just a few of the information available on Twitter. This information is tied to users and tweet messages. Tweets and Twitter users were both influenced by one another [8–10]. Each tweet has an author, a post, a one-of-a-kind id, the date and time it was tweeted, and, on rare occasions, the user's spatial position data. The majority of users are needed to have a Twitter handle, an identifier, followers, and, in certain cases, a bio. Tweets are the fundamental components of each Twitter account. Twitter, too, uses "status alerts." Username, timestamp, and message are among the "root-stage" attributes of the tweet functionality. Identity, in this meaning, denotes a distinct identification. The tweet and timestamp both provide UTC time, indicating when the tweet was sent, and the text contains the UTF-8 text of the status update. The post is explained using the metadata from the user's public Twitter account. A user could be anyone or anything. Each user's metadata has a lot in common. The account identity is one of the few fields that will never change when the account is being set up.

✉ Rathinasamy Muthusami
r.muthusami@gmail.com

[1] Department of Computer Applications, Dr. Mahalingam College of Engineering and Technology, Coimbatore, Tamil Nadu, India

[2] Department of Mathematics, P.A. College of Engineering and Technology, Coimbatore, Tamil Nadu, India

The contents of tweets and user bios were analysed to reveal how the two have influenced one another in terms of finding high-quality subjects. In the Tweet text, which is blended with inferred traces [11], short and less than 140 characters (now 280 characters) are employed. Twitter profiles are less than 160 characters long and provide precise information about what the user truly does, allowing one to pitch Twitter users' true identities [12]. Users of Twitter can tell the world about themselves in far less than 160 characters. This box can be filled in a variety of ways, including expressing a desire, an interest, or other personal information, as well as presenting facts like age, family status, location, or occupation. In this bio description, users have a few things in common, and they use specific phrases to express their function or occupation. Their accomplishments are highlighted in flattering Twitter profiles, which characterise them as legitimate and humanising. It may include an invitation to invite others to join them. It will tell more about the user's desires and affiliations, similar to tweets [13, 14], and it will be exciting to learn more about this.

The best technique for analysing short texts is topic modelling, according to a prior study [15–26]. In this study, Twitter topic models, such as LDA and CTM, were used to review tweets that were topically equivalent, to receive documents that were not only convenient but also topically coherent, and where better topics were discovered, using a Twitter user's bio and data on tweets for each model with the number of topics 10, 20, and 30, and hashtag and user pooling schemes have been used to review tweets that were topically equivalent. The computation of Ward-D2 and Euclidean distance methods on hierarchical clustering on LDA and CTM topic models produced a collection of dendrograms for bios and tweets text to evaluate and compare the subjects given by topic modelling. The entanglement value was calculated after visual comparison of the dendrograms. Dendrogram correlation matrices have also been computed. The findings showed that both the user bio and the tweet text had an impact on topic quality discovery.

The most significant contributions of this research are;

- Twitter topic models, i.e., LDA and CTM have been implemented with Twitter user's bio and data on tweets for each model with the number of topics 10, 20, and 30.
- On each model, the topics are discovered, and dendrograms are built to evaluate them.
- On dendrograms, the various metrics were analyzed to show that both the user bio and the tweet content had an impact on topic quality discovery.
- On both tweets and bios, the performance of topic models was investigated.

## Related Work

Semertzidis investigated how Twitter users express themselves in their profile bios and discovered that the analysis of user bios is capable of predicting the linkages between Twitter users [27]. Wagner investigated several types of user-related information, such as tweets, retweets, and user list memberships, to gain a better understanding of Twitter users' expertise [28]. These studies use quantitative methods to characterise Twitter users by assessing user-related information such as tweets, retweets, and biographies; however, qualitative technique is not explored. The proposed study utilises both qualitative and quantitative methodologies to achieve better outcomes.

Rodriguez proposed a fuzzy logic-based followee recommendation mechanism on Twitter. This system approaches recommendation as a link prediction problem and makes use of three types of resemblance between two users: similarity of tweets, similarity of followee ids, and similarity of followee tweets. These commonalities are computed in the extraction of user profiles [29]. Tran developed a hashtag recommendation approach based on a study of tweet content, user attributes, and very popular Twitter hashtags, which dramatically enhances the effectiveness of hashtag recommendation systems [30]. Corcoglioniti proposed a recommendation system that uses attributes ranging from core social media features to specialised domain relevant user profile traits inferred from knowledge using machine learning methods [31]. The approaches used in these researches solely investigate Twitter user profiles or tweets; however, they do not evaluate tweet content and user profiles in tandem, and hence lose out on discovering appropriate topics. In order to uncover high-quality topics, it has been assessed the influence of user profiles on tweet content.

Ding investigated methods to obtain new knowledge exposing people's preferences from Twitter bios. A progressive labeling model has been trained by autonomously created labeled facts to establish a messy training sample set using a series of seed sequences and intuitive criteria, and then a CRF model is trained on this labeled data set. They also investigated the relationship between interest tags derived from user bios and tweet text using tf-idf frequencies of lexical items as traits and discovered a weak association between them, indicating that bios might possibly serve as a complementary resource of data to tweets. In contrast, the proposed work implies that tweet content and user profiles have a strong connection to discovering quality topics [32].

Jones presented a publicly accessible pooled database as an examination of temporal changes in individually articulated identity; the collection permits evaluation of the predominance of phrases preferred by Twitter users in the United States of America for inclusion in those bios. They employed longitudinal online profile sampling approach for examining people's preferences while identifying individuals using terms [33]. Rogers examined Twitter bios and found that the average American user is progressively incorporating politics into their social identity. They employed longitudinal online profile sampling to provide measurable insights on how people change their identities over time [34]. Pathak utilized phenomenological strategies to discover and describe the idea of a distinctive identifier, that's essentially indicative of how identity is conveyed in Twitter bios. In addition, a strategy for extracting all individual identifiers contained in a given bio was developed. In addition, the researchers assessed the dependability, authenticity, usefulness, and usefulness of employing terms derived from Twitter bios to explore socioeconomic ethnicity [35]. These recent studies used longitudinal online profile sampling approaches to analyze Twitter users' biographies and tweet content for public social identification; however, they did not use any unsupervised learning methodologies and validation methods to discover quality content, which the proposed method performs.

## Materials and Methods

Figure 1 depicts the conceptual model of this research, which includes data preparation, pre-processing of data, specification of the topic model, selection of the topic model, and visualization of the topic modeling, as well as the construction of dendrograms, performance investigation of the topic modeling, and applications. The following sections discuss each phase.

### Dataset

The dataset #App (tweets and bios) have been gathered, based on hashtag and user pooling schemes with the support of Twitter Streaming APIs and R library. The Twitter Streaming APIs gave access to all tweets as they published on Twitter which has sent out the real-time tweets. The R library has supported to make a connection to Twitter Streaming APIs, which extracted tweets with specific search terms (hashtag, #), language, duration etc. A hashtag (#) is a Twitter tradition used to streamline inquiry, indexing and pattern disclosure. Users can incorporate exceptionally designed terms that begin with # into the body of every post. For the proposed work, the dataset has been collected with hashtag #App, language is English and the duration is between 29th June'19 and 22nd July'19 in which 1,06,868 tweets have been gathered from Twitter. After the data have been gathered based on hashtag, ten thousand users have been grasped by using R libraries in which 4,000 user's bio have been collected.

#### *Data Pre-Processing*

Data pre-processing enabled the production of high-quality text categorization while also reducing computing complexity. On the dataset, the following pre-processing techniques were used: remove the "RT" (retweet), links, hashtags, punctuation, white spaces, stop words, and convert text to lower case. Stop words in this context are words that convey an interfacing capacity in the sentence, such as relational words and prepositions like "the," "is," "at," "which," and "on." The document-term matrix (DTM) was generated after pre-processing to build the topic model.

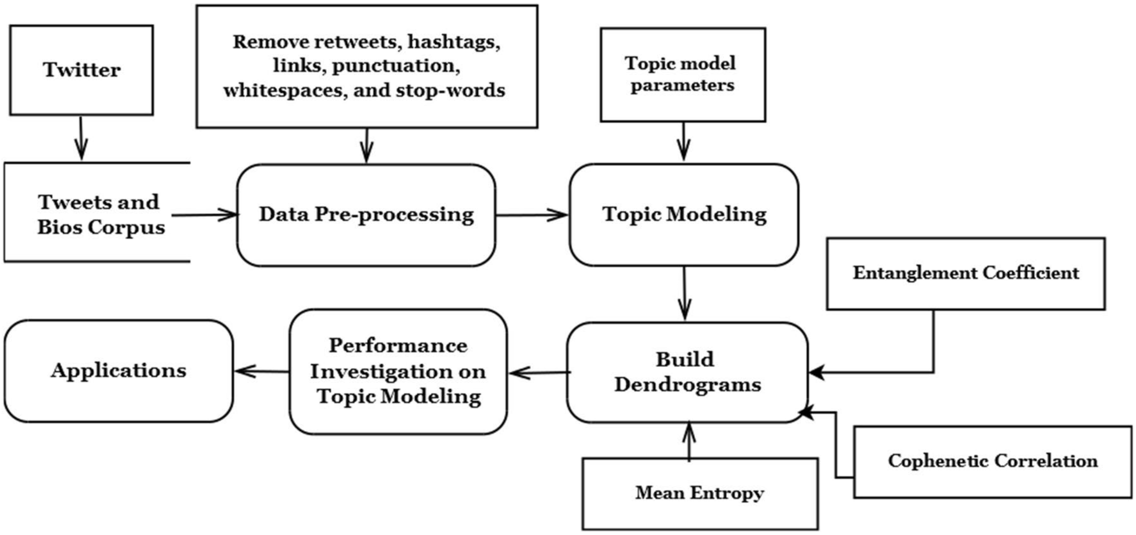

**Fig. 1** The conceptual model of this research

## Topic Models

The objective of topic modeling is described in the introduction as inevitably finding the topics in a set of documents. The arrangement of topics per document and terms per topic is the topic structure's concealed structure. A document often covers a number of topics in diverse degrees of depth. Topic modeling approaches establish categories of comparable terms as topics [36–38]. A topic model encapsulates this intuition in a statistical model that enables for the assessment of a series of documents and the discovery of how well the topics may be relying on the statistical of terms for each, as well as the proportion of topics in each document. The most efficient experimental challenge for topic modeling is inferring the concealed topic structure from the observed texts. In this work, the topic models such as LDA and CTM have been implemented on both datasets, i.e., Tweets and Bios.

### *Latent Dirichlet Allocation (LDA)*

In latent Dirichlet allocation (LDA), the topic model is a part of the larger field of probabilistic modeling. LDA could be described further legitimately with the representation below [15]. The topics are $\beta_{1:t}$, where every $\beta_t$ is a jargon distribution. $\theta_s$ is the topic scope of the $s^{th}$ document in which $\theta_{s',t}$ is the topic scope for the topic t in document $s$. $c_s$ is the $s^{th}$ document of topic consignment in which $c_{s,m}$ is the topic undertaking for the $m^{th}$ word in document $s$.

There, $o_s$ are the observed words of document $s$, wherein $o_{s,m}$ is the $m^{th}$ word in document $s$, it is an element of the stable jargon. The LDA generative method focuses on particular accumulated dispersion of the concealed and observable factors with this notation (Eq. 1),

$$p\left(\beta_{1:T},\theta_{1:S},c_{1:S},o_{1:t}\right) = \prod_{i=1}^{t} p(\beta_i) \prod_{s=1}^{S} p(\theta_s) \left( \prod_{m=1}^{M} p(c_{s,m}|\theta_s \right) p(o_{s,m}|\beta_{1:T}c_{s,m}) \tag{1}$$

It is to be kept in mind that this distribution specifies some of the dependencies. For example, the assignment of the topic $o_{s,m}$ depends on the scope of the topic for each document $\theta_s$. The example of this is the term $o_{s,m}$ which is reliant on the topic consignment $c_{s,m}$ and all topics $\beta_{1:t}$

The number of topics should be pre-set in advance for fitting the LDA model to a given document-term matrix. Further, the appraisal, utilizing Gibbs sampling, needs the information of values for the parameters of the prior dispersals, i.e., 10/*k* for $\alpha$, the topic distribution of documents $\theta$, drawn from a Dirichlet prior with parameter $\alpha$ and 0.1 for $\phi$, each $\phi$ is drawn from a Dirichlet prior with parameter $\beta$, the term distribution of the topics.

Now looking into computational problem, the provisional dispersal of the topic structure present in the observed documents is computed. This is called the posterior. Using the notation, the posterior is, Eq. 2,

$$p\left(\beta_{1:T},\theta_{1:S},c_{1:S}|\,o_{1:S}\right) = \frac{p\left(\beta_{1:T,}\theta_{1:S},C_{1:S},O_{1:S}\right)}{p\left(O_{1:S}\right)} \tag{2}$$

In the equation above, the combined dispersal of all contingent factors is the numerator which is the hidden factors' impact. The minimal probability of the remarks is the denominator in the implemented topic model.

## Correlated Topic Models (CTM)

The dispersals of topics and terms for specified documents are characterized by the topic models. The topic quantities for a document $s$ is $\theta_s$, which is a vector of all possible topics $T$, considered for the topic dispersal. The correlated topic model makes it possible to compare topics by pretending a multivariate dispersal for a transmuted version of $\theta_s$. The data come as the occurrence of the word, i.e., the number of times each word from a certain jargon seems in a specified document is significant to the topic model [39]. It is to be supposed that $S$ documents, each having $M_s$ words are created from a jargon of $J$ words. The $m^{th}$ word in document s is denoted $o_{s,m} \in \{1,\ldots,J\}$, $s = 1,\ldots,S$, $m = 1,\ldots, M_s$. Understanding the topic attributed to that word, $o_{s,m}$ accompany a multinomial distribution over the vocabulary of $J$ words. The topic itself is assigned, $c_{s,m}$ follows a multinomial distribution over the possible $T$ topics. The chance of the term seems in a document is commendably resolute by the term occurrences, and the likelihood of the topic is also a hidden variable to be assessed.

In the proposed work, an LDA is fitted for demonstration with 10 topics, utilizing Gibbs sampling with the appropriate values have been set to control parameters. Wherein the initial iteration has been set as zero and then every1000 iterations have been returned for 1000 cycles. The underlying $\alpha$ has been set to 10, 20, and 30/*k*, and the best model regarding the log-likelihood *log(p(o|c))* has been observed when Gibbs sampling is returned.

The CTM is fitted for demonstration with 10 topics, utilizing the variational inference and Expectation–Maximization (EM) algorithm with control parameters setting as 500 to the maximum number of iterations. The tolerance for the relative change in the likelihood set as 0.0001 for variational inference step as well as for the EM algorithm. And for the same parameters, the values have been set as 1000 and 0.001, respectively, trying to optimize the fit to

**Table 1** Mean entropy of 10, 20, and 30 topics from the fitted topic models LDA and CTM on Tweets dataset

| No. of topics/ Model | $N=10$ | $N=20$ | $N=30$ |
|---|---|---|---|
| LDA | 2.298525 | 2.9881518 | 3.3906809 |
| CTM | 2.019442 | 2.4268366 | 2.6268160 |

**Table 2** Mean entropy of 10, 20, and 30 topics from the fitted topic models LDA and CTM on Bios dataset

| No. of topics/ Model | $N=10$ | $N=20$ | $N=30$ |
|---|---|---|---|
| LDA | 2.2990445 | 2.988750 | 3.391125 |
| CTM | 1.8353607 | 2.355219 | 2.842286 |

**Table 3** Sample topics from the fitted topic model LDA on Tweets dataset

| Topic 1 | Topic 3 | Topic 6 | Topic 9 |
|---|---|---|---|
| "Fiverr" | "Ziprecruiter" | "Backup" | "Parkmobile" |
| "Dayforce" | "Zoom" | "Gif" | "Hotschedules" |
| "Blackberry" | "Shoporgdxtop" | "Traveling" | "Entrepreneurs" |
| "Reuters" | "Lead" | "Onavo" | "Indeed" |
| "Hotschedules" | "Vlc" | "Workday" | "Ctr" |
| "Fullscreen" | "Psychiatry" | "Moovit" | "Classpass" |
| "Careerbuilder" | "Voxer" | "Deadpool" | "Medium" |
| "Traveling" | "Heineken" | "Vyclone" | "Ups" |
| "Mobilereading" | "Managing" | "Careerbuilder" | "Noise" |
| "Joined" | "Casestudy" | "Snapshot" | "Tapas" |

the data observed. The resultant approximate mean entropy is given below.

Tables 1 and 2 show the mean entropy of a number of topics 10, 20, and 30 rendered by the LDA and CTM topic models on dataset #App (tweets) and #App (bios), respectively. It shows that topics and mean entropy values are instantly proportionate to the number of topics and that the greater value of the fitted topic model relatively suggests that the distribution of topics is sensible to the dataset.

Tables 3 and 4 show the sample topics rendered by the implementation of LDA with number of topics 10 on dataset #App (tweets) and #App (bios), respectively. From Table 3, the topic 1's some terms include “dayforce,” "reuters," and "hotschedules" that recommend distinctive apps. The topic 6's some terms include "moovit," "vyclone," and "careerbuilder" which recommend other sets of apps. From Table 4, the topic 3's some terms include "hacker," "testers," and "developer" that suggests some identity. The topic 9's include "musician," "traveling," "evofuse," and "networking" which recommends other sets of identity.

Tables 5 and 6 depict the topics rendered by the implementation of CTM with number of topics 10 on dataset #App (tweets) and #App (bios), respectively. From Table 5, the topic 1's some terms include "ziprecruiter," "mobileiron," and "jobseeker" that recommend distivtive apps. In the topic 9's some terms include "shopping," "vyclone," and "kronos" which recommend other sets of apps. From Table 6, the topic 6's some terms include "developer," "calendars," and "icbc" that suggest some identity. The topic 9's include "musician," "appanalytics," and "hacker" which recommends other sets of identity.

## Results and Discussion

The topics rendered by the topic models LDA and CTM are reviewed and contrasted in the proposed work, which involves creating dendrogrms. Dendrograms are a form of "bottom-up" agglomerative hierarchical clustering. To put it another way, each entity (topic) has only one item cluster at first (leaf). The two very similar clusters are merged into a huge new cluster at every stage of the set of rules (nodes).

The similarity between the objects was determined using the Euclidean distance method. A distance or dissimilarity

**Table 4** Sample topics from the fitted topic model LDA on Bios dataset

| Topic 1 | Topic 3 | Topic 6 | Topic 9 |
|---|---|---|---|
| "Fun" | "Hacker" | "Stupid" | "musician" |
| "Philippines" | "Networking" | "Center" | "Day" |
| "Alliance" | "Calendars" | "Locationintelligence" | "Tips" |
| "Awesome" | "Indian" | "Sectors" | "Insurance" |
| "Tecnológico" | "Day" | "Musician" | "Traveling" |
| "Locationintelligence" | "Testers" | "Canarias" | "Evofuse" |
| "Evofuse" | "Forward" | "Ate" | "Networking" |
| "Hacker" | "Developer" | "Thai" | "Testers" |
| "Networking" | "Beakerhead" | "Vfx" | "Forward" |
| "Day" | "Mark" | "Traveling" | "Developer" |

**Table 5** Sample topics from the fitted topic model CTM on Tweets dataset

| Topic 1 | Topic 3 | Topic 6 | Topic 9 |
|---|---|---|---|
| "ziprecruiter" | "Guidebook" | "Adp" | "Shopping" |
| "Mobileiron" | "Chromebook" | "Chromecast" | "Shoporgdxtop" |
| "Indeed" | "Hotschedules" | "Tapas" | "Workday" |
| "Jobseeker" | "Gif" | "Classpass" | "Deadpool" |
| "Seeker" | "Heineken" | "Jobr" | "Vyclone" |
| "Weighing" | "Nps" | "Psychiatry" | "Concur" |
| "Medium" | "Iran" | "Magic" | "Kronos" |
| "Employees" | "Ian" | "Fromshows" | "Moovit" |
| "Cfo" | "Medium" | "Linking" | "Announced" |
| "Tuesdaymotivation" | "Hitman" | "Bigcommerce" | "Enterprisemobile" |

**Table 6** Sample topics from the fitted topic model CTM on Bios dataset

| Topic 1 | Topic 3 | Topic 6 | Topic 9 |
|---|---|---|---|
| "Attractions" | "Hacker" | "Center" | "Musician" |
| "Networking" | "Sectors" | "Developer" | "Stupid" |
| "Projects" | "Forward" | "Tips" | "Underestimate" |
| "Vfx" | "Adaptable" | "Beakerhead" | "Appanalytics" |
| "Insurance" | "Alliance" | "Calendars" | "Locationintelligence" |
| "Atat" | "Traveling" | "Mark" | "Sectors" |
| "Ate" | "Philippines" | "Stoked" | "Hacker" |
| "Sectors" | "Goes" | "Icbc" | "Forward" |
| "Hacker" | "Messaging" | "Thai" | "Goes" |
| "Traveling" | "Developer" | "Messaging" | "Alliance" |

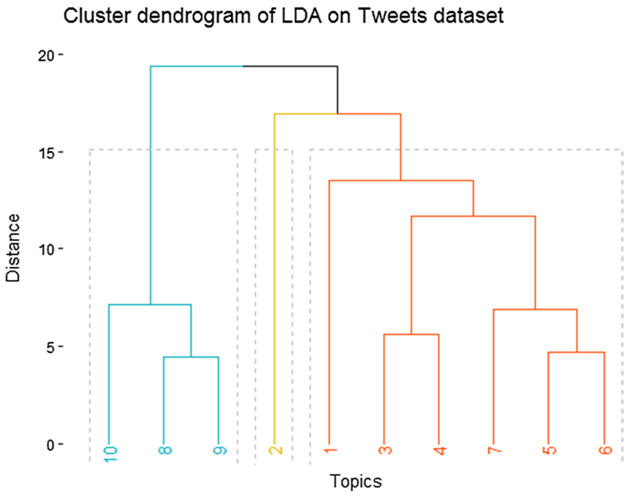

**Fig. 2** Display the dendrogram of 10 topics from the fitted topic model LDA on Tweets dataset

matrix is the result of this calculation. The linkage function then uses the distance information to group things into clusters based on their resemblances. The construction of new clusters is then connected to one another to form larger clusters. This method is repeated until a hierarchical tree is formed with all of the objects in the actual dataset. Various cluster agglomeration methodologies exist (i.e., linkage strategies).

The most often used linking methods are full or full linkage, minimum or single linkage, mean or average linkage, centroid linkage, and Ward's minimum variance method. In this study, Ward's method, as well as the Euclidean distance method, are recommended. In the dendrogram below, each leaf represents a single entity. The (dis)similarity/distance between two objects/clusters is shown by the vertical axis height of the merge. The higher the elevation of the mixture, the less comparable the objects are. The cophenetic distance between the two objects is defined as this height.

The distances (i.e., heights) within the tree should be calculated after connecting the items in a dataset into a hierarchical cluster tree, portraying the actual distances as they should be. The cophenetic distances and the actual distance data obtained with the distance method are then compared to see how well the cluster tree performed. The initial distance matrix's distances between points have a strong correlation with the cluster tree's object linking. Clustering validity has been kept. The value of the correlation coefficient that is closest to 1 is right, whereas the value found is 0.75, which is considered appropriate.

On each #App (tweets) and #App (bios) dataset, the process began with the generation of a collection of two dendrograms using the Ward-D2 method of hierarchical clustering on LDA and CTM topic modelling. Figures 2, 3, 4, and 5 show them individually.

Figure 2 shows that the topics are grouped into 3 clusters, the topics 8, 9, and 10 are grouped as cluster 1, the topic 2 grouped as cluster 2 and the topics 1, 3, 4, 5, 6, and 7 are grouped as cluster 3, respectively. Among the 3 clusters, the topics 5, 6, 7, 8, 9, and 10 are almost of the same distance. Fig. 3 shows that the topics are grouped into 3 clusters, the topics 8, 9, and 10 are grouped as cluster 1, the topic 2 grouped as cluster 2 and the topics 1, 3, 4, 5, 6, and 7 are grouped as cluster 3. Topics 5, 6, 7, 8, 9, and 10

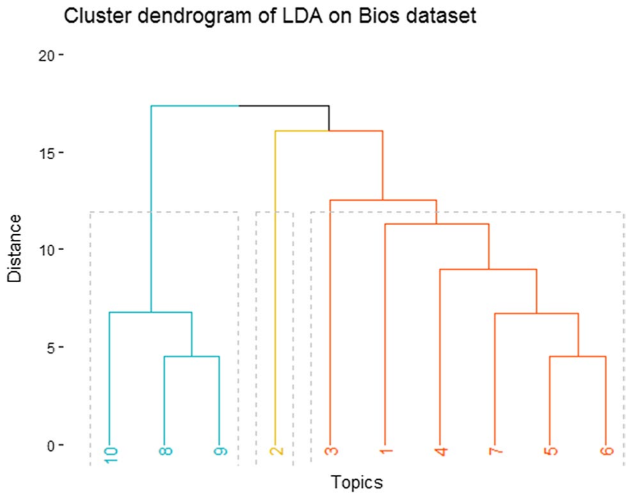

**Fig. 3** Display the dendrogram of 10 topics from the fitted topic model LDA on Bios dataset

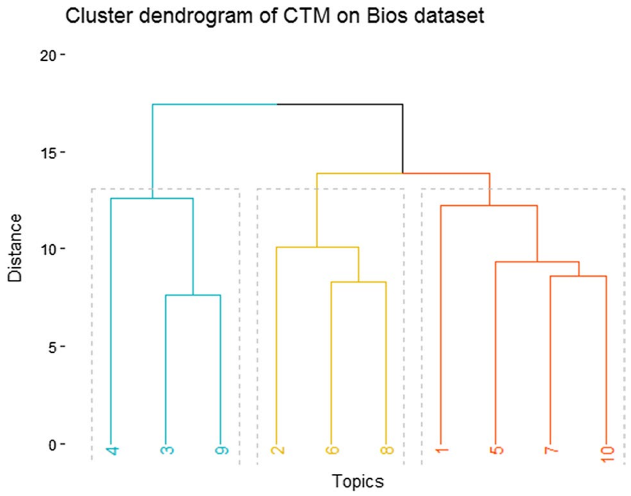

**Fig. 5** Display the dendrogram of 10 topics from the fitted topic model CTM on Bios dataset

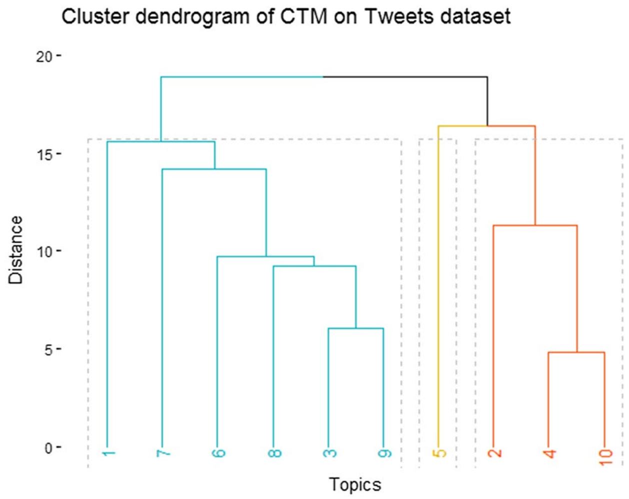

**Fig. 4** Display the dendrogram of 10 topics from the fitted topic model CTM on Tweets dataset

are almost the same distance apart among the three clusters. Figures 2 and 3 show that the cluster dendrogram of 10 LDA topic model topics on dataset #App (tweets) and the cluster dendrogram of 10 LDA topic model topics on dataset #App (bios) are almost identical.

Figure 4 shows that the topics are grouped into 3 clusters. The topics 1, 3, 6, 7, 8, and 9 are grouped as cluster 1, topic 5 grouped as cluster 2 and the topics 2, 4, and 10 are grouped as cluster 3, respectively. Among the 3 clusters the topics 3, 4, 9, and 10 are almost of same distance. Figure 5 shows that the topics are grouped into 3 clusters, the topics 3, 4, and 9 are grouped as cluster 1, topics 2, 6, and 8 are grouped as cluster 2 and the topics 1, 5, 7, and 10 are grouped as cluster 3, respectively. Among the 3 clusters the topics 3, 6, 7, 8, 9, and 10 are almost of the same distance. From both Figs. 4 and 5, it is understood that the cluster dendrogram of 10 topics of CTM topic model on dataset #App (tweets) and the cluster dendrogram of 10 topics of CTM topic model on dataset #App (bios) are distinctive. After computing hierarchical clustering, two dendrograms are compared visually and a correlation matrix between the dendrograms is computed.

In an experiment, the two dendrograms are visually measured and plotted beside each other, having respective labels bound by arcs. Based upon on measures of an entanglement value among tree, the quality of the alignment or separation has been measured. Entanglement is calculated by assigning values to the left tree's labels ranging from 1 to tree length and then matching such values to the right tree. Entanglement is defined as the L norm distance among both two feature vector, L indicates the panelty level; for acute angles, choose a large panelty. It was evaluated between 1 (complete entanglement) and 0 (no entanglement).

The reduced coefficient of entanglement shows an effective alignment. In the diagram, the highlighted line indicates unique nodes. Figures 6 and 7 show that the entanglement coefficient of cluster dendrograms of dataset #App (tweets and bios) on 10 topics of LDA topic model is 0.02 and 0.64 for cluster dendrograms of dataset #App (tweets and bios) on 10 topics of CTM topic model, i.e., LDA has 99.98% and CTM has 36% good, indicating that LDA has outperformed CTM, which has proved that tweets and bios are mutually influenced to discover the topics. The entanglement coefficient for the number of topics 20 and 30, as shown in Table 7, has also been calculated. As a consequence, it was recognized that LDA is performing better for the number of topics 10, 20, and 30. The same could be interpreted with the outcomes shown in Figs. 2 and 3, respectively.

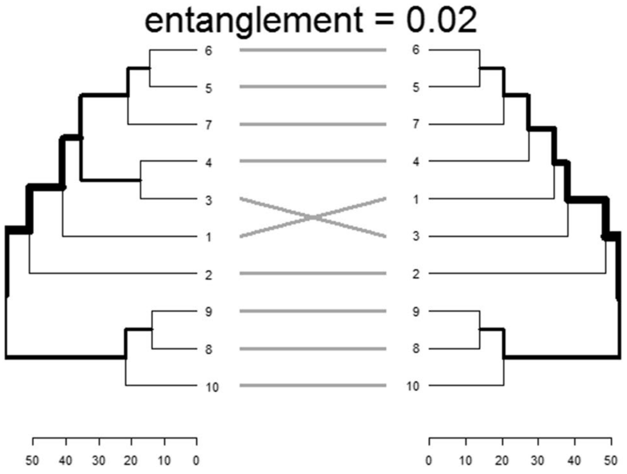

**Fig. 6** Comparison of cluster dendrograms of 10 topics from the fitted topic model LDA on Tweets and Bios datasets

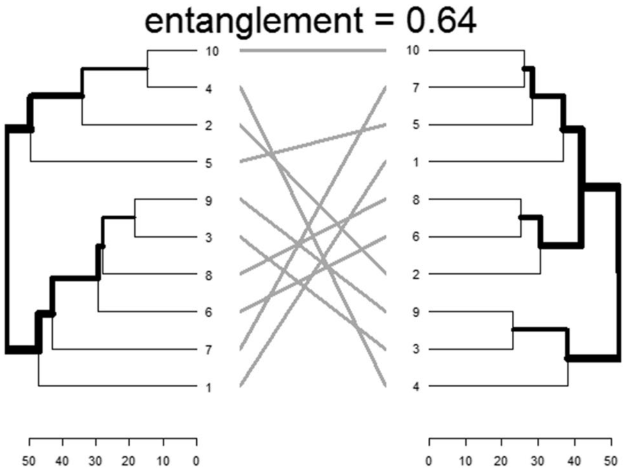

**Fig. 7** Comparison of cluster dendrograms of 10 topics from the fitted topic model CTM on Tweets and Bios datasets

**Table 7** Entanglement co-efficient of dendrograms of topics 10, 20, and 30 built from the fitted topic models LDA and CTM on the Tweets and Bios datasets

| Model | Number of topics(N) | | |
|---|---|---|---|
| | $N=10$ | $N=20$ | $N=30$ |
| | Entanglement coefficient | | |
| LDA | 0.02 | 0.13 | 0.09 |
| CTM | 0.64 | 0.38 | 0.60 |

Next, the cophenetic correlation matrix between the cluster dendrograms of #App (tweets and bios) dataset on 10 topics of LDA topic model and also the cluster dendrograms of #App (tweets and bios) dataset on 10 topics of CTM topic model have been computed. The linear connection among the variance for every range of parameters with respective associated cophenetic intervals is described by the cophenetic correlation constant. The cophenetic intervals are a metric of outgroups variances between two factors that were integrated on a similar cluster. This approach was used to determine if a dendrogram is an adequate method or not; the strong correlation (closer to 1) between the initial and cophenetic intervals indicates a high good for a particular dendrogram. The resulting values were shown in Figs. 8 and 9, respectively. Also, the cophenetic correlation coefficient for the number of subjects 20 and 30 is calculated, as shown in Table 8. As a result, it was understood that LDA performs better for the number of topics 10, 20, and 30, as the coefficient of cophenetic correlation is close to 1 for both tweets and bios dendrograms, indicating that tweets and bios are mutually influenced for the discovery of topics.

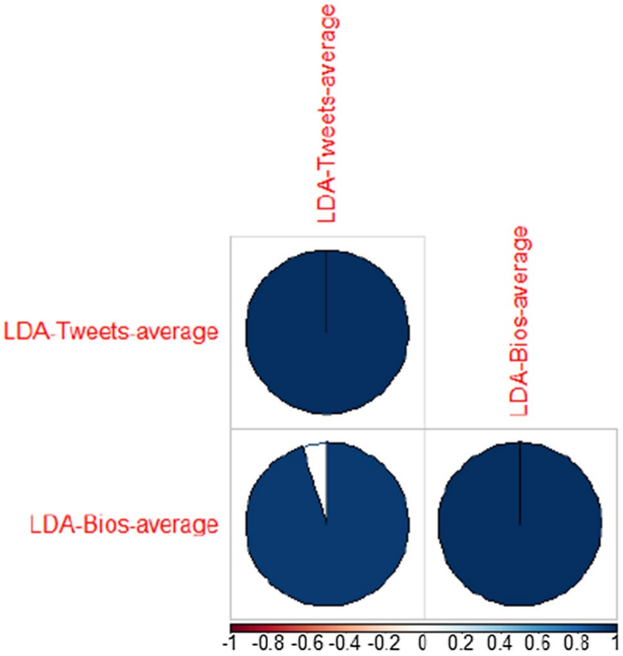

**Fig. 8** Shows the cophenetic correlation coefficient between dendrograms of Tweets and Bios datasets on 10 topics of LDA topic model

In an overview, the proposed technique was contrasted with another method, namely perplexity. Perplexity has been used as the most common method for assessing topic models [40, 41]. In the case of LDA, it is not so good because it is hard to comprehend. LDA works well when the topics are soft clustered. Sometimes it shows better on the topic model when learning works smoothly, but it shows very poor overall model quality. Therefore, it is evident that the suggested technique was better than the other technique, i.e., perplexity.

Furthermore, since the generative probabilistic model, LDA, does not have implicit clustering properties, hierarchical clustering methods have been explicitly applied, resulting in agglomerative clusters on various topics that can be evaluated

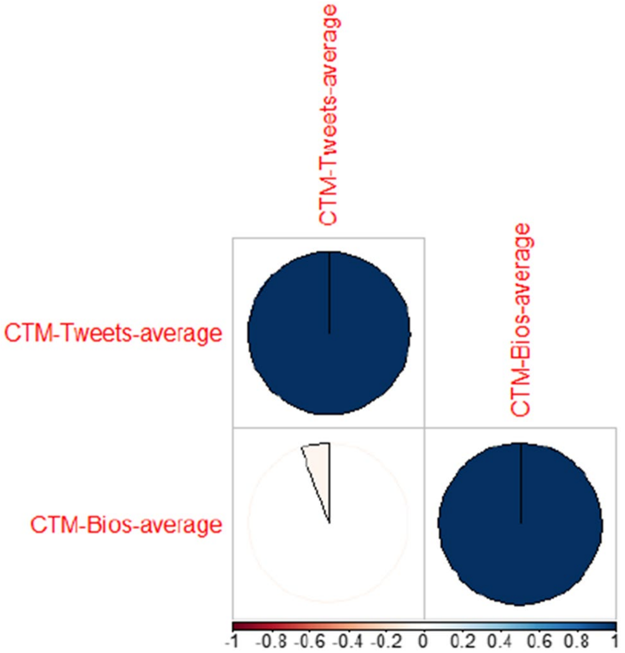

**Fig. 9** Shows the cophenetic correlation coefficient between dendrograms of Tweets and Bios datasets on 10 topics of CTM topic model

**Table 8** Cophenetic correlation co-efficient of dendrograms of topics 10, 20, and 30 built from the fitted topic models LDA and CTM on the Tweets and Bios datasets

| Model | Number of topics (N) | | |
|---|---|---|---|
| | $N=10$ | $N=20$ | $N=30$ |
| | Cophenetic correlation coefficient | | |
| LDA | 0.9521602 | 0.937985 | 0.9440211 |
| CTM | −0.05584715 | 0.1029337 | 0.1033684 |

uniquely in this proposed work by comparing the topics discovered from various corpora such as tweets and bios.

The limitations of the research work are the quantity of user bios and tweets, as well as the single dataset used, are too limited. And also, only three metrics with two techniques were used to assess the performance. All of these will be expanded in future work.

## Conclusion

Twitter topic models, i.e., LDA and CTM, have been deployed with Twitter user bios and tweet data for each model with 10, 20, and 30 topics. A set of dendrograms for bios and tweets text has been created in order to analyse and compare the topics that have been rendered using topic modelling. The entanglement value was determined by comparing the dendrograms visually. The cophenetic correlation coefficient between dendrograms has also been computed. The findings revealed that both the user bio and the tweet text had an impact on topic quality discovery. Based on the following metrics, the LDA topic model outperforms the CTM topic model.

1. The mean entropy of 10, 20, and 30 topics produced by the LDA and CTM topic models on datasets #App (tweets) and #App (bios) has been computed (Tables 1 and 2). It demonstrates that the number of topics and mean entropy values are directly proportional to the number of topics, and that the higher the value of the fitted topic model, LDA reveals that the distribution of topics is responsive to the dataset.
2. The entanglement coefficient for the number of topics 10, 20, and 30, as shown in Table 7, was determined to assess the quality of the alignment or separation among dendrograms of dataset #App (tweets and bios) on LDA and CTM topic models. As an outcome, it was determined that LDA performs better for the number of topics 10, 20, and 30.
3. The cophenetic correlation coefficient was then determined for the number of topics 10, 20, and 30, as shown in Table 8. As a consequence, it was revealed that LDA performs better for the number of topics 10, 20, and 30, since the coefficient of cophenetic correlation for both tweets and bios dendrograms is nearer to one.

In future, researchers intend to contribute the proposed method to additional Twitter topic models that incorporate more information, such as follower profiles and retweets, for social identities and recommendation applications.

**Funding** The authors have not disclosed any funding

**Declarations**

**Conflict of interest** The authors declare there is no conflict of interest

## References

1. A. Java, X. Song, T. Finin, B. Tseng, Why we twitter: understanding microblogging usage and communities, in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis, San Jose California, pp. 56–65, (2007)
2. R. Muthusami, A. Bharathi, K. Saritha, Covid-19 outbreak: tweet based analysis and visualization towards the influence of coronavirus in the world. Gedrag Organ. **33**(2), 534–549 (2020)

3. S. Shugars, A. Gitomer, S. McCabe, R.J. Gallagher, K. Joseph, N. Grinberg, L. Doroshenko, B.F. Welles, D. Lazer, Pandemics, protests, and publics: Demographic activity and engagement on twitter in 2020. J. Quantit. Descr. Digit. Media **1**(1), 1–68 (2021)
4. A.P. Rodrigues, R. Fernandes, A. Bhandary, A.C. Shenoy, A. Shetty, M. Anisha, Real-time twitter trend analysis using big data analytics and machine learning techniques. Wirel. Commun. Mob. Comput. **2021**(3920325), 1–13 (2021)
5. S. Fazel, L. Zhang, B. Javid, I. Brikell, Z. Chang, Harnessing twitter data to survey public attention and attitudes towards covid-19 vaccines in the UK. Sci. Rep. **11**, 23402 (2021)
6. E. Elakiya, N. Rajkumar, In text mining: detection of topic and sub-topic using multiple spider hunting model. J. Ambient Intell. Hum. Comput. **12**, 3571–3580 (2021)
7. A. Pradhan, M.R. Senapati, P.K. Sahu, ABET: an affective emotion-topic method of biterms for emotion recognition from the short texts. J. Ambient Intell. Hum. Comput. (2022). https://doi.org/10.1007/s12652-022-03799-9
8. D. Ramage, S. Dumais, D. Liebling, Characterizing microblogs with topic models, in Proceedings of the 4th international AAAI conference on weblogs and social media (ICWSM'10), Stanford University, pp. 130–137, (2010)
9. L. Hong, B.D. Davison, Empirical study of topic modeling in twitter, in Proceedings of the first workshop on social media analytics, Lehigh University, pp. 80–88, (2010)
10. N. Rogers, J.J. Jones, Using twitter bios to measure changes in self-identity: Are americans defining themselves more politically over time. J. Soc. Comput. **2**(1), 1–13 (2021)
11. W. Dong, M. Qiu, F. Zhu, Who am i on twitter?: a cross-country comparison, in Proceedings of the companion publication of the 23rd international conference on world wide web companion, Seoul, Korea, pp. 253–254, (2014)
12. R. Muthusami, A. Bharathi, Stance detection and mobile app recommendation discourse on tweets. Comput. Intell. **4**(35), 1043–1060 (2019)
13. J. Shima, M. Yoshida, K. Umemura, When do users change their profile information on twitter?, in IEEE International conference on big data(Big Data), Cornell University, pp. 3119–3122, (2017)
14. J. Li, G. Longinos, S. Wilson, W. Magdy, Emoji and self-identity in twitter bios, in Proceedings of the fourth workshop on natural language processing and computational social science, Association for Computational Linguistics University of Edinburgh, pp. 199–211, (2020)
15. D.M. Blei, Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
16. R.A. Levine, G.M. Richardson, J. Bowers, A.J. Woodill, J.R. Barr, J.M. Gawron, Topic models: a tutorial with R. Int. J. Semant. Comput. **8**(1), 85–98 (2014)
17. R. Muthusami, K. Saritha, Global analysis of covid-19 clinical related trials. J. Microbiol. Infect. Dis. **10**(4), 183–187 (2020)
18. S.J. Blair, Y. Bi, M.D. Mulvenna, Aggregated topic models for increasing social media topic coherence. Appl. Intell. **50**(1), 138–156 (2020)
19. F. Yi, B. Jiang, W. Jianjun, Topic modeling for short texts via word embedding and document correlation. IEEE Access **8**, 30692–30705 (2020)
20. D. Alvarez-Melis, M. Saveski, Topic modeling in twitter: aggregating tweets by conversations, in Proceedings of the tenth international AAAI conference on web and social media, Germany, pp.519–522, (2016)
21. A.O. Steinskog, J.F. Therkelsen, B. Gamback, Twitter topic modeling by tweet aggregation, in Proceedings of the 21st nordic conference of computational linguistics, Sweden, pp. 77–86, (2017)
22. K.W. Lim, C. Chen, W. Buntine, Twitter-network topic model: a full bayesian treatment for social network and text modeling. NIPS 2013 topic models: Computation, application and evaluation, arXiv preprint arXiv: 1609.06791, (2016)
23. R.C. Belwal, S. Rai, A. Gupta, A new graph-based extractive text summarization using keywords or topic modeling. J. Ambient Intell. Hum. Comput. **12**, 8975–8990 (2021). https://doi.org/10.1007/s12652-020-02591-x
24. T. Jose, S.S. Babu, Detecting spammers on social network through clustering technique. J. Ambient Intell. Hum. Comput. (2019). https://doi.org/10.1007/s12652-019-01541-6
25. M. Eldib, F. Deboeverie, W. Philips et al., Discovering activity patterns in office environment using a network of low-resolution visual sensors. J. Ambient Intell. Hum. Comput. **9**, 381–411 (2018). https://doi.org/10.1007/s12652-017-0511-7
26. D.C. Edara, L.P. Vanukuri, V. Sistla et al., Sentiment analysis and text categorization of cancer medical records with LSTM. J. Ambient Intell. Hum. Comput. (2019). https://doi.org/10.1007/s12652-019-01399-8
27. K. Semertzidis, E. Pitoura, P. Tsaparas, How people describe themselves on twitter, in Proceedings of the ACM SIGMOD workshop on databases and social networks, New York, pp. 31–36, (2013)
28. C. Wagner, V. Liao, P. Pirolli, L. Nelson, M. Strohmaier, It's not in their tweets: modeling topical expertise of twitter users, in ASE/IEEE international conference on social computing, amsterdam, Netherlands, pp. 91–100, (2012)
29. F.M. Rodríguez, L.M. Torres, S.E. Garza, Followee recommendation in twitter using fuzzy link prediction. Expert. Syst. **33**(4), 349–361 (2016)
30. V.C. Tran, D. Hwang, N.T. Nguyen, Hashtag recommendation approach based on content and user characteristics. Cybern. Syst. **49**(5–6), 368–383 (2018)
31. F. Corcoglioniti, Y. Nechaev, C. Giuliano, R. Zanoli, *Twitter user recommendation for gaining followers, in AI*IA 2018 – advances in artificial intelligence* (Springer, New York, 2018), pp. 539–552
32. Y. Ding, J. Jiang, Extracting interest tags from twitter user biographies, in Information retrieval technology: information retrieval technology - 10th asia information retrieval societies conference, AIRS 2014, Kuching, Malaysia, pp.268–279, (2014)
33. N. Rogers, J.J. Jones, Using twitter bios to measure changes in self-identity: Are americans defining themselves more politically over time. J. Soc. Comput. **2**(10), 1–13 (2021)
34. J.J. Jones, A dataset for the study of identity at scale: annual prevalence of American twitter users with specified token in their profile bio 2015–2020. PLoS One **16**(11), e0260185 (2021)
35. A. Pathak, N. Madani, K. Joseph, A method to analyze multiple social identities in twitter bios. Proc. ACM Hum. Comput. Interact. **5**(CSCW2), 1–35 (2021)
36. D. Chehal, P. Gupta, P. Gulati, Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations. J. Ambient Intell. Hum. Comput. **12**, 5055–5070 (2021). https://doi.org/10.1007/s12652-020-01956-6
37. A. Srivastav, S. Singh, Proposed model for context topic identification of english and hindi news article through LDA approach

with NLP technique. J. Inst. Eng. India Ser. B **103**, 591–597 (2022). https://doi.org/10.1007/s40031-021-00655-w

38. R. Singh, S. Singh, Text similarity measures in news articles by vector space model using NLP. J. Inst. Eng. India Ser. B **102**, 329–338 (2021). https://doi.org/10.1007/s40031-020-00501-5
39. D.M. Blei, J.D. Laerty, A correlated topic model of science. J. Ann. Appl. Stat. **1**(1), 17–35 (2007)
40. R. Mehrotra, S. Sanner, W. Buntine, L. Xie, 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, Dublin, Ireland, pp.889–892, (2013)
41. P. Yali, Y. Jian, L. Shaopeng, L. Jing, A biterm-based dirichlet process topic model for short texts, in Proceedings of the 3rd conference on computer science and service system, Bangkok, Thailand, pp.301–304, (2014)

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