Received: April 28, 2020. Revised: May 19, 2020.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Two-Stepped Feature Engineering Process for Topic Modeling Using Batchwise LDA with Stochastic Variational Inference Model

Sujatha Arun Kokatnoor[1]* **Balachandran Krishnan[1]**

[1]*Department of Computer Science and Engineering, School of Engineering and Technology, CHRIST (Deemed to be University), Bangalore, India*
* Corresponding author's Email: sujatha.ak@christuniversity.in

Abstract: Online ratings and customer feedback on hotel booking websites support the decision-making process of the customer as the reviews provide a deeper understanding about all aspects of a hotel. Consequently, review and rating analyses are of great interest to consumers and hotel owners for the hotel related social media services. The key challenge, however, is to make the wide variety of information accessible in a simple, fast and relevant way and the solution is Topic Modelling and Opinion Mining. Common approaches like Latent Semantic Analysis (LSA) and Hierarchical Dirichlet Process (HDP) have order affects. If the input dataset is shuffled then different topics are generated leading to misleading results. To overcome this, a two-stepped feature engineering process is used: first step is to use a TF-IDF with modified trigrams calculation followed by the second step in removing weak features from the corpus thereby reducing the dimensionality of the Vector Space Model (SVM) for efficient Topic Modeling and sentiment analysis of the considered corpus. Sentiment score is calculated using VADER tool and Topic Modeling is done with Batch Wise Latent Dirichlet Allocation (LDA) using Stochastic Variational Inference (SVI) model. The modified trigrams included calculation of probabilities of words not only in the backward direction but also the probability calculation of the next two words of the target word thereby retaining its context information. The proposed method using Batchwise LDA with SVI along with two-stepped feature engineering process considerably improved its performance when compared to LSA and HDP models due to the fact of identifying hidden and relevant topics in terms of their optimized posterior distribution in hotel reviews dataset. The Batchwise LDA with SVI improved its performance by 3% in terms of its coherence values by using two-stepped feature engineering process and by 9% and 4% increase when compared with LSA and HDP models respectively.

Keywords: Sentiment analysis, Topic modeling, Latent dirichlet allocation, Hierarchical Dirichlet process, Latent semantic analysis, Feature engineering, Vector space model, Stochastic variational inference, Feature engineering.

1. Introduction

Online data has expanded exponentially across the internet over the last few years. People today use various forms of social media, news outlets, forums, and so on as information sources and to communicate in a number of ways. Together with this huge amount of data, it is important to identify and sort these data. Comprehensive methods are used to organize the classification of these results. In the past few years, identifying new events and monitoring current developments is an area of concern for many researchers from the vast amount of data streams from different available social media. The major result in the huge increase in user-generated content in the tourist field is the substantial reduction in the asymmetries of information between demand for tourism and supply. Useful users have to trust the numerous intermediaries and high quality web sites associated with tourism and tourism facilities before review sites like Holidaycheck.com or Booking.com. Rating options on different review sites have significantly changed the situation to allow travellers to share their experiences on their journeys and stay online with other potential travellers. Since then, transparency has dramatically increased in the quality of tourism services, such as hotels, restaurants,

destinations and all travel deals [1]. Various information extraction, information reduction, and data mining techniques can be used to automatically collect and process user reviews based on the publicly available user feedback [2]. The collected and processed data can thus be used from the user generated content as a valuable insight into various strategic or operational business choices. With this an important work is of researcher's interest is Topic Detection or Topic Modeling which is a part of the Sentiment Analysis [3].

Topic Modeling is defined as the task of allocating the topics to the unlabelled textual dataset. The Topic Modeling is a branch of unsupervised machine learning approach that uses a text document that can best explain the information available with the assistance of several other topics. Topic Modeling can be used for several applications as shown in Fig. 1.

Before proceeding with Topic Modeling, the unstructured and unlabelled text corpus has to be converted into a Vector Space Model (VSM). Commonly used methods like Bag of Words (BoW) and CountVectorizer have limitations of generating a high dimensional VSM. This is due to the total number of occurrences of each word in a text document and the model fails to appropriately map these words to the chosen topics thereby providing poor representation of the feature vector.

There are many ways to do Topic Modeling. They are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Hierarchical Dirichlet Process (HDP) and Latent Dirichlet Allocation (LDA) with Gibbs Sampling approaches but with certain limitations. LSA vectors require comprehensive storage. It performs very well for long documents since each document is represented in a limited number of context vectors. Nevertheless, it requires additional storage and calculation times because of the large volume of data, which reduces LSA performance [4]. Also LSA's outcomes, words associated to topics are difficult to interpret and LSA partially captures polysemy, the words which has multiple meanings. PLSA and LDA with Gibbs Sampling produces different topics if the order of the text data is changed leading to misidentification of topics of interest. PLSA uses large Eigenvectors, it cannot be attributed to the only applicable groups and documents in smaller communities. The PLSA model doesn't generalize to new and unseen documents and is prone to overfitting. With LDA using Gibbs sampling, the shape of the distribution depends on the long convergence period, in particular with the dimensionality and convergence of data. LDA also has difficulty locating the posterior for each variable with Gibbs sampling.

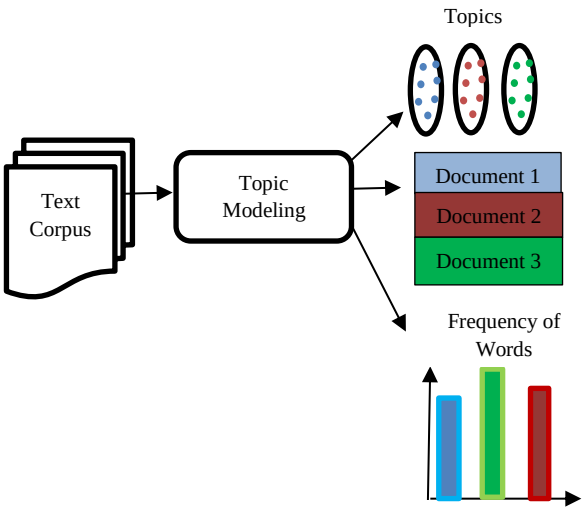


Figure. 1 Applications of topic modeling

In this paper, a two-stepped feature engineering is proposed to address the limitations of BoW and CountVectorizer methods for pre-processing the text corpus. In the first step, TF_IDF with trigrams is used to reduce the size of the dataset considered followed by the calculation of sentiment scores using VADER tool as the customer's reviews are directly associated with their sentiments about their stay in the hotel. In the second step, weak feature vectors are removed as they are constituted using low frequency terms which rarely occur in the document. This feature engineered VSM is now provided as an input to LDA using Stochastic Variational Inference (SVI) for better tracking and detection of the topics.

The rest of the paper is organized as follows. Related literature review work carried with respect to feature engineering and Topic Modeling is covered in section 2. The proposed architecture and methodology is covered in section 3. Section 4 gives the results and discussions about the proposed work and section 5 gives conclusion and future scope of the work proposed in this paper.

2. Related work

In this section, the literature review was carried out various Topic Modeling approaches.

The combination of the topic model and the Continuous bag of words (Cbow) process, known as the Cbow Topic Model (CTM) [5] to detect topics was proposed to provide summarization of the text corpus collected from social networks. This method introduced a new idea for the embedding topic model. A classic Cbow word vectorising approach was also used to cluster the target social network text dataset.

This helped to learn efficiently about internal associations between words and minimized the size of the input texts. Cbow unfortunately cannot pick up polysemy as they appear to represent one word as one vector. Cbow also fails to identify the combined word phrases in a text corpus.

The different methods of extracting hidden topics from the text (corpus) were investigated using the topic models such as LSA, PLSA and LDA [6]. The researchers expressed their challenges in dealing with short texts having difficulties in interpreting the topics present in the corpus due to the limited space available on social media web pages which resulted in sparse and noisy text reports. Since social media sites have a small space, text analyses were loaded with noisy and sparse words. The LDA with parallel-supervised model was proposed to achieve high accuracy results with quicker response time along with supervised Latent Dirichlet Allocation (SLDA) and Parallel Latent Dirichlet (PLDA) [7]. When the process time of Topic Modeling is increased, while the speed is increased, accuracy is decreased. Also LDA is an unsupervised tool which does not properly categorize and predict the topics.

To geographically map online discussions over time and in order to recognize subject situations in unstructured text, a distributed geo-aware streaming LDA was proposed [8]. To test the model it was used and implemented during portions of the US presidential election in 2016 for the automated identification and spatial monitoring of electoral topics. The locations were shown to be related to the actual polling places and the model offered a better description of geo-location than a keyword-based method. The disadvantage was that the proposed method suggested a closed world presumption, that new places of interest weren't known and that they were not context-sensitive.

A topic model, the twitter hierarchy Latent Dirichlet Allocation (thLDA) [9] was proposed to automatically mine the hierarchy of tweet topics, which can also be used on tweets for text Online Analytical Analysis (OLAP). Seeking a more efficient aspect also included the word2vec used by thLDA to evaluate the semantic relation between words in tweets. It was observed that the proposed model failed to capture correlations present in the corpus.

A topic-specific sentiment lexicon was constructed and the final sentiment similarity was obtained by using the Spectral Clustering model for clustering the words present in the corpus. It was based on the sentiment scores and the sentiment relationship graph for obtaining the topic-specific sentiment lexicon, namely, STCS (Spectral Clustering based Topic – Specific) lexicon [10]. The spectral clustering needs the construction of Eigenvectors. If K-Nearest Neighbours graph is used then all the created Laplace matrices will be sparse and this speeds up the convergence process else with smaller Eigen gap, the algorithm computing the first set of Eigenvalues doesn't converge leading to a slower process.

In social networks, words' sentiments are different from their useful topics which leads to a problem in building a sentiment lexicon. The Latent Dirichlet allocation [11] was used for the extraction and validation of the topics of interest in the dataset as a generative, Bayesian, hierarchical statistical model. The findings supported existing literature, including location and quality of service. In addition, the variations in the topics of concern between the different characteristics of the accommodations in terms of metropolitan vs. rural and type of accommodation were noticed. The inherent limitation of a country-based study, however, meant that other countries did not generalize the results.

A probabilistic Supervised Joint Aspect and Sentiment Model (SJASM) was proposed to model user generated reviews and to classify semantic topics and topic level sentiments from review data and also to predict overall sentiments from the reviews [12]. SJASM represented the opinion pairs of each review document and was used to construct feature vectors and the corresponding opinion words from the hidden topic responses and sentiments. The drawbacks with this approach are that probabilistic regression for sentiment analysis which necessitates determining in advance the number of latent topics for analysis of user reviews in the corpus.

A hierarchical supervision topic model was proposed for the construction for higher-level classification tasks of a Topic-adaptive Sentiment Lexicon (TaSL). This proposal used sentiment lexicons as a useful prior knowledge for Topic Modeling and Sentiment Analysis [13]. However, lexicons of sentiments were constructed to ignore the variation of the sentiments polarities in various subjects or fields.

A probabilistic graphic model called Sentiment-aware Multi-modal Topic Model (SMTM) was proposed [14] to leverage the latent semantics of multi-modal knowledge on the web site in order to solve the question of customized travel by using the data available from social media in the real world. The proposed method distinguished itself from previous methods, in that the topics from tourism and attraction domains were independently analysed for the communication of semantics for tourist topics and

themes. The main limitation was the consideration of a fixed number of topics on three sentiment opinions: neutral, positive and negative.

Similar tweets were combined together into single documents for the purpose of pooling [15]. This overcame the traditional LDA model in terms of its topic coherence. The limitation of this approach was that it worked only for a single hashtag.

Parallel Online Supervised LDA was proposed for large scale textual datasets [16]. This approach used SVI learning method for increasing the speed of training time and parallel computing through map reduce framework model was used for endorsing cloud computing capacity and the processing of huge dataset. The input to this model is the labelled dataset which is a limitation as most of the reviews obtained for this research work is unstructured and unlabelled. Without feature engineering, the accuracy of the model poorly performs.

Various types of topic detection techniques were identified and the results were calculated. Topic detection methods were classified in five categories: clustering, frequent pattern mining, simulation, matrix factoring and probabilistic models [17]. For clustering techniques, nine different approaches were evaluated. They were Sequential K-Means, Spherical K-Means, Kernel K-Means, Scalable Kernel K-Means, Incremental Batch K-Means, DBSCAN, Spectral Clustering, Document Pivot Clustering and Bngram. Moreover, for matrix factorization techniques, five different techniques were analysed and they were Sequential Latent Semantic Indexing (LSI), Stochastic LSI, Alternating Least Squares (ALS), Rank-One Downdate (R1D), and Column Subset Selection (CSS). However, the discovered topics were hidden and did not have clear meaning.

In order to identify and monitor topics in conversational contents of social media networks, a Computational Dynamic Latent Dirichlet allocation (CDLDA) model was proposed. Topic detection and tracking is essential, especially for spoken interactions and for conversational communication. During conversational communication, language processors must identify various topics in these materials, as the topic transitions occur frequently. In view of its structure, the dynamic models were used to capture the sequence in spoken content of two adjacent topics [18]. However, CDLDA models were not sufficiently efficient enough for processing the unstructured corpus of conversational contents comprising ambiguous topics.

A Time-User Sentiment / Topic Latent Dirichlet Allocation (TUS-LDA) was proposed which aggregates posts as a pseudo-document at the same time to lessen the question of context sparsity [19]. But the application of the model to short articles on social media directly affects the sense of sparsity, despite the brief and informal characteristic of messages.

Topics from short texts were detected using Semantics Assisted Non-negative Matrix Factorization (SeaNMF) approach [20]. The approach effectively incorporated the semantic correlations of the word-context into the model, where the semantic relationships between the words and their contexts were learned from the corpus in terms of skip-grams. The SeaNMF model was solved by means of a descent algorithm for block coordinates. Also Non-negative Matrix Factorization (NMF) was used for detection of relationships [21] between the phenotypes which are infectious and the various types of related genes. CVD and hyperlipidemia enriched topics had positive correlations with rs10455872 ($P < 0.001$), replicating a previous finding. Whereas there was a detection of a negative correlation between LPA and a topic enriched for lung cancer ($P < 0.001$) that was not previously identified through phenome-wide scanning. The NMF solution yields a normal representation for the data dependent on sections. A major drawback when applying NMF for data representation is that it fails to consider the geometric structure in the data. If there are several attributes and the attributes are undefined or have poor predictability, NMF is useful else models like LDA and LSA are considered.

The topic models as discussed in the literature review has few pitfalls. They cannot pick up polysemy as they appear to represent one word as one vector and fails to identify the combined word phrases in a text corpus, they are not context sensitive, fails to capture the correlations present in the corpus, doesn't capture the sentiments associated with the topics and the models converges slowly. To overcome these drawbacks, the proposed model is used.

3. Proposed work

The main objective of this research work is to detect hidden, relevant topics and aspects on the reviews given by the users for the hotels they stayed. The hotel review dataset was extracted from booking.com. This website was considered for the study and analysis of customer's reviews on various hotels across the globe.

3.1 Corpus creation

A customized scraper was implemented using the libraries provided by the Python tool. This scraper takes an URL (Uniform Resource Locator) as an input to a particular hotel on booking.com website. The reviews were crawled and were saved in a CSV (Comma Separated Values) file. With these 115,000 customer reviews were extracted, comprising 354 luxury hotels across the globe. The processed CSV file consisted of the following details: Hotel Name, Hotel Location, Review Date, Rating, Reviewer Nationality, Hotel Score, Negative Reviews and Positive Reviews. The architecture of the proposed method is as shown in Fig. 2.

3.2 Data pre-processing

The preliminary data processing used in the research work is described in this section. The corpus created for the hotel reviews were pre-processed for cleaning the dataset for further analysis. If processed as it is was leading to inappropriate detection of topics. The suggested approach used Natural Language pre-processing tools. For creating a standard text dataset the following processes were used.

- The textual format of both positive and negative reviews collected was converted to lowercase for reducing the volume of the text dataset.
- The punctuations were removed to avoid the various forms of a same word followed by the removal of whitespaces.
- The dataset comprising stop words with little insight into the semantical content of the document were removed. E.g. "is," "the," "and", "a," "an," etc.

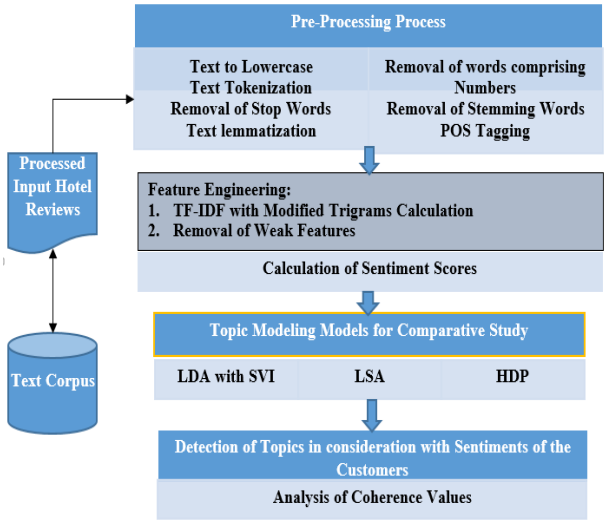


Figure. 2 Proposed architecture for topic modeling

- The stemming words with similar semantic features but different forms had been reduced to a common root name. Examples include for the word think: thinking, thought, thinker likewise.
- Text tokenization was done to split the reviews into tokens. Text comprising of numbers too were removed.
- Part-of-Speech (POS) tagging was done to assign different parts of speech to each word present in the processed corpus.

3.3 Feature engineering

Feature Engineering is the domain information method used to extract features from raw data using technologies in data mining. These features can be used to improve algorithms for machine learning. Its main objective is to prepare the correct input data set, in accordance with the required machine learning algorithm and to increase the performance of machine learning models. In this research work, a two stepped feature engineering process was used to convert both the positive and negative reviews of the customers on the hotel reviews to a vector space model.

In the first step, a TF-IDF with trigrams model was used. Generally the trigrams are calculated as shown in Eq. (1):

$$p\,(x_{1,}x_{2,}x_3) = p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_1\,x_2) \qquad (1)$$

where p gives the probability of x in the document and $x_{1,}x_{2,}x_3$ is a trigram sequence.

$$p(x_3 \mid x_1x_2) = \frac{count\ (x_1x_2x_3)}{count\ (x_1x_2)} \qquad (2)$$

In Eq. (1), for the target word x3, the probabilities of x_1 and x_2 *are* considered as shown in Eq. (2). But users post their tweets using natural language and doesn't follow appropriate grammar in framing the statements. Therefore in order to retain the context and meaning of the trigram sequence, the probabilities of next two words are also considered and calculated as shown in Eq. (3).

$$p(x_1,x_2,x_3)=p(x_1)p(x_2|x_1)p(x_3|x_1x_2)p(x_4|x_3)p(x_5|x_3x_4) \quad (3)$$

This modified model was used for Topic Modeling, which included analyzing and using the relationships between the individual word content-bearing and the topic categories to decide which category a document containing these words belongs to. The Eq. (3) was further processed using TF-IDF

vectorizer. TF, IDF and TF-IDF were calculated using the following Eqs. (4), (5) and (6) respectively.

$$tf\,(x, t) = f_t\,(x) \tag{4}$$

where x is the word, t is the text document and $f_t\,(x)$ is frequency of occurrence of x in the text document t.

$$idf\,(x, T) = log\,\frac{1+|T|}{1+df\,(t,x)} \tag{5}$$

where T is the set of all text documents present in the corpus and $df\,(t, x)$ represents the total number of text documents word x is present.

$$tfidf\,(x, t, T) = tf\,(x, t)\ \text{x}\ idf\,(x, T) \tag{6}$$

Feature vector generated using trigrams indicates how many times it appeared in a specific text document and the TF-IDF values were then replaced by trigrams, rather than mere counts. The first step with TF-IDF using trigrams helped in reducing the high dimensional vector space thereby enhancing the computational capability of the proposed method. Fig. 3 shows the sample trigrams generated for the positive reviews.

Words with higher frequencies are more likely to appear than low frequency words in the textual dataset. The low frequency words are essentially poor characteristics of the corpus, which makes it a good practice for its removal. In the second step of feature engineering process all weak features were removed from the pre-processed corpus.

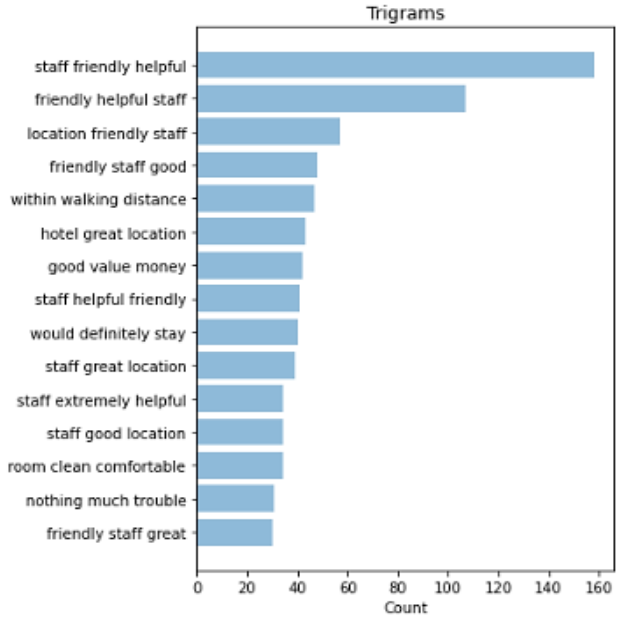


Figure. 3 Sample trigrams of positive reviews generated during experimentation process

Table 1. Python code snippet to remove weak features

```
define remove_high_freq_word(text):
#Output: Optimal words without high frequency words
# Compute median
for value, occurences in values_sorted.items():
   index - = occurences
   if index < 0 and x is True:
      median_manual = value
      break
   elif index == 0 and even is True:
      median_manual = value/2
      x = False
Do all the pre-processing steps()
# remove words with only one letter
text = [t for t in text if length(t) > 1]
# join all
text = " ".join(text)
return(text)
```

The steps for removing the weak features are as follows:

- Create a dictionary of word frequency.
- Arrange every word x according to its frequency.
- If the word frequency count is below the calculated median value, then this word x is considered as a weak feature.
- Remove the identified weak features from the pre-processed corpus.

The steps for Removing Weak Features is as shown Table 1.

It was observed that removal of weak features from the pre-processed corpus created an efficient VSM thereby reducing its dimensionality and improvement of model's computational capability.

3.4 Topic modeling

The hotel reviews dataset after going through initial pre-processing and two-stepped feature engineering process, the created VSM was given as an input to three topic detection models for comparative studies. They were Latent Semantic Analysis (LSA), Hierarchical Dirichlet Process (HDP) and the proposed method Batchwise LDA (Latent Dirichlet Allocation) with Stochastic Variational Inference (SVI) model. Also a sentiment analysis was done on the pre-processed text corpus for positive and negative reviews. This sentiment analysis helped user's reviews in appropriately and accurately detecting relevant topics as user's sentiments were directly associated with their reviews written for a particular hotel based on their experience.

The sentiment score was calculated using VADER (**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner) which is a natural language processing tool. For the classification of text VADER is a rule-based, unsupervised system. VADER is used mainly for textual data from online or social media. VADER tool helps in using a method named SentimentIntensityAnalyzer feature for finding the text's sentiment. This method upon giving a text input returns a four parameter tuple: 'compound', 'neg', 'neu' and 'pos'. The 'compound' value lies between [-1 and 1]. Percentage value of 'neg', 'neu' and 'pos' defines the text's negative, neutral or positive polarity. Their value lies between the range [0, 1].

The Compound Score is analysed as follows:
If its value is >= 0.05 then its polarity is considered as positive.
If its value is > - 0.05 and < 0.05 then its polarity is considered as neutral
If its value is <= - 0.05 then it is considered as negative polarity.

3.4.1. Latent semantic analysis model

LSA is an unsupervised statistical approach and is one of the commonly used technique for Topic Modeling. The core idea is to take a matrix of documents and terms and then decompose it into a separate document-topic matrix and a topic-term matrix. In the first step the document-term matrix is generated as follows: Given *m* documents and *n* words in the vocabulary, an *m x n* matrix *A* is constructed using TF-IDF scores in which each row represents a unique word and each column represents a document. In the second step, matrix *A* is factorized to three other matrices *X, Y* and *Z* using Singular Value Decomposition (SVD) process for further reducing the dimensionality. It is as shown in Eq. (7). *X* and *Z* are orthonormal matrices and *Y* is a singular matrix whose determinant cannot be found.

$$A = X\,Y\,Z^T \tag{7}$$

Where X is the document-term matrix which gives the vector representation of a document. The length of this vector is equal to the total number of preferred topics for modeling process. Z gives the vector representation for the terms and is called term-topic matrix. SVD therefore provides us with vectors in our data for every document and term. Every vector is k in length. Similar terms and related documents are then identified using the cosine similarity test by using certain vectors.

3.4.2. Hierarchical dirichlet process model

A Dirichlet (DP) process can be called a distribution of probability within the field of probability measurements as a stochastic process. The method is named precisely because the DP results in partial, finite-dimensional distributions of Dirichlet, which is close to Gaussian, which has finite-dimensional, conditional distributions distributed in Gaussian [22]. A DP is defined by a base measurement and a parameter in terms of concentration. Every draw from the DP is a measure itself. As it is certain that a previously drawn attribute can be drawn again, the drawn values are discrete with probability 1. This allows them to cluster into DP mixtures. The HDP is an alternative of DP hierarchical. The hierarchy offers an elegant framework for sharing parameters. This process determines a set of X_i likelihood measurements for D pre-specified data groups and a global X_0 probability measure. It is shown in the Eq. (8).

$$X_0 |\, \alpha, H \sim DP\,(\alpha, H\,) \tag{8}$$

Where H gives the measure for Base Probability and α is the parameter for concentration. The HDP measures for simple probability permit countless multinomial drawings and therefore an infinite number of topics. This allows the number of topics to increase or decrease by data. In this study, the on-line HDP method provides the versatility of modeling the HDP to the speed of on-line variational Bayes model. The idea behind the usage of Bayes online variation is to maximize the objective function of variation with stochastic optimization model.

3.4.3. Batchwise latent dirichlet allocation model with stochastic variational inference

Batchwise LDA with SVI is the proposed method for this research work. LDA is an unsupervised machine learning approach and it is pLSA (Probabilistic LSA) with Bayesian version. pLSA uses probabilistic approach and not SVD. A probabilistic model with latent topics can be found which can produce the data that observes in the document-term matrix. LDA takes text corpus as input and produces topics as the output. Also it gives the percentage of each text document and its association with each found topic. For Topic Modeling, LDA doesn't consider the order of words in the given corpus. The LDA process starts with an assumption that there are *n* topics across text corpus and thereby documents are produced based on those assumed *n* topics. This in turn generates word based on the distribution of the probability. The value of n

should not be a high value as it produces granular topics.

Given a document *t* in the textual dataset *T*, then *t* is calculated as follows:

- The total number of words present in the document *t* as represented by X_t is drawn from the Poisson's Distribution model.
- The number of topics to be found present in the document *t* as represented by θ_t is drawn from topic-document distribution also known as Dirichlet Distribution.
- Allocate each word w_i to a topic z_i, where i is ranging from 1 to X_t, in such a way that it is consistent with the previous step.
- The word w_i is drawn from the topic-word distribution.

A corpus can be divided into batches of fixed sizes to obtain most relevant topic terms. LDA procedure can produce different results on these batches many times, but the best topics are generated using the intersection results of all the batches. This process is called as the Batchwise LDA.

Stochastic Variational Inference is a scalable approach used for the approximation of posterior distribution of the probabilistic model of LDA. It is useful in analyzing large text corpus which other topic models fails [23]. Table 2 gives the LDA algorithm using SVI.

3.5 Topic coherence measure

It is a widely used metric to evaluate topic models and gives a realistic measure to identify the total number of topics. It uses the latent variable models. Each generated topic has a list of words. This measure finds the average of pairwise word similarity scores of the words associated with the topic. The topic model with high Coherence Measure value is considered a good topic model. When the initial number of topics (*k*) chosen was 33 in this paper, it resulted in good coherence scores but the number of words appeared per topic were high and repeating. Therefore during the experimentation process, *k* was limited to and varied between 1 and 8.

Table 2. LDA with SVI

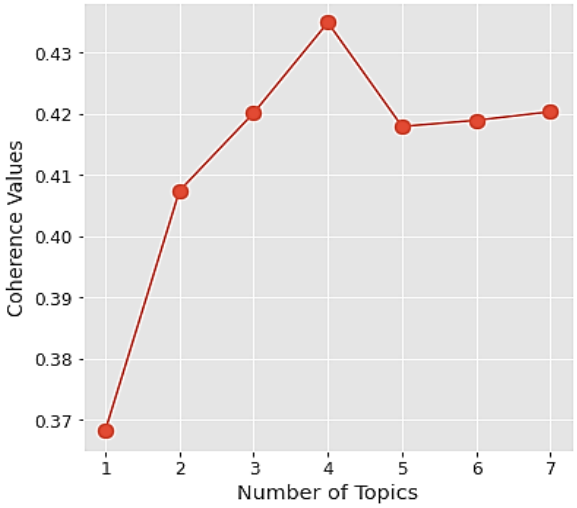
1. Initialize Global per Topic Dirichlet Parameters (initialize topics at 0^{th} iteration)
2. Set the Step-Size schedule (Maximizes the Optimization Function comprising of Step 1)
3. Repeat:
    a. Sample a document *t* uniformly from the text dataset.
    b. Initialize a K-vector Dirichlet Parameter to 1 (per-document parameters)
    c. Repeat:
        i. For each document *t* compute optimal local variational parameters θ by updating step 3b until the θ converges by maximizing the Evidence Lower Bound (ELBO) (until word is assigned to the correct topic).
    d. Find and update intermediate topics using optimal local variational parameters. (Update variational Dirichlet for each topic).
    e. Set the topics at the next iteration to be a weighted combination of the intermediate topics and current topics.
4. Until forever

Fig. 4 (a), (b), and (c) are the results of Topic Modeling using LSA, HDP and Batchwise LDA with SVI without using Feature Engineering. Fig. 5 (a), (b), and (c) are the results of Topic Modeling using LSA, HDP and Batchwise LDA with SVI using Feature Engineering with TF-IDF trigram model. Fig. 6 (a), (b), and (c) are the results of Topic Modeling using LSA, HDP and Batchwise LDA with SVI using the proposed method (two-stepped feature engineering process).

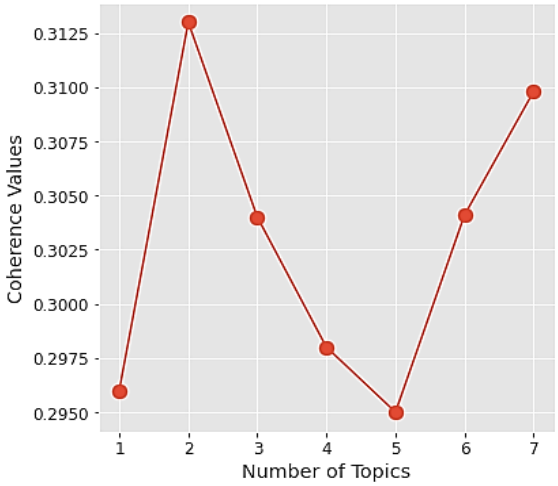
4. Experimental results and discussions

This section provides performance evaluation of the proposed research methodology. The dataset considered for the study was the reviews on hotels across the globe. This was done with the help of a customized scraper using Python tool. The reviews in the text format was unstructured and had lexical, syntax and semantic errors. If processed as it is was leading to inappropriate detection of topics. Therefore it was first pre-processed as discussed in 3.2 and the corpus collection is mentioned in the section 3.1. Fig. 7 gives the statistics of the reviews collected.

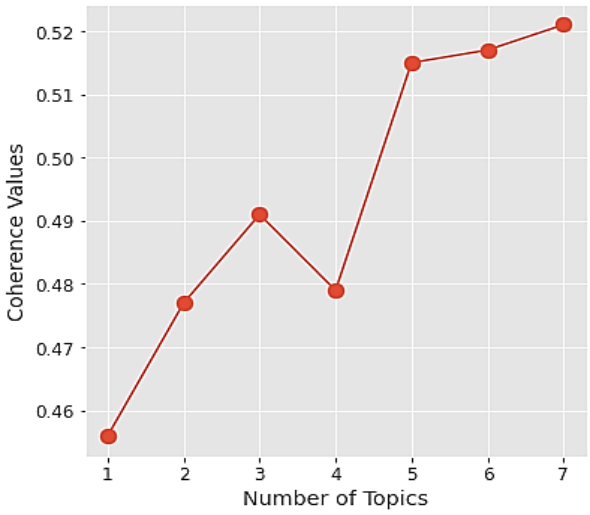
Natural Language Toolkit 3.0 was used for building python programs as the work involved was with human generated textual dataset. After pre-processing the dataset, the input was given to three topic detection models namely LSA, HDP and Batchwise LDA with SVI. The results are as shown in Fig. 4 (a), (b) and (c). Later the pre-processed dataset was converted into a vector space model using TF-IDF with trigrams model. It was observed that the coherence values for the topic models considered was improved and this is shown in Fig. 5 (a), (b) and (c) for LSA, HDP and Batchwise LDA with SVI models respectively. Further the experimental results improved with good coherence values when the



(a)

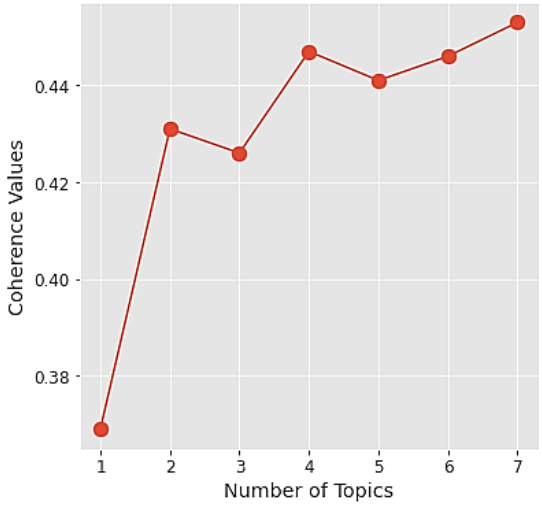


(b)

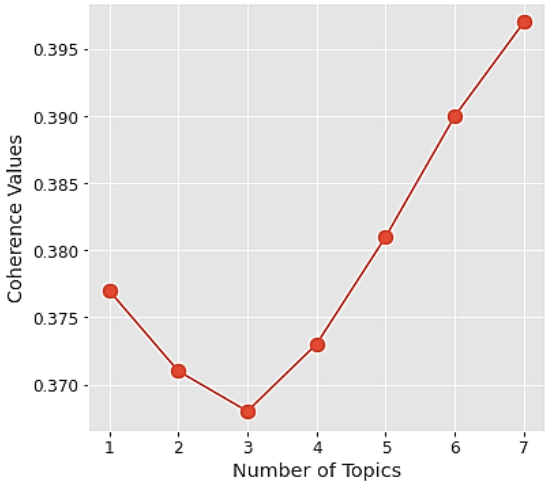


(c)

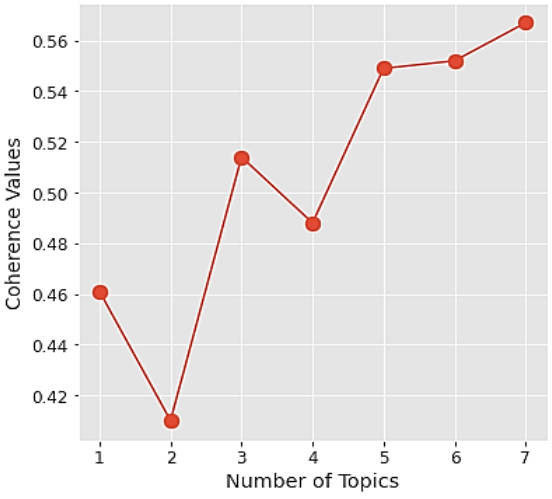
Figure. 4 Results of topic modeling: (a) LSA without feature engineering, (b) HDP without feature engineering, and (c) Batchwise LDA with SVI without feature engineering



(a)

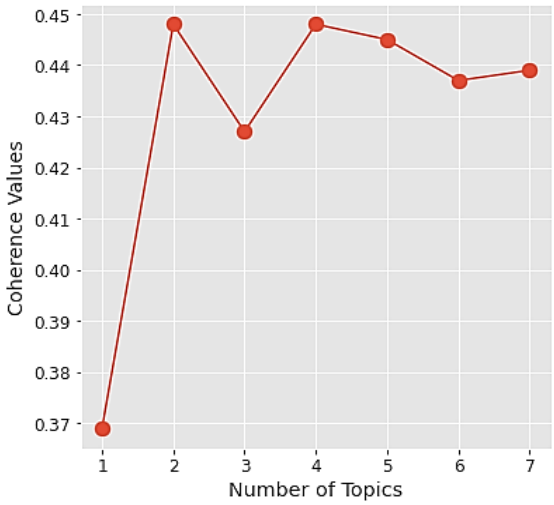


(b)

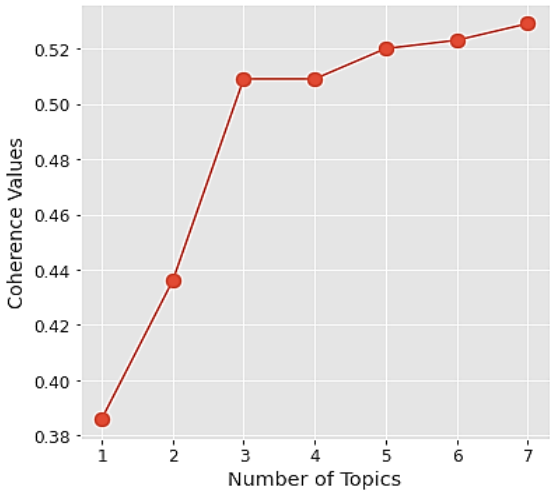


(c)

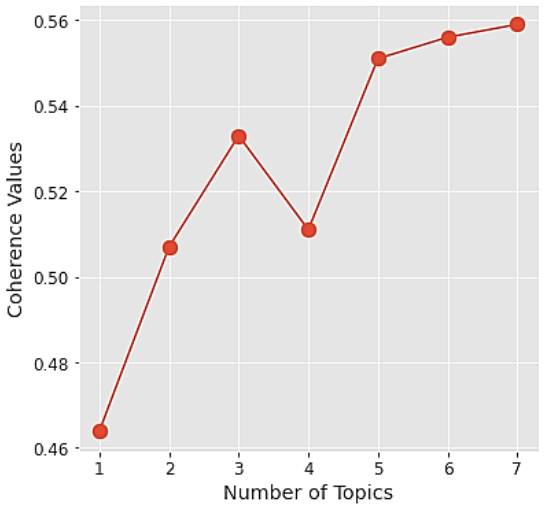
Figure. 5 Results of topic modeling: (a) LSA using feature engineering with TF-IDF trigram model, (b) HDP using feature engineering with TF-IDF trigram model, and (c) batchwise LDA with SVI using feature engineering with TF-IDF trigram



(a)



(b)



(c)

Figure. 6 (a) LSA using two-stepped feature engineering process, (b) HDP using two-stepped feature engineering process, and (c) batchwise LDA with SVI using two-stepped feature engineering process

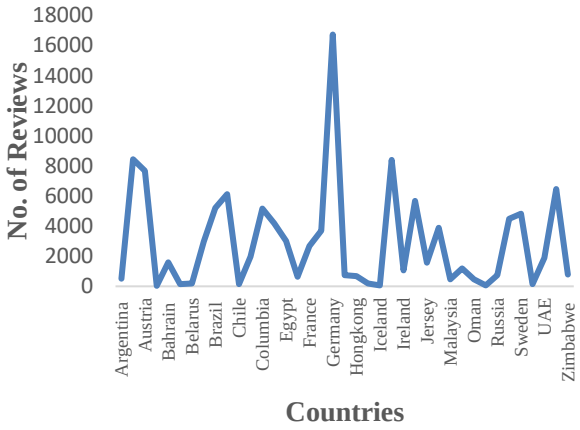


Figure. 7 Hotel reviews statistics

dataset was further processed by removal of weak features as they weren't contributing much in detection of hidden relevant topics with respect to the reviews on hotels given by the customers and this can be seen in Fig. 6 (a), (b) and (c) for the models considered for the study. Table 3 gives the comparative results of all the three topic models where *K* being the total number of topics being chosen.

Since all three models used for the comparative studies were created with a variety of different topics, a mixture of general and more specific topics were created because there was an increasing number of topics. The benefit was that these engineering models had more general topics which led to the topic model being more representative of the corpus, as shown by the numerical coherence value results as shown in the Table 3. It was also observed that the proposed approach Batchwise LDA with SVI with two-stepped feature engineering method worked better in Topic Modeling with optimum number of topics than the existing methods like LSA and HDP. From the Table 3, it was observed that LDA with SVI model with two-stepped feature engineering had attained the high average coherence value of 0.513 when compared to LSA with an average coherence value of 0.4273 and HDP with an average coherence value of 0.472. However their coherence score was less without feature engineering process. The graphical representation of the proposed method is shown in Fig. 8.

Table 3. Comparative coherence values for batchwise LDA-SVI, LSA and HDP topic models

K (No. of Topics)	Without Feature Engineering			One - Step Feature Engineering Process			Two - Stepped Feature Engineering Process		
	LDA - SVI	LSA	HDP	LDA - SVI	LSA	HDP	LDA - SVI	LSA	HDP
1	0.456	0.368	0.296	0.461	0.369	0.377	0.464	0.369	0.386
2	0.477	0.407	0.313	0.410	0.431	0.371	0.507	0.448	0.436
3	0.491	0.420	0.304	0.514	0.426	0.368	0.533	0.427	0.509
4	0.479	0.435	0.298	0.488	0.447	0.373	0.511	0.448	0.509
5	0.515	0.418	0.295	0.549	0.441	0.381	0.551	0.445	0.520
Average	0.483	0.409	0.301	0.484	0.423	0.374	0.513	0.427	0.472

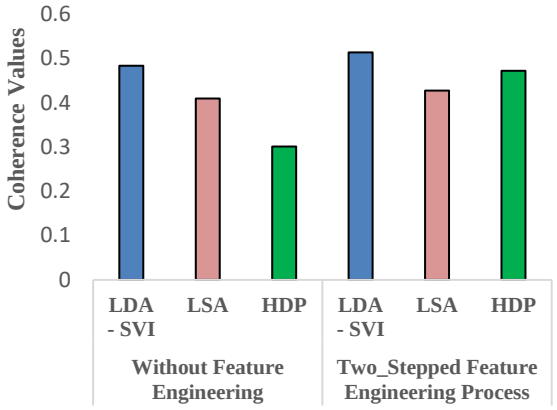


Figure. 8 Topic models with and without feature engineering process

5. Conclusion and future work

The analysis and description of the hotel user reviews listed on the Booking.com website was a major focus of this report. The overview description provided users with a better understanding of what they want when reserving their hotel of choice. The hotel reviews analysis provided hidden topics, relevant and useful words. This analysis is important for further usage in discovering more valuable information. The research also examined the sentiments of hotel users on their aspects, which provided a clearer understanding of which aspects of the hotel being studied are clearer than the others according to the feedback from the user and which of these aspects should be considered.

In this proposed work, a two-stepped feature engineering was proposed. In the first step, modified calculation of trigrams was used which helped in retaining the context and meaning of the sequence considered and this information was passed on to TF-IDF model thereby creating an efficient Vector Space Model. In the second step, additional removal of the weak features from the corpus based on Median Value added enhancement in detecting hidden topics with simultaneously increasing the model's coherence values. The general variance model added more weightage for the outliers as they were far from the calculated mean, so Median Value was considered for weaker features removal. Batchwise LDA with SVI when compared with LSA and HDP yielded better experimental results. This was due to model's probabilistic capability of identifying interpretable topics. The Batchwise LDA with SVI improved its performance by 3% in terms of its coherence values by using two-stepped feature engineering process and by 9% and 4% increase when compared with LSA and HDP models respectively.

Batchwise LDA assumes that the input dataset includes documents relevant to a certain number of topics and that each document derives from a combination of probabilistic samples: first, the distribution of possible topics for the dataset considered and second, the list of potential words in a chosen topic. This generative presumption confers one of Batchwise LDA's main advantages over other approaches to modeling. LDA with SVI easily distinguished between global variables and local. The global values are topics and local variables are document-level topic indices and breaking proportions. Finally Batch Wise method helped in the identification of the important topics relevant in decision making of hotel booking by customers based on the hotel reviews.

Many reviews will literally defame the hotel and undermine the hotel rating by leaving the negative aspects of their comments. Therefore, using data such as "good feedback" or the number of positive votes, the legitimacy of this analysis can be further improved and increase its weightage on the final

outcome. A significant way for this research is to ask the customer about the qualities they like in hotels, e.g. Economical, for family, good food, air condition, Wi-Fi, pools and so on. These information would help the researchers to work upon a more customized and personalized Hotel Recommender System for customers to choose good hotels based on their preferences. Many customers express their opinions using images or the photos they had captured during their stay. With this an image processing can be thought of integrating with Topic Modeling. Also integrating dynamic time series data for converting the proposed model to a powerful predictive model can be this paper's future scope.

Conflicts of Interest

The authors declare no conflicts of interest.

Author Contributions

The contributions by the authors for this research article are as follows: "Conceptualization, Methodology, Formal Analysis, Corpus Creation, Visualization and Writing - Original Draft Preparation by Sujatha Arun Kokatnoor; Result Validation, Data Curation, Formal Analysis Writing—Reviewing and Editing and Supervision by Balachandran Krishnan".

Acknowledgments

Authors wishes to acknowledge the technical and infrastructural help rendered by the faculty members of CSE department of CHRIST (Deemed to be University), Bangalore, India.

References

[1] M. Zhang, S. Ding, and Y. Bian, "The Online Reviews' Effects on Internet Consumer Behavior: An Exploratory Study", *Journal of Electronic Commerce in Organizations*, Vol. 15, No. 4, pp. 83-94, 2017.

[2] M. Reisenbichler and T. Reutterer, "Topic Modeling in Marketing: Recent Advances and Research Opportunities", *Springer Journal of Business Economics*, Vol. 89, No. 3, pp. 327-356, 2019.

[3] W. Hopken, M. Fuchs, and M. Lexhagen, "Business Intelligence for Cross-Process Knowledge Extraction at Tourism Destination", *Journal of Information Technology and Tourism*, Vol. 15, No. 2, pp. 101-130, 2015.

[4] D. Valdez, A. C. Pickett, and P. Goodson, "Topic Modeling: Latent Semantic Analysis for the Social Sciences", *Social Science Quarterly*, Vol. 99, No. 5, pp. 1665-1679, 2018.

[5] L. Shi, G. Cheng, S. Xie, and G. Xie, "A Word Embedding Topic Model for Topic Detection and Summary in Social Networks", *Measurement and Control*, Vol. 15, No. 9, pp.1289-1298, 2019.

[6] S. Likhitha, B. S. Harish, and H. M. Keerthi Kumar, "A Detailed Survey on Topic Modeling for Document and Short Text Data", *International Journal of Computer Applications*, Vol. 178, No. 39, pp. 1-9, 2019.

[7] M. Mukherjee and E. Poovammal, "Improved Topic Modeling with Parallel-Supervised LDA", *International Journal of Recent Technology and Engineering*, Vol. 8, No. 3, pp. 5692-5696, 2019.

[8] M. G. Lozano, J. Schreiber and J. Brynielsson, "Tracking Geographical Locations using a Geo-Aware Topic Model for Analyzing Social Media Data", *Decision Support Systems*, Vol. 99, No. 1, pp. 18-29, 2017.

[9] D. Yu, D. Xu, D. Wang, and Z. Ni, "Hierarchical Topic Modeling of Twitter Data for Online Analytical Processing", *IEEE Access*, Vol. 7, No. 1, pp. 12373-12385, 2019.

[10] B. Zhang, D. Xu, H. Zang, and M. Li, "STCS Lexicon: Spectral-Clustering-Based Topic-Specific Chinese Sentiment Lexicon Construction for Social Networks", *IEEE Transactions on Computational Social Systems*, Vol. 6, No. 6, pp. 1180-1189, 2019.

[11] I. Sutherland, Y. Sim, S. K. Lee, J. Byun, and K. Kiatkawsin, "Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation", *Sustainability*, Vol. 12, No. 5, pp. 1-15, 2020.

[12] Z. Hai, G. Cong, K. Chang, P. Cheng, and C. Miao, "Analyzing Sentiments in One Go: A Supervised Joint Topic Modeling Approach", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29, No. 6, pp. 1172-1185, 2017.

[13] D. Deng, L. Jing, J. Yu, S. Sun, and M. K. Ng, "Sentiment Lexicon Construction with Hierarchical Supervision Topic Model", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 4, pp. 704-718, 2019.

[14] X. Shao, G. Tang, and B. K. Bao, "Personalized Travel Recommendation Based on Sentiment-Aware Multimodal Topic Model", *IEEE Access*, Vol. 7, No. 1, pp. 113043-113052, 2019.

[15] A. O. Steinskog, J. F. Therkelsen, and B. Gamback, "Twitter Topic Modeling by Tweet Aggregation", In: *Proc. of the 21st Nordic Conf.*

of Computational Linguistics, Gothenburg, Sweden, pp. 77-86, 2017.

[16] Y. Li, W. Z. Song, and B. Yong, "Stochastic Variational Inference-Based Parallel and Online Supervised Topic Model for Large-Scale Text Processing", *Journal of Computer Science and Technology,* Vol. 33, No. 5, pp. 1007-1022, 2018.

[17] R. Ibrahim, A. Elbagoury, M. S. Kamel, and F. Karray, "Tools and Approaches for Topic Detection from Twitter Streams: Survey", *Knowledge and Information Systems*, Vol. 54, No. 3, pp. 511-539, 2018.

[18] J. F. Yeh, Y. S. Tan, and C. H. Lee, "Topic detection and Tracking for Conversational Content by using Conceptual Dynamic Latent Dirichlet Allocation", *Neurocomputing*, Vol. 216, No. 3, pp. 310-318, 2016.

[19] K. Xu, G. Qi, J. Huang, and T. Wu, "A Joint Model for Sentiment-Aware Topic Detection on Social Media", In: *Proc. of the 22nd European Conf. on Artificial Intelligence*, Hague, Netherlands, pp.338-346, 2016.

[20] T. Shi, K. Kang, J. Choo, and C. K. Reddy, "Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word Context Correlations", In: *Proc. of the 2018 World Wide Web Conf.*, Lyon, France, pp. 1105-1114, 2018.

[21] J. Zhao, Q. Feng, P. Wu, J. L. Warner, J.C. Denny, and W. Wei, "Using Topic Modeling via Non-negative Matrix Factorization to Identify Relationships between Genetic Variants and Disease Phenotypes: A Case Study of Lipoprotein(a) (LPA)", *PloS One*, Vol. 14, No. 2, pp. 1-15, 2019.

[22] A. M. Dai and A. J. Storkey, "The Supervised Hierarchical Dirichlet Process", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 2, pp. 243-255, 2015.

[23] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic Variational Inference", *Journal of Machine Learning Research*, Vol. 14, No. 1, pp. 1303-1347, 2013.