Journal of Medical Systems (2019) 43: 93
https://doi.org/10.1007/s10916-019-1225-5

MOBILE & WIRELESS HEALTH

A Robust User Sentiment Biterm Topic Mixture Model Based on User Aggregation Strategy to Avoid Data Sparsity for Short Text

Nimala K[1] · Jebakumar R[1]

Received: 8 January 2019 / Accepted: 21 February 2019 / Published online: 5 March 2019

Abstract
Sentiment analysis is a process of computationally finding the opinions that are expressed in a short text or a feedback by a writer towards a particular topic, product, service. The short piece of review from the user can help a business determine or understand the attitude of the user thereby predict the customer's behaviour and itsubstantiallyimproves the quality of service parameters. The proposed Robust User Sentiment Biterm Topic Mixture (RUSBTM)model discovers the user preference and their sentiment orientation views for effective Topic Modelling using Biterms or word-pair from the short text of a particular venue. Since short review or text suffers from data sparse, the user aggregation strategy is adapted to form a pseudo document and the word pairset is created for the whole corpus. The RUSBTM learns topics by generating the word co-occurrence patterns thereby inferring topics with rich corpus-level information. By analysing the sentiments of the paired words and their corresponding topics in the review corpus of the particular venue, prediction can be done that exactly portrays the user interest, preference and expectation from a particular venue. The RUSBTM model proved to be more robust and also, the extracted topics are more coherent and informative. Also the method uses accurate sentiment polarity techniques to exactly capture the sentiment orientation and the model proves to be outperforming better when compared to other state of art methods.

Keywords Sentiment · Topic modelling · Biterm · Short text · Sparsity

Introduction

Social Media is an interactive and internet based Technology which facilitates in sharing of opinions, ideas, information, interest via virtual communitynetworks. People try to share information, reviews, photos etc. The social media technology takes many varieties of forms including blogs, micro blogs, forum, products/service reviews, video sharing, book markings, gaming etc. Social Medias normally carry short text such as micro blogs which is increasing in a rapid face. Hence performing sentiment analysis and simultaneously discovering topics is a great challenge and such type of scenarios is often used in wide variety of applications which involves content characterizing, content recommendation, user interest profiling, emerging topic detection and semantic analysis. Short text or messages suffer severe data sparsity problem since most of the words occur only once and there is no enough contexts to identify the sense of ambiguous words. Previous work suggested various ways to overcome the challenges of data sparsity problem namely Latent Dirichlet Allocation (LDA) with document aggregation strategy, use of mixture of unigrams and Sparse topic models. Since topics are mainly group of words correlated together and the effect of correlation is exposed by word occurence, better revealing of topics is possible with the use of rich global word occurrence pattern.

The online reviews are nothing but the Opinions or feelings generated by the users to comment on some entity. Many users are concerned with different aspects or topics and use different expressions to express the sentiments behind it. Since the sentiments are highly involved with topic determination, probabilistic topic models are widely used to explore sentiment Analysis. To better model the sentiments and topics,the user and the sentiment attached with the reviews are taken into consideration. The Robust user sentiment Biterm Topic model

This article is part of the Topical Collection on *Mobile & Wireless Health*

✉ Nimala K
nimskt@gmail.com

[1] School of Computing, SRM University, Chennai, India

(RUSBTM) is a novel approach which incorporates users and their sentiment orientation views for effective Topic Modelling using Biterms or word-pair which produces both positive topic-words, negative topic- words, user- positive topic, user-negative topics, venue item- topic distribution simultaneously.

The proposed approach uses the traditional LDA technique with document aggregation strategy for same user and directly models the word co –occurrences for topic learning, which is a better way to infer topics. In this novel approach of RUSBTM,it extends the user sentiment topic model and Author Topic model by including an added sentiment layer between the author and the topic,where the topics are obtained by Biterm pairs. The framework of the RUSBTM is more coherent model that clearly captures the users fascinated topics both likes and dislikes i.e. positive and negative topics with sentiment oriented information. The significant feature of the model is that it is an unsupervised model which correctly captures the sentiment trend of the user. The Fig. 1 depicts the scenario of the RUSBTM that incorporates an effective sentiment analysis into topic extraction by investigating the opinions of the user rather than just modelling to know what people discuss. The proposed model demonstrates the user likes or dislikes topics about the venue by cultivating the topic extraction with varying sentiment trends. The model considers the extraction of Biterms and then computes the Polarity based on sentiments. Sentiment polarity classification is done and fed to LDA, where the topics both positive and negative are extracted based on users opinion under the correct sentiment tag. Therefore it is sensible that the topics extracted by the model has better precision on the user's analysis of sentiment towards it.

The contribution of the work are,

- We proposed the RUSBTM model to identify sentiment and topics simultaneously for venue related datasets.
- We adopted collapsed Gibbs sampling to infer parameters.
- We provided both the quantitative and qualitative measure to estimate the sentiment and topic distribution.

Related works

Micro blogs or Micro reviews are the latest development in the field of communication,where people share short text in the digital form like pictures, video or audio on internet. It differs from blogs since it is of short form. Extracting a topic from the short text is quite challenging since traditional LDA could not be applied directly on micro blogs. Usually the generative models have their existence in the LDA, It assumes documents to have distribution of topics and topics to have distribution of words. LDA is a probabilistic generative model which is a multi-level hierarchical Bayesian Model in which the model as a predetermined mixture over an essential set of topics [1]. Yuan Zuo et al. proposed a novel model which is probabilistic in nature called Pseudo document based Topic

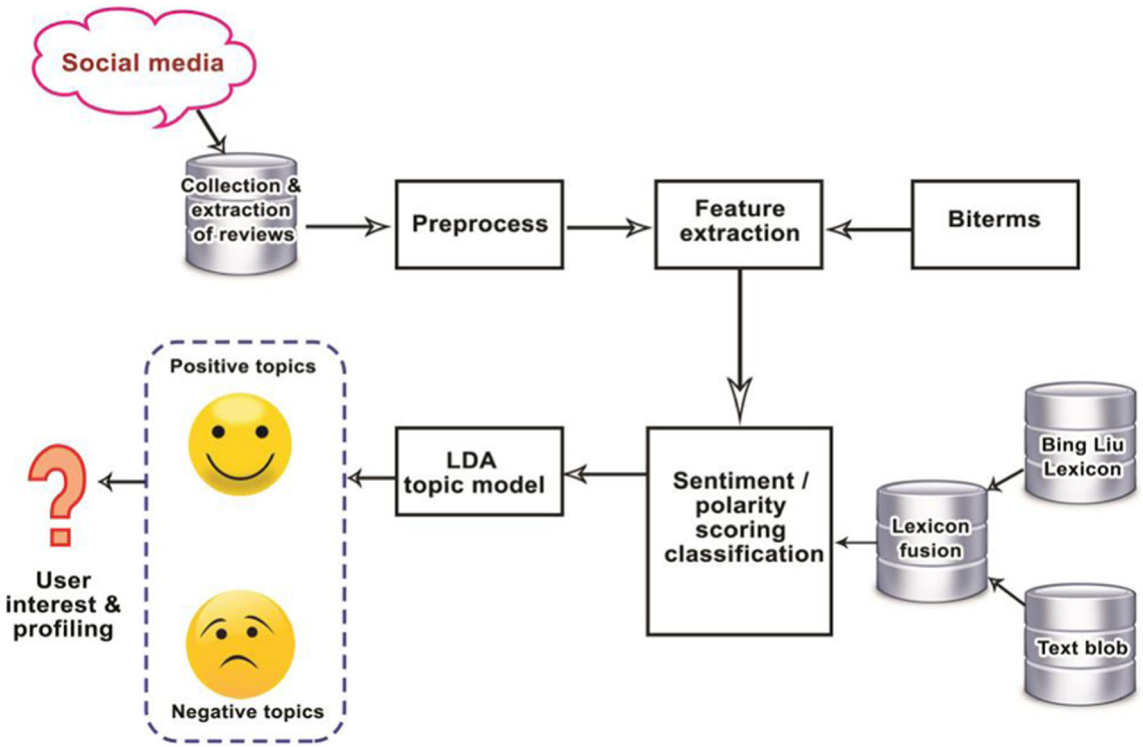

Fig. 1 Scenario depicting the RUSBTM model

Model for short text,which implicitly aggregates short text to overcome data sparsity problem. A yet another model called Sparsity enhanced PTM was also designed which involves applying priors with the intention of removing undesired correlations that exist between pseudo documents and latent topics [2]. In [3] the authors have framed three topic modelling schemes and one extended model of Author topic (AT),based on the standard LDA. The work clearly illustrated that the message aggregated by same user leads to remarkable accomplishmentin comparison with other baseline state -of art models.

Xiaohui et al. proposed a novel approach recommended as BTM stating Biterm Topic Model which learns topic directly by modelling the formation of word co-occurrence pattern. The empirical results states that BTM determines worth topics and also performs best when compared with other previous approach [4]. Short text or messages suffer severe data sparse,which the traditional or conventional model cannot cope up with, So [5] proposed a novel approach by capturing the words that co-occur as a pattern in the whole corpus. To effectively handle a large scale datasets,Xueqi introduced two online algorithms for BTM,such that the model learns higher quality topics and document's topic proportion when compared to other state-of-art methods [5]. Yunqing et al. came up with a new Discriminative Bi-term Topic model which removes low indicative bi-terms by differentiating the topical terms from general and document specific ones. His work when applied to merely headlines, induced latent topics which is as good as LDA applied to the full news Text. The authors major contribution was discriminating terms into topical terms, document-specific terms and general terms. The d-BTM(discriminative bi-term topic models) model was introduced to achieve clustering on news headlines. [6]. In the paper [7], the author proposed a twitter -BTM model,that considers user level personalization, that aggregates biterm for the user, to know the user specific topic distribution. The work also considers user's preference by incorporating a background topic and estimating the background words and topical words and the experimental results portrays that twitter-BTM outpace several state of -the art baseline methods.

Weijiang et al. experimental work employs in integrating the BTM Topic model with K-means clustering algorithm for effective topic discovery. [8]. The proposed combination of BTM model with K-means solves the sparsity problem as well as improves the result of clustering. In [9], Yali Pan proposed a Dirichlet process based on coexistence of words in which topics are mined from short text automatically. The results shows that the model outplays the other methods in terms of quality of topics, perplexity etc. ChengtaoLi et al. in his work decomposed the multiplicative process of sentiment polarity of words in to two layer hierarchy.ie the first layer computes whether a word is to be a sentiment word or an ordinary word and in the layer two,if it is found to be a sentiment word,it automatically concludes the polarity of that words. [10].

In the reference [11],the experimenter discusses the various pooling schemes which aggregates tweets in order to overcome the data sparsity problem. His experimental result analysis states that the novel method of tweet aggregation by hash tags leads to extensive improvement of higher measure for topic coherence and also the automatic hash tag labelling further improves the pooling results, thereby improving LDA topic models on twitter content. Qiaozhu et al. proposed TSM which is Topic sentiment Mixture model where he reveals the inherent topical aspects in a weblog pool. The model could also describe the sentiment models that are very much suited for ad hoc topics. The empirical results shows that the approach is very effective for sentiments and topic facets. [12]. Summarizing the micro review is of great challenge, as the summary of the micro review should be representative,compact and readable. To produce effective summary of minimum description length framework, the author used approximation and heuristic algorithms on the dataset collected from Foursquare and Yelp [13].

The nature of short Text paves way for the sparsiy problem which brings a new challenge for topic modelling. Ziyu Lu et al. addressed the fragmented nature of short text by using pooling strategies,which aggregated user reviews based on venues and user. They put forward the finest topic model which integrates the influence of both venue elementary properties and user preferences and as superior significance on the topics of micro-reviews [14]. The Location based dynamic sentiment topic model was proposed which collectively models sentiment, topic, time and geopositioning particulars [15]. The model out performs well by discovering sentiments and topics from social forum and also examining their variations in different geolocations during natural disasters.

Ximing et al. in his work on short text modelling proposed a comprehensive topic model for short text named LTM which advances an flexible clustering process of short text and concurrently evaluates the inherent variables of interest [16]. [17] describes the weakly supervised system, where it performs the sentiment analysis on the movie review. The work involves the incorporation of knowledge from Wikipedia to filter out the non relevant text such that the accuracy of sentiment classification can be improved further. Paulo et al.devised a scheme to create Pseudo-document that is suited for topic modelling. His framework was based on word co-occurrence pattern to illustrate on the metric space and to calculate the distribution based on word vector representations [18–20]. Introduces the author Topic model which includes the authorship information. The author demonstrates the ATM model and author model and computes the similarity between authors and entropy of authors output. The JAST is a procreant process of considering the writing behaviour of an author.

JAST uses LDA to know topic preferences specific to the author and their emotional bond towards the topic [20, 21].

In [22], the model is named as Joint sentiment topic model where it as an additional layer called sentiment sandwiched between document and topic layers. In this model,words are related with both sentiment and topic,and the topics are linked with sentiment tags and the sentiment tags are linked with documents. The model (ASUM) was proposed by Y.jo et al. and it had the similar layers as JST,but the only distinctness between ASUM and JST is that, the prior assumes that the words or terms in the sentence exists to the alike language model whereas JST allows the terms to exist from disparate language model [23]. In paper [10] the sentiment layer in the model is decomposed into two levels. The first step is to determine whether the word is sentic word or topic word. And secondly if it is supposed to be a sentic word, discover its polarity oreintation. Also the model STDP requires a proir knowledge to differentiate the sentiment and topic words and this assumption will not suit all languages and domains. Shufeng et.al in his work proposed a WSTM i.e. joint sentiment for short text, which detects sentiment and topic jointly [24]. The model says that all the terms in the sentence belongs to same sentiment labels and also the words in the word pair set is also falling under the same topic [25]. Our model devised is a probabilistic generative model adapting the traditional flow of LDA model to learn and understand the sentiment and topic for short reviews of a specific user by straightly modelling the word pair co-occurrence pattern in a global scale.

Robust user sentiment biterm topic mixture model

RUSBTM focuses on the topic modelling which is an unsupervised technique and sentiment polarity categorization of corpus level. However the data sparse problem of short text in our model is overcomed by generating pseudo document by aggregating text based on the user and directly model the word pair co-occurrence pattern. The global generative process jointly models the sentiment and topic simultaneously. In this paper,the weakly supervised RUSBTM concentrates on the corpus level sentiment categorization and topic modelling. The model devised is a probabilistic generative model adapting the traditional flow of LDA model to learn and understand the sentiment and topic for short reviews of users by straightly modelling the word pair co-occurrence pattern in a global scale. As per our model, the entire corpus is transformed into a corpus of co-occurring word pairs, where each instance is sampled from the mixture model, that involves both sentiment and topic language models. By this model,the user sentiment topic distribution of the corpus is learnt. Experimental findings shows that RUSBTM model discovers the topics accurately by identifying the polarity of the reviewers than the state-of-the- art baseline methods.

Biterm Topic model

For many content analysis application,uncovering topic within the short text such as tweets, reviews, microblogsetc becomes an challenging task as normal traditional topic modelling will not work well with short text, as short text contains unedited and idiomatic text. To overcome the problems faced by short text,aBiterm model was introduced which directly discovers the topics by the creation of word-co-occurrence patterns in the whole of the corpus. A Biterm denotes a fully unorganized word- pair set Co-occurring in a short text .A short text always denotes to a correct text window holding a meaningful word co-occurrence. Suppose the collection of text consist of *n*words,then the model generates C_i^2 biterms in total.

Biterm extraction

The word bitermindicates an unordered word pair co-occurrence in a short text. Here a short text or context always denotes to a short definite window size over a sequence term. We can simply assume each reviews or tweets as an individual text unit. In such a scenario,in a document level any two distinctive words form a biterm. For instance, a document with four distinctive words would generate six biterms as follows:

$$BP(w1,w2,w3....wn)\Rightarrow\Big\{(w1,w2),(w1,w3).....(w1,wn),$$
$$(w2,w3),.....(w2,wn),$$
$$.$$
$$.$$
$$.$$
$$(wn-1,wn)\Big\}$$

where (-,-) is unordered. Once after the extraction of biterms is complete for an document, the entire corpus collection now changes into a set of biterms, which is completed in a single scan over the document.

Robust BTM

BTM models the multiplicative process of involving co-occurrence pattern of words in short text. However while generating the word co-occurrence patterns for corpus level words, user individualities are ignored. Hence user level personalization is included in BTM, where user based review aggregation is done to learn user specific topics from user based biterm aggregation. To some extent words are

distinguished into back ground words and topical words thereby a reduced biterm word sets is formed which better learns the topics. The proposed technique leverages the aggregation of user reviews and also incorporates the back ground topics in BTM.

Model description

Consider the corpus consists of a collection of users, locations, and a group of documents . Formally for the model, we use V to denote the sets of locations and U to denote the set of users, respectively. Any document $d \in D$ is a short text or message scripted by an author $u \in U$ in location $v \in V$.

Also, consider S *to* be the no of discrete sentiment labels, and T to be the total no of topics, assuming S and T to be the predefined constant values. Since each review is a short in nature,understanding them individually is not very effective and informative. Hence we introduced pooling methods to form a pseudo document of aggregated documents for each venue for each user. A_v is defined as the set of all users who have written reviews for the location v and d_v is refered as location document which is supposed to be the collection of all the documents written for location v. $N_d$is the number of words in location document d_v.The notations used in the model are represented in the Table 1. Our model generalizes traditional LDA by combinedly modelling user,sentiment, topic, etc. (Fig. 2).

The model represents the generative process of B,where B is the set containing extracted word pairs for the entire corpus. The next step in our proposed approach is to estimate the user sentiment topic distribution for each document.

Table 1 Meanings of the notation

Symbol	Description
$\boldsymbol{V}$	Set of locations
$\boldsymbol{U}$	Set of authors
$\boldsymbol{T}$	Total no of topics
$\boldsymbol{S}$	Sentiment Labels
$\boldsymbol{A_v}$	Set of all authors written for location V
$\boldsymbol{d_v}$	Set of all documents written for V
$\boldsymbol{z}$	Topic variable
$\boldsymbol{u}$	User variable
$\boldsymbol{c}$	Switch variable
$\boldsymbol{N_d}$	No of words in d_v
$\boldsymbol{\theta^v}$	Distribution of topic specific to V
$\boldsymbol{\phi_z}$	Distribution of words specific to topic z
$\boldsymbol{X^a}$	Distribution of sentiment specific to u
$\boldsymbol{\theta^a}$	Distribution of sentiments over topics
$\boldsymbol{X^v}$	Distribution of sentiment specific to V
$\boldsymbol{\alpha,\beta,\sigma,\eta}$	Dirichlet priors
$\boldsymbol{\gamma}$	Bernoulli prior
$\boldsymbol{w_i,w_j}$	Word pair set
$\boldsymbol{M}$	No of Word Pair

Model generative process

The generative process described captures the user preference,thereby including sentiment orientation which is a known label. To obtain the sentiment polarity for the term 퓌 in the review, Bing Liu sentiment Lexicon is used to obtain the polarity values for the words in the documents . From this we can derive a collection of <퓌,*s*> pairs of words and their polarity values for the venue document $\boldsymbol{d_v}$ for the set of all users $\boldsymbol{A_v}$.The sentiment of the authority users determine the predicted topic.

In the above process,for each topic Z under S, $\phi_{s,z}$is the Dirichlet distribution of sentiment over topic,drawn from the Dirichlet prior β. For each user ω_u is the Beta distribution of γ.

Generative process

1. For each topic z under s

   – Draw $\phi_{s,z} \sim Dir(\beta_{s,z})$.

2. For every user u

   a) Draw $\omega_u \sim Beta(\gamma)$
   b) Draw $x_u^{(a)} \sim Dir(\tau)$
   c) For every sentiment label s

   – Draw topic distribution

$\theta_{u,s}^{(a)} \sim Dir(\rho)$

3. For every location document d_v

   a) Draw sentiment distribution $x_{d_v}^{(v)} \sim Dir(\alpha)$

b) For every sentiment label s,

   – Draw topic distribution $\theta_{d_v,s}^{(v)} \sim Dir(\alpha)$

c) For each word pair $b \in B$ in location doc d_v

   – Draw $u \sim A_v$
   – Draw a switch $c \sim Bernoulli(\omega_u)$

If $c = 0$

   – Draw sentiment $s_{d_v i} \sim Mult\left(x_u^{(a)}\right)$
   – Draw a topic $z_{d_v i} \sim Mult\left(\theta_{u,}^{(a)} s_{d_v i}\right)$

If $c = 1$g

   – Draw sentiment $s_{d_v i} \sim Mult\left(x_{d_v}^{(v)}\right)$
   – Draw a topic $z_{d_v i} \sim Mult\left(\theta_{d_{vi}}^{(v)}, s_{d_v i}\right)$

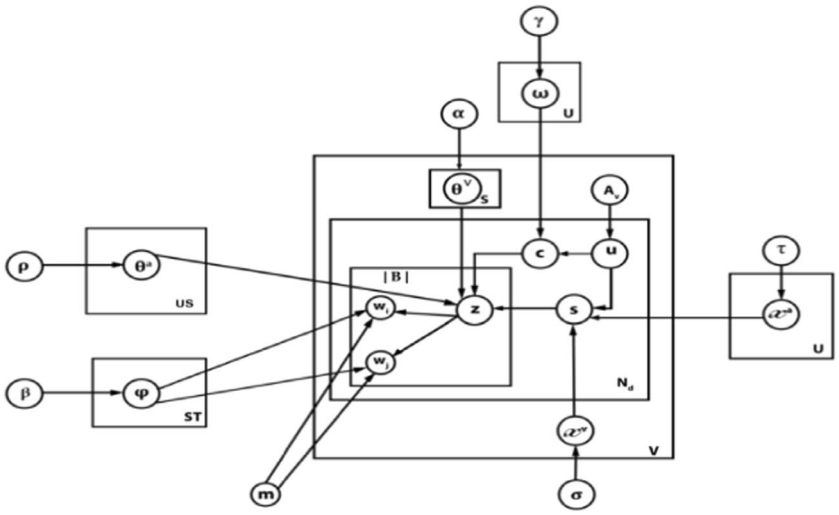

Fig. 2 Plate notation of the RUSBTM model

Draw two words:$\omega_i, \omega_j \sim Mult\left(\phi_{s_{dvi}, z_{dvi}}\right)$

Model inference

In the RUSBTM,the parameter α, β, σ, η and γ are hyper parameters .The latent topics and sentiments depend on the venue and the preference of the user. The variable ω is the parameter of Bernoulli distribution which influence the control between venue and the user. The switch variable c is introduced to determine whether the location document is inclined by the location or by the user.

For parameter inference,collapsed Gibbs Sampling technique is chosen to findout the unknown hidden variables $\{\phi, \omega, \theta^a, \theta^v, X^a, X^v, \psi\}$.Posterior distribution of the latent variable for every word is computed as follows.

Hidden Parameter $\{\phi, \omega, \theta^a, \theta^v, X^a, X^v, \psi\}$

$$P\left(U_{dvi,} = U, C_{dvi} = 0, S_{dvi,} = S\, Z_{dvi,} = z/\theta\right)$$

Where $\theta < U^{\neg dvi}, C^{\neg dvi}, S^{\neg dvi}, Z^{\neg dvi}, w, t, A_v >$.The symbol $\neg dv, i$ denotes a term excluding the current word dv, i

$$\alpha = \frac{{n_{uc}}^{\neg dvi}(0) + \gamma}{{n_{uc}}^{\neg dvi} + \gamma + \gamma'} \times \frac{{n_{us}}^{\neg dvi} + \tau_s}{\sum_s \left(n_{us} S^{\neg dvi} + {\tau_s}'\right)} \tag{1}$$

$$x = \frac{{n_{uc}}^{\neg dvi} + \rho z}{\sum_{z'} \left({n_{usz}}^{\neg dvi} + \rho z'\right)} \times \frac{{n_{szw}}^{\neg dvi} + \beta\omega}{\sum_{\omega'} \left({n_{sz\omega^1}}^{\neg dvi} + \beta\omega'\right)} \tag{2}$$

$$P\,(Udvi = U, Cdvi = 1, Sdvi = S, Zdvi = Z/\theta)$$

$$\alpha = \frac{{n_{uc}}^{\neg dvi}(1) + \gamma}{{n_{uc}}^{\neg dvi} + \gamma + \gamma'} \times \frac{{n_{vs}}^{\neg dvi} + \sigma_s}{\sum_{s'} \left(n_{vs} S^{\neg dvi} + {\sigma_s}'\right)} \tag{3}$$

$$x = \frac{{n_{vsz}}^{\neg dvi} + \alpha z}{\sum_{z'} \left({n_{usz}}^{\neg dvi} + \alpha z'\right)} \times \frac{{n_{szw}}^{\neg dvi} + \beta\omega}{\sum_{\omega'} \left({n_{sz\omega^1}}^{\neg dvi} + \beta\omega'\right)} \tag{4}$$

The $n_{uc}(0)$ and$n_{uc}(1)$ are the no of counts that $c = 0$ and $c = 1$ are sampled for the user. $n_{u,\,s}$is the no of times the sentiment $s$sampled from the Distribution of sentiment specific to user $\boldsymbol{X^a}$ and $n_{v,\,s}$is the no of times the sentiment s sampled from the Distribution of sentiment specific to venue $\boldsymbol{X^v}$. $n_{u,\,s,\,z}$ is the no of times the topic z is sampled from the distribution specific to user u and sentiment s . $n_{v,\,s,\,z}$is the no of times the topic z is sampled from the distribution specific to location v and sentiment label s. $n_{s,\,z,\,w}$ is the no of times the word w sampled from the distribution specific to sentiment s and topic z .

After collapsed Gibbs Sampling, $\{\phi, \omega, \theta^a, \theta^v, X^a, X^v, \psi\}$ are estimated as follows:

$$\hat{\theta} v, s, z = \frac{nvsz + \alpha z}{\sum_{z'} \left(nvsz' + \alpha z'\right)} \tag{5}$$

$$\hat{\theta} u, s, z = \frac{nusz + \rho z}{\sum_{z'} \left(nusz' + \rho z'\right)} \tag{6}$$

$n_{u,\,s,\,z}$ is the number of times the topic z is sampled from the distribution specific to user u and sentiment s . $n_{v,\,s,\,z}$is the number of times the topic z is sampled from the distribution specific to location v and sentiment label s.The hidden parameter X^a, X^v which is nothing but the distribution of sentiment specific to user u and venue v

$$\hat{X}d, v, s = \frac{nvs + \sigma s}{\sum_{s'}\left(nvs^{'} + \sigma s^{'}\right)} \tag{7}$$

$$\hat{X}\alpha, u, s = \frac{nus + \tau s}{\sum_{s'}\left(nus^{'} + \tau s^{'}\right)} \tag{8}$$

The distribution of the word specific to sentiment label s under the topic z is given by the equation below

$$\hat{\phi}_{\alpha,s,z,w} = \frac{n_{szw} + \beta\omega^{'}}{\sum_{\omega'}\left(n_{szw'} + \beta\omega^{'}\right)} \tag{9}$$

$$\hat{W}^{au} = \frac{n_{uc(1)} + \gamma}{nuc + \gamma + \gamma^{1}} \tag{10}$$

After necessary training of the above RUSBTM model,the objective is to estimate $p(z|\, u,\, v)$,i.e. the probability for all topics z conditioned on a new pair $\boldsymbol{w_i}$, $\boldsymbol{w_j}$ of the user for the venue.

Since word pair is used. It is necessary to present an approximate estimation of both sentiment and topic for a document as an extension step.

$$S = \arg\max X^{a}X^{v} \tag{11}$$

We find out the probability of topic z for the word w_i in doc d as

$$P(z/w_i) = \frac{\sum_B P\left(\frac{z}{b}\right) XP\left(\frac{wi}{b}\right)}{\sum_B P\left(\frac{wi}{b}\right)} \tag{12}$$

Where

$$P(z/b) = \frac{\sum_1 P(s)\;\; P\left(\frac{z}{l}\right) P\left(\frac{w_i}{l,z}\right) P\left(\frac{w_i}{l,z}\right)}{\sum_z\left(\sum_s P_d(S).P^{\left(\frac{z}{s}\right)}.P\left(\frac{wi}{S,z}\right).P\left(\frac{wi}{l,z}\right)\right.} \tag{13}$$

l will be replaced with S

$$P\left(\frac{z}{b}\right) = \frac{\sum_z P_d(s) P\left(\frac{z}{s}\right) P\left(\frac{w_i}{s,z}\right) P\left(\frac{w_i}{s,z}\right)}{\sum_z\left(\sum_s P_d(S).P\left(\frac{z}{s}\right).P\left(\frac{w_i}{s,z}\right).P\left(\frac{w_j}{s,z}\right)\right.} \tag{14}$$

Similarly probability for an word w_i with its sentiment label S can be calculated as

$$P\left(\frac{s}{w_i}\right) = \frac{\sum_{MM} P_d(s) P(s) P\left(\frac{w_i}{m}\right)}{\sum_M P\left(\frac{w_i}{m}\right)} \tag{15}$$

where $_M$is the total no of word pair in the corpus and $P\left(\frac{w_i}{m}\right)$ is the probability of the word w_i appeared in m.

Experimental analysis and evaluation

Data sets

We performed the experiments in the Restaurant review domains, where the dataset is obtained from trivago, India. The data is in the form of a csv file which contains over 1000 reviews. The csv file is structured into 4 columns with title of the review,Review text, Sentiment of the review and finally the rating percentage. The dataset chosen for the work contains 1000 reviews with 192 authors with atleast 1 review per author. In order to proceed with the work, the authors are retained who have given reviews for about 10 times. By this way the dataset is reduced to have 857 reviews with 18 authors. Each reviews is rated by the user on the rating scale of 1 to 3.

Data pre-processing

In order to reduce the low quality and incomplete reviews, we Pre-processed the reviews through the following sequence of steps: (a) removing non-English characters, stop words in the review (b) transforming letters into lower case (c) deleting duplicate reviews (d) filteringout reviews with word length less than 5 (e) The model uses almost 85% of the data per user to know the parameters during inference. The Table 2 depicts the statistics of the dataset used.

1) No of users $|A|$: 18
2) Number of terms $|W|$ after removing the stop words:857

Model priors

In the implementation, the number of topics K is set to 15 and the values of prior hyper parameter such as α and β are taken as 50/K and 0.01 respectively . The model requires the initialization of the parameter K=15, the number of Topics. The number of users, review ratings and percentage of reviews are defined as per the dataset used. Model perplexity is used to initialize the value of K which is considered to be an essential measure for Topic modelling. Lexical classification is considered to be an baseline for the work. Textblob library is also used to arrive the polarity values for an distinct words in the review. It is an library for diving into the common natural

Table 2 Dataset statistics

Dataset	Authors	Avg Rev./ Author	Rev, pos,Neg,Total			Avg Rev. Length	Avg Words/ Rev
Restuarant reviews	18	7	R1 64	R2 130	R3 806	200	25

language processing task such as sentiment analysis, noun phrase,part-of-speech tagging etc. Bing Liu sentiment Lexicon is used to derive the polarity values of terms both positive and negative terms. RUSBTM uses the initialization of the parameter K i.e. the no of Topics. Usually model perplexity is used to initialize the value of Kwhich is the important measure to topic modelling .The value of the perplexity indicates the likelihood of the model, so the value of the parameter of the model is chosen based on the minimum perplexity of the model.

Evaluation metrics

The main task of Topic modelling which is an unsupervised technique is extracting topics from the distribution of words. The topics extracted can be evaluated both in a qualitative and Quantitative way. The qualitative approach is to observe the extracted words by the topic model. The quantitative approach can be evaluated by various topic measures such as perplexity, Topic Coherence both Intrinsic and Extrinsic measure, purity etc. The Sentiment topic mixture model which relies on the discovery of sentiment specific topic discovery can be measured or checked by qualitative evaluation method by observing the extracted topics based on the sentiment orientation of the words. The quantitative approach of measuring the discovered topic is examined by checking the various measures such as Perplexity,coherence etc.

Perplexity

Perplexity is defined as the measures which tells how well the given model fits the test data ie how well the probability model predicts the sample, thus it tries to evaluate or asses the predictive power of the topic model. A lower the value of the perplexity score indicates increase in predictive power. For the given value of K,estimate the LDA model. Then given the theoretical word distribution for a topic, compare that to the actual distribution of words in our documents. Perplexity is a function that calculates the value of the given model. The measure is computed for varying values of K for the model.

The lower perplexity value for the model is considered the best. The graph is plotted with perplexity against the number of Topics,the lower the perplexity score better the number of Topics. But lowest score at times leads to the decision making process, where the structure of topics for the given documents may not make sense. So domain knowledge of the dataset is required for effectively evaluating the number of Topics of the LDA models.

$$PP(WW) = \sqrt[N]{1/P(w1, w2....wN)} \tag{16}$$

Topic coherence:(PMI)

Topic modelling enables us with the technique to understand, systematize and summarize large number of short text. Topic

Table 3 List of positive words discovered by RUSBTM

Rooms	Variety	Stay	Comfortable
relatively	furnished	service	service
offer	enjoyed	good	neat
greater staff	best	gym	spacious
windows	gracious	condition	served
amenites	quality	visible	friendly
attenders	conducted	amenities	experienced
happy	experience	choices	business
multiple	parking	staff	center
loved	professional	price	area
prime	lights	personnal	courteous

Table 4 List of negative words discovered by RUSBTM

Service	Stay	Cost	Food
ambience	close	costlier	nasty
unkept	short	high	smell
unexperienced	unpleasant	overpriced	bad
awful	uncomforatable	nominal	poor
less	time	price	meals
bad	usage	increase	drenched
problem	place	expensive	taste
change	dirty	worthless	odour
service	left	costly	buffet
poor	wait	money	less

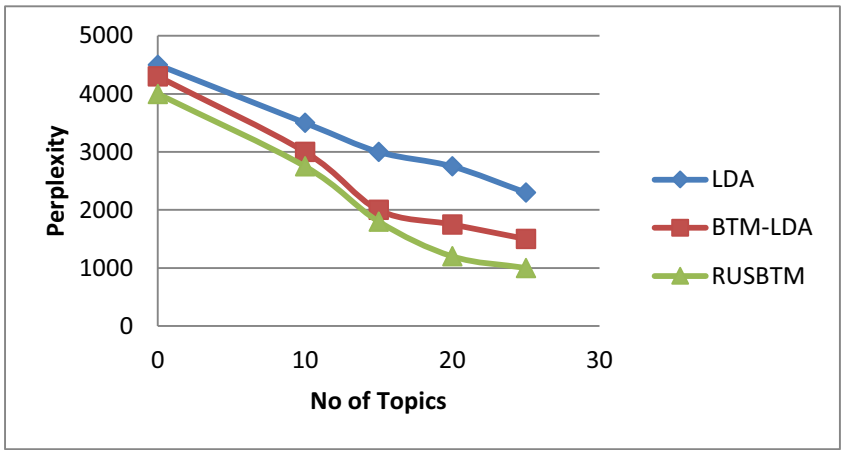

Fig. 3 Perplexity vs no of topics

model learns topic in an unsupervised way by automatically arranging set of words from an unlabelled documents. The LDA technique used to extract the topics does not guarantee to be well interpretable, so an metric called coherence measures are computed to differentiate between good and bad topics. The measure coherence is computed from the coherence framework which is defined as an compositional parts that can be summed. The parts are combined into dimensions that it spans the whole configuration space of the coherence measure. The coherence measures are broadly classified into two measures .

Intrinsic measure For an ordered word set,the intrinsic measure UMass is used to compare a distinct word with its predecessor and successor words respectively, ie it makes use of pairwise score function. The pairwise score function is computed as the conditional log probability towards the smoothing count inorder to avoid computing the log of zero problem.

Extrinsic measure For an unordered set the UCI measure defines as, each single distinct word is paired up with every other distinct word. The UCI measure uses Pointwise mutual information .The coherence score is computed to be the summation of pairwise scores of the words $w_1, \ldots w_n$ for the topic. This measure is an important evaluation metric used to determine the number of topics. The LDA model should be built with different values of K,and the value of K is chosen such that it has the highest coherence score.

The graph is plotted with coherence score against the number of Topics and the value of number of Topics should be accordingly chosen such that it has the highest coherence score. Always Topic coherence is used in topic modelling to take better decision. The UMass and the C_V topic coherence tells the optimal no of topics by suggesting a number called coherence score which gives the interpretability of these topic.

Sentiment identification

The sentiment identification is always measured by the accuracy of the sentiment by different models. Sentiment accuracy is always computed against human label or it

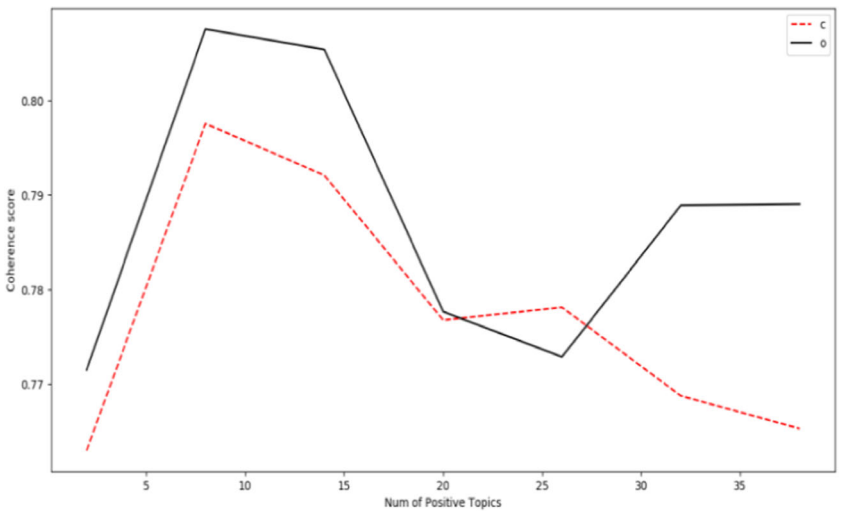

Fig. 4 Coherence score vs positive topics

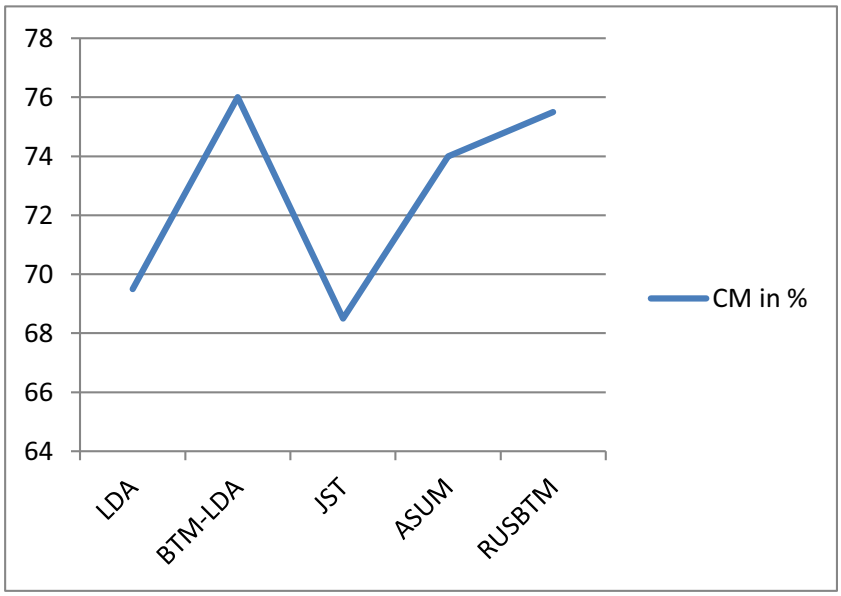

Fig. 5 coherence measure for models

can be measured by conducting paired t-test for sentiment label and the value of p value is less than 0.05.

Results and discussion

The performance of the model is analysed by three different ways .i.e. 1) Based on the discovered topics 2) Based on the sentiment specific topic discovery and 3) Sentiment identification. The discovered topic is assessed both by the qualitative as well as quantitative way. For topic models, since the model RUSBTM is designed for review analysis on venue dataset, it is observed that the topic number setting is set to 15 for discovering topics. Tables 3 and 4 shows the sentiment topic discovered words by RUSBTM.

The Tables 3 and 4 states the list of top 10 words listed under the positive and the negative topics. For easier understanding the labels are assigned for the topics. The table states that the words are properly well listed under the topics and also the words are more coherent within the topic. For example column 4 in Table 3,the positive topic is "comfortable" and the listed words are inferred as the place is comfortable for "service", "neat", "spacious", "served", "friendly", "center", "area" etc. and the column 4 in Table 4 states the negative topic as "food ", and the list of words distributed under the topic food are "nasty", "smell", "bad", "poor", "meals", "taste", "odour", "less" etc. Usually some noise words would occur in the reviews, which should be appropriately removed in the pre-processing steps of the data. It is also necessary to remove more common topics for fine grained analysis.

Tables 3 and 4 classified words based on their sentiment polarity and their co-relations. The Tables 3 and 4 clearly shows the sentiment specific positive and Negative topics. The restaurant review taken for the experiment as positive topics such as rooms,stay, comfortable,varietyetc. The words under the positive topic rooms are relative,offer, amenities, multiple, loved, prime etc. Hence the listed words under the topic rooms gives the reason why it is positive. Similarly the words under the negative topic food are nasty, smell, bad,poor,meals, taste,odour,lessetc. The stated model RUSBTM clearly captures the positive and negative aspect respectively. The positive part of the restaurant are rooms,stay,variety,comfortable and the negative part are service,cost,food and stay .

Explicitly, we selected 500 documents from the venue dataset, and categorised each document or reviews under positive, negative and neutral labels manually. We measured the overall performance of the proposed model under the positive and negative labels.

Perplexity is a function that calculates the value of the given model. The measure is computed for varying values of K for the model. Lower perplexity value for the model is considered the best. Figure 3 shows the graph plotted with perplexity against the number of Topics,the lower the perplexity score better the number of Topics. It clearly states that the RUSBTM model arrives a lesser perplexity values when compared to other baseline models.

Figure 4 shows the coherence score plotted against the postive topics and c and o in the graph shows the plotting of any two users taken at random. The graph clearly states that the value of K is chosen such that it has the highest coherence score.

Table 5 Coherence measure

Models	LDA-baselines	BTM-LDA	JST	ASUM	RUSBTM
Coherence	69.5	76	68.5	74	75.5

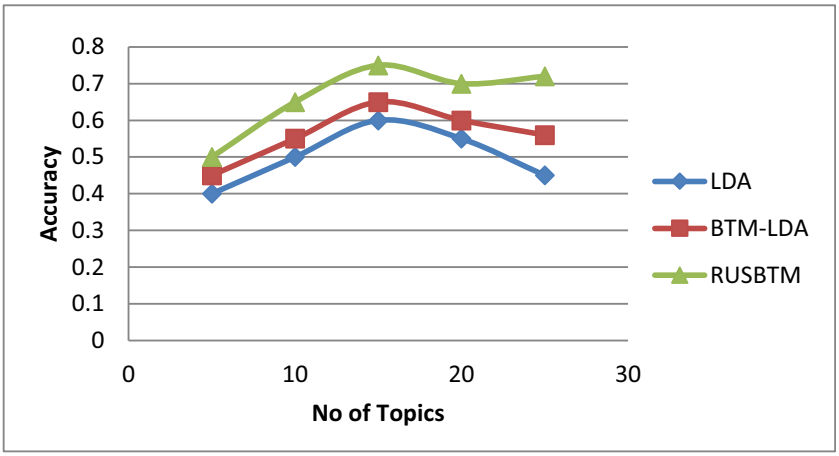

Fig. 6 Accuracy vs no of topics

Coherence measure is stated as the proportion between the total no of words which are relevant to the total count of all words. The performance of RUSBTM is almost comparable with BTM and both of them outperform with other state of art methods. The RUSBTM accomplishes the same measure as BTM model and it tackles the data sparse problem of short text to a greater extent. The RUSBTM assures that the two words of the pair set belong to the same topic and sentiment. The constrains that the two words belonging to the same topic and sentiment is more applicable to the restaurant, product reviews.

The Fig. 5 and Table 5 shows the coherence measure of RUSBTM as it is almost comparable with BTM and both of them outperform when compared to other state of art models. The RUSBTM assures that the two words of the word pair belong to same topic and sentiment

In ASUM model,complete words in the review belongs to the same topic whereas RUSBTM allows that the generated word pairs are from different topics and sentiments.

Figure 6 reports the accuracy of the model on influence of the topic numbers and Fig. 7 discusses the sentiment accuracy of the models and to have a unified setting,we have set the topic number as 15. The work involved Bing Liu sentiment Lexicon and Textblob so when conducted paired-t-test for sentiment label,the value of p value is less than 0.05.

The RUSBTM model has better accuracy compared to other baseline models,the reason behind is that the words in the pair set belong to the same topic and sentiment. as seen in Table 6. The JST model has lesser accuracy because of the fact that every word in the model have different topics. The accuracy of the ASUM model is comparatively same as the RUSBTM model beause it states that the every word from the

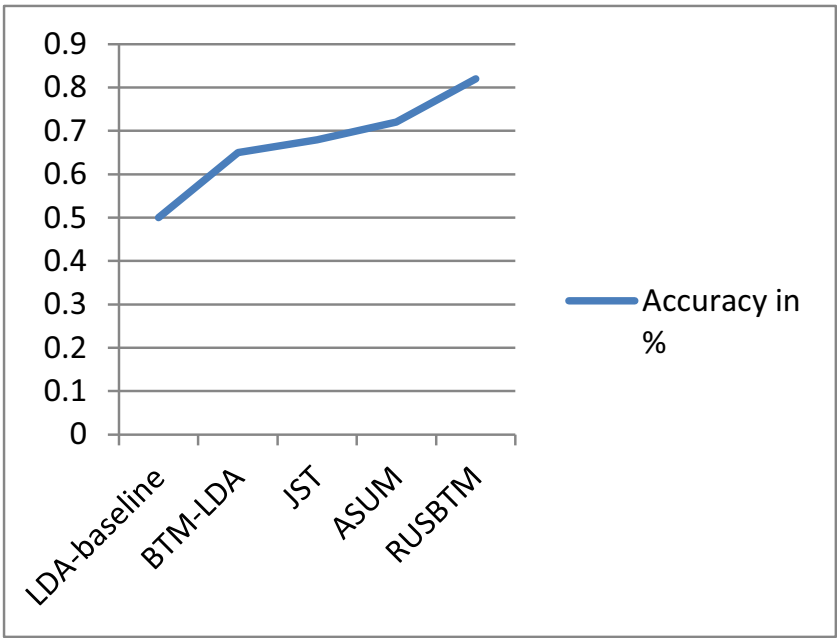

Fig. 7 Sentiment identification of the models

Table 6 Sentiment accuracy

Models	LDA-baseline	BTM-LDA	JST	ASUM	RUSBTM
Accuracy	0.5	0.65	0.68	0.72	0.82

same sentence belong to the same topic. So it is said that the two models are more sensitive to the topic numbers. The performance of JST and ASUM model depend on the dataset that is being used for the experiment and they are not suitable for the short text,due to text sparse problem and they does not sample enough sentence to obtain the estimated parameters. RUSBTM outperforms other models because it allows that the words in a sentence can have different topics. Hence it demonstrates the effectiveness of the method.

Conclusion and future works

The Robust user sentiment Biterm Topic model (RUSBTM) is a novel approach which incorporates users and their sentiment orientation views for effective Topic Modelling using Biterms or word-pair which produces both positive topic-words, negative topic- words, user- positive topic, user-negative topics, venue item- topic distribution simultaneously. Extensive Analysis on the venue datasets reveals that our model is more effective and advantageous. The results obtained by our model reveals that the extracted topics are more coherent and informative. Also our method uses the accurate sentiment polarity techniques to exactly capture the sentiment orientation of the user. So the model correctly reveals the user preference which could be utilized for variety of applications such as content characterising,content recommendation, user interest profiling, semantic analysis etc.

Compliance with ethical standards

Conflict of interest The authors declare that this article content has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Blei, D. M., Andrew, Y. N. G., and Jordan, M. I., Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022, 2003.
2. Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., and Xiong, H., Topic modeling of short texts: a pseudo- document view. KDD '16, 2016.
3. Hong, L., and Davison, B. D., Emprical study of topic modeling Twitter. In: *Proceedings of SOMA*, pages 80–88. ACM, 2010.
4. Devi, G. U., Priyan, M. K., & Gokulnath, C. (2018). Wireless camera network with enhanced SIFT algorithm for human tracking mechanism. International Journal of Internet Technology and Secured Transactions, 8(2), 185–194.
5. Cheng, X., Yan, X., Lan, Y., and Guo, J., BTM: topic modelling over short text. *IEEE Trans. Knowl. Data Eng.* 26(12), 2014.
6. Xia, Y., Tang, N., Hussian, A., and Cambria, E., Discriminative bi-term topic model for headline-based social news clustering. *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research society conference.*
7. Vijayakumar, V., Priyan, M. K., Ushadevi, G., Varatharajan, R., Manogaran, G., & Tarare, P. V. (2018). E-health cloud security using timing enabled proxy re-encryption. *Mobile Networks and Applications*, 1–12.
8. Li, W., Feng, Y., Li, D., and Yu, Z., Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm. *Autom. Control. Comput. Sci.* 50(4):271–277, 2016.
9. Pan, Y., Yin, J., Liu, S., and Li, J., A biterm-based Dirichlet process topic model for short texts. International Conference on Computer Science and Service system, CSSS, 2014.
10. Li, C., Zhang, J., Sun, J. T., and Chen, Z., Sentiment topic model with decomposed prior. In: *Proceedings of SDM*, pages 767–775. SAIM, 2013.
11. Mehrotra, R., Sanner, S., Buntine, W., and Xie, L., Improving lda topic modelsfor microblogs via tweet pooling and automatic labeling. In: *Proceedings of SIGIR*, pages 889–892. ACM, 2013.
12. Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C., Topic sentiment mixture:modeling facets and opinions in weblogs. In: *Proceedings of WWW*, pages 171–180. ACM, 2007.
13. Nguyen, T.-S., Lauw, H. W., and Tsaparas, P., Review synthesis for micro-review summarization. In: *Proceedings ofthe Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, 2015.
14. Lu, Z., Mamoulis, N., Pituoria, E., and Tsaparas, P., Sentiment-based topic suggestion for micro reviews. Proceedings of the Tenth International AAAI conference on Web and social media, 2016.
15. Kumar, P. M., Devi, U., Manogaran, G., Sundarasekar, R., Chilamkurti, N., & Varatharajan, R. (2018). Ant colony optimization algorithm with Internet of Vehicles for intelligent traffic control system. *Computer Networks, 144*, 154-162.
16. Manogaran, G., Shakeel, P. M., Hassanein, A. S., Priyan, M. K., & Gokulnath, C. (2018). Machine-learning approach based gamma distribution for Brian abnormalities detection and data sample imbalance analysis. IEEE Access.
17. Mukherjee, S., and Bhat-tacharyya, P., Wikisent: weakly supervised sentiment analysis through extractive summarization withwikipedia, ECML PKDD'12, pp. 774–793, 2012.
18. Bicalho, P., Pita, M., Pedrosa, G., Lacerda, A., and Pappa, G. L., A general framework to expand short text for topic modeling. *Inf. Sci.* 393:66–81, 2017.
19. Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P., The author-topic model for authors and documents. In: *Proceedings of the 20th Conference on Uncertainty in ArtificialIntelligence*, UAI '04, 487–494, 2004.
20. Mukherjee, S., Basu, G., and Joshi, S., Joint author sentiment topic model. In: *Proceedings of the 2014 SIAM International Conferenceon Data Mining*, pages 370–378. SIAM, 2014.
21. Balan, E. V., Priyan, M. K., Nath, C. G., & Devi, G. U. (2014, December). Efficient energy scheme for wireless sensor network application. In 2014 IEEE International Conference on

Computational Intelligence and Computing Research (pp. 1–5). IEEE.
22. Lin, C., He, Y.. Joint sentiment/topic model for sentiment analysis. In: *Proceedings of CIKM*, pages 375–384, Hong Kong, China. ACM, 2009.
23. Jo, Y., and Oh, A. H., Aspect and sentiment unification model for online reviewanalysis. In: *Proceedings of WSDM*, pages 815–824. ACM, 2011.
24. Sundarasekar, R., Thanjaivadivel, M., Manogaran, G., Kumar, P. M., Varatharajan, R., Chilamkurti, N., & Hsu, C. H. (2018). Internet of things with maximal overlap discrete wavelet transform for remote health monitoring of abnormal ECG signals. *Journal of medical systems, 42*(11), 228.
25. Priyan, M. K., & Devi, G. U. (2019) A survey on internet of vehicles: applications, technologies, challenges and opportunities. International Journal of Advanced Intelligence Paradigms 12 (1–2):98