Topic Extraction using Mixture Model from Online Reviews and Articles

Sana Zawar, Salabat Khan
COMSATS University
Islambad Attock Campus
Pakistan
{sanazawar8@gmail.com} {salabat.khan@ciit-attock.edu.pk}

ABSTRACT: *Topic recognition and tracking is one of the problems in the field of Natural Language Processing (NLP). Many methods and techniques have been introduced currently for information retrieval and extracting topic from user generated reviews. Finding out topics from the group of documents is very valuable for several real-world applications.. LDA (Latent Dirichlet Allocation) and PLSA (Probabilistic Latent semantic Analysis) are used for this purpose. This research aims to extract topic from online reviews and articles using mixture model. Mixture Model is probability based model. One thousand articles from four hundred refereed journals are collected and analyzed textually through topic mining techniques. These articles are collected from six major databases namely IEEE, Science Direct, Wiley, SAGE, Cambridge and Springer. The experimental results of proposed method clearly shows that proposed model yields very promising results.*

Keywords*:* Mixture Model, Clustering, Natural Language Processing

Received: 11 October 2019, Revised 18 January 2020, Accepted 2 March 2020

DOI: 10.6025/jcl/2020/11/2/51-59

1. Introduction

The web is flooded with reviews generated by users and articles. The reviews generated by users are considered the significant source of how someone develop an opinion while making decisions. As number of reviews is growing it has become gradually significant to provide users some tools to make many opinions about a topic. For this purpose our center of attention is to mining topics from online reviews generated by users and articles.

Although timely information retrieval is becoming progressively more important now a days in the economy based on the information provided on web. To gain access is no longer a difficulty due to the extensive accessibility of broadband in businesses and homes. Inconsistently, high speed connectivity and sudden increase in quantity of digitized text available online created problem called information overload. Noticeably, the capability of humans to incorporate such huge amounts of knowledge is limited. Topic recognition has emerged as a capable research field that can deal with this problem.

Now a days customers are frankly enable to share product reviews and their experiences over the e-commerce websites and on social networks. This fact makes it possible to have the vast amount of information produced by users in order to assist the future customers. Topic extraction and sentiment analysis have been the dynamic research areas during the recent years. These abilities make it achievable to automatically recognize the emotions of the users about the specific topics they used to discuss. In particular, the processing of data related to opinion of customers as observed in reviews and articles has been key interest area. Existing work regarding this study has paying attention on providing helpful information to the customers who desire to make sure the sentiment communicated by the others users with respect to the business [1]. While the difficulty to finding out the sentiments polarity and topic extraction from online reviews and articles has been directed in literature. There is no existing technique to observe how related sentiments and topics can be used to providing gaudiness for businesses. This provided gaudiness should be able for the businesses to improved understanding about the opinions of their clients regarding services and products The reviews generated by users are the significant source of knowledge for clients and are Considered to play a dynamic role in making decision in the number of fields [2]. So far many researchers have described techniques for taking out topical information from reviews generated by users. It is being demonstrated that these features help to get better classification performance which is possible using more predictable techniques of feature extraction. Mixture Model is probability based model. Work with the clustering of data points, so each data point is a distribution of mentioned cluster, where each data point may be carrying in one cluster with a specific probability. However number of cluster will be specified by the users .The proposed method has hierarchical structure which will work in two levels. In the first stage data will be collected from online reviews, articles and newspapers. In the next stage collected data will be processed by Mixture model which works in two levels. In the first level data will be distributed in clusters and the word having high probability will be selected as seed word. In the next level seed word having high probability will be selected as the topic.

The rest of the article organized as follows. Section 2 represents the literature review of the topic modeling techniques. The proposed approach is presented in section 3. Section 4 presents experimental setup and results in detail. Finally conclusion and future work is discussed in section 5.

2. Literature Review

Researchers have proposed different techniques used for extracting topic from online reviews and articles. Proposed hybrid term relations analysis technique for topic detection in online reviews and articles [3]…used correlation, clustering and classification technologies for topic detection [4]. Yang et al developed incremental clustering framework to detect new topics and also used temporal and content features to detect emerging topics [5]. Melies et al Proposed a dual-sparse topic model for topic extraction that allowed an individual to selects a few focused topics to in a document and a topic to select focused terms [6, 7] proposed an analytical method and identify two things first one cluster associated terms phrases comprise meaningful topics and their association and second is identify changing in topical emphases [8] .

Chen et al introduced Existing graph-based ranking technique for key phrase mining compute a single significance score for every word through a single arbitrary walk [9]. Motivated by the fact that both documents and words can be presented by a mixture of topics, the technique is proposed to decompose conventional arbitrary walk into many arbitrary walks particular to a variety of topics. Zang et al thus assemble a TPR (Topical Page Rank) on word to calculate word importance for different topics [10]. After that, the topic allocation of document, authors further calculate the ranking scores of words and extract the top ranked word as key phrases. The results show that topical page rank outperforms then the existing key phrase extraction techniques on various datasets under a variety of evaluation metrics.

CTM is a probabilistic model used to combine the statistical model along with hierarchy of semantic concepts defined by human. In this introduced topic model, authors give the idea to use concepts to topics of topic model .This topic model produce effective set of topics and word distribution of concepts for the mentioned documents.

Opinion mining and Sentiment analysis aims to apply automated tools for the detection of subjective information e.g. opinions, feelings and attitudes expressed in the text. In this method author suggest a new probabilistic modeling agenda based on LDA (Latent Dirichlet Allocation) which is called JST (Joint Sentiment Model). JST is used to detect topic and sentiment at the same time from text [11]. Contrasting from remaining machine learning techniques for the classification of sentiment which is often required to labeled corpora used for training of classifier, the proposed JST method is unsupervised. This method has been tested on the dataset of movie review for the classification of prior information review sentiment polarity have also been explored to improve the accuracy of sentiment classification. The experiments have exposed capable results achieved by JST.

Many researchers work on Finite Mixture model ..Mixture Model is probability based model. Works with the clustering of data points, In Mixture Model each data point is considered a distribution for mentioned cluster, and each data point may be carrying in one cluster with a specific probability [12]. LDA (Latent Dirichlet Allocation) and PLASA (Probabilistic Latent semantic Analysis) are used for topic extraction. CorrTM is also introduced for topic extraction Correlated method has capability to model difficult structure of fundamental topics and provides covariance matrix used form atopic graph The mentioned models have some limitations for example, the current models do not capture correlations between labels, ignore the word ordering, do not identify the polarity of text at diversity of levels [13]. The following table describes the summary of topic modeling techniques.

Sr. No	Year	Topic Models	Methodology Applied	Specification
1	2015	LSA	Singular Decomposition Value	LSA can extract topic if there is any synonym word.
2	2016	AHMM	EM	Identify topical dependency among the words
3	2016	HLDA	Gibbs Sampling	Number of topics are not fix
4	2016	Core TM	Variational EM	Create relation between topics by Applying logistic distribution
5	2016	HLSI	Hierarchy graph	It maintains the inherent structure of categorization
6	2017	SCNTM	Gibbs Sampling (Collapsed)	Describes bibliographic analysis for topics authors and documents
7	2018	LDA	Variational EM	It does not represent the relationship between the topics From a single topic, it can generate every word
8	2018	PLSI	EM (Tempered)	even various word can be generated from the document.
9	2018	DTM	Gaussian Model	Evolution of the topic according of time

Table 1. Summary of Topic Extraction Techniques

3. Proposed Approach

This research aims to study multiple techniques for topic modeling. The main focus of research is on extracting topics from online reviews and articles using Mixture Model. Mixture Model will be used on online reviews and articles in order to extract meaningful topic more accurately and efficiently. The proposed method will be compared with multiple topic modeling techniques e.g LDA (Latent Dirichlet Allocation) and PLASA (Probabilistic Latent semantic Analysis) the main concern is to gain the maximum accuracy of assigning topic.

Mixture Model is a probability based model. Works with the clustering of data points, so each data point is a distribution of mentioned cluster, where each data point may be carrying in one cluster with a specific probability. However number of cluster will be specified by the users. The proposed method has hierarchical structure which will work in two levels. The levels of mixture model are discussed below.

In the first level probability of words is calculated from the clusters of words. The following equation 1 will be used to calculate probability of words from the clusters. Formally we use random variables $x1, x2, x3..., x\,(n-1), xn$, using mixture model which have K components. Assume that each k^{th} component of Mixture Model be a distribution having parameters $(\theta\, k)$ in the form of $F(x\,|\theta\, k)$, and let $\pi\, k$ $(\pi\, k \geq 0$ and $\_k\ \pi\, k = 1)$ be the weight of k^{th} component which denotes the probability. Therefore, probability of xi can be written as:

$$p(x_i) = \Sigma_{k=1}^{K} \pi_k f(x_i | \theta_k) \tag{1}$$

The second level is the probability distribution of topic. After generating seed words (the words having higher probability within each cluster). We will select a seed word with higher probability as a topic. The equation 2 will be used to calculate the probability of seed words.

$$p(x_1, x_2, ..., x_n) = \Pi_{i=1}^{n} \Sigma_{k=1}^{K} \pi_k f(x_i | \theta_k) \tag{2}$$

3.1 A Research Methodology Flow Diagram

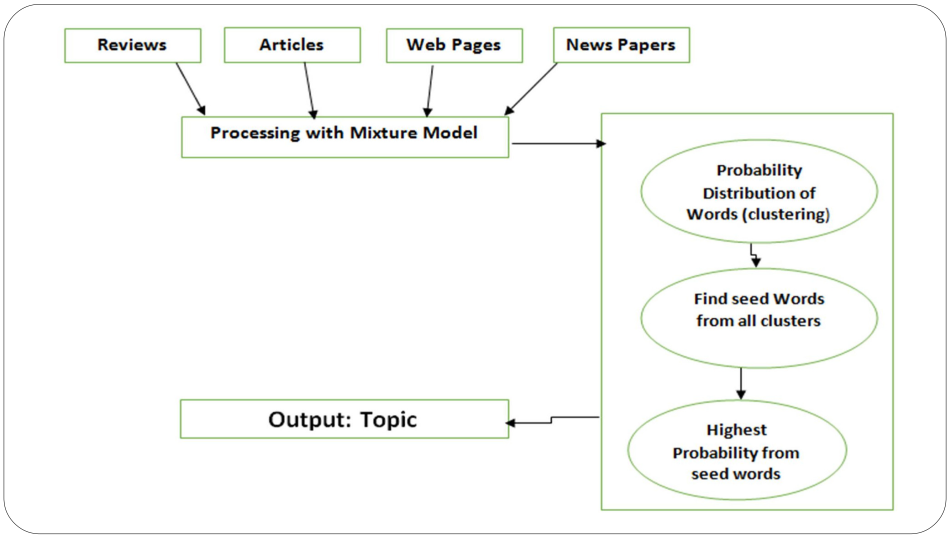

Reviews
Articles
Web Pages
News Papers
Processing with Mixture Model
Probability Distribution of Words (clustering)
Find seed Words from all clusters
Highest Probability from seed words
Output: Topic

Figure 1. Research Methodology

Figure 1 describes the flow of research. In the first stage data will be collected from online reviews, articles and news papers. In the next stage collected data will be processed by Mixture Model which works in two levels. In the first level data will be distributed in clusters and the word having high probability will be selected as seed word. In the next level seed word having high probability will b selected as the topic.

3.2 *k*- Mean Clustering

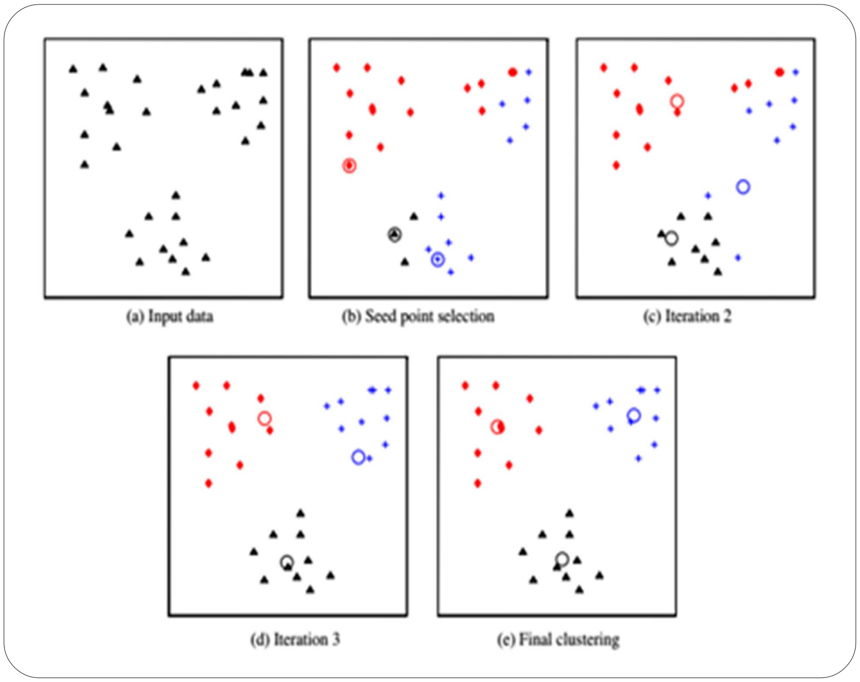
Let *X* ¼ fxig, *I* ¼ 1; ... ; *n* be the arrangement of *n* d-dimensional focuses to be clustered into an arrangement of *K* groups, *C* ¼ *fck*; *k* ¼ 1; ... ; *Kg*. *K*-mean calculation finds a section with the end goal that the squared error between the exact mean of a cluster and the focuses in the collection is limited. Give *lk* a chance to be the mean of group *ck*. The squared error amongst *lk* and the focuses in group *ck* is characterized as:

$$j(c_k) = \Sigma_{xi(c_k)} \,||\, x_j - \mu_k \,||^2 \tag{3}$$

The objective of *K*-mean is to limit the whole of the squared error over all *K* groups:

$$j(c) = \Sigma_{i=1}^{k} \Sigma xi(c_k) \,|\,| x_j - \mu_k | \,|\, 2 \quad (4)$$

Limiting this target work is known to be a *NP*- difficult issue (notwithstanding for *K* = 2). Therefore *K*-means, which is an a avaricious calculation, can just merge to a nearby least, despite the fact that ongoing examination has appeared with a huge likelihood *K*-means could join to the worldwide ideal when groups are all around isolated. *K*-implies begins with an underlying segment with *K* groups and appoint examples to bunches to diminish the squared mistake. Since the squared blunder dependably diminishes with an expansion in the quantity of bunches *K* (with *J*(*C*) = 0 when *K* = *n*), it tends to be limited just for a settled number of groups.


(a) Input data (b) Seed point selection (c) Iteration 2

(d) Iteration 3 (e) Final clustering

Figure 2. *k*-Mean clustering

Figure 2 demonstrates a definition of the *K*- implies calculation on a 2-dimensional dataset with three groups.

3.3 Number of Clusters

Naturally deciding the quantity of groups has been a standout amongst the most difficult issues in information gathering. Most techniques for consequently deciding the quantity of groups cast it into the issue of model determination. More often than not, grouping calculations are kept running with various estimations of *K*; the best estimation of *K* is then picked in view of a predefined measure. A few researchers utilized the base message length (ML) criteria in conjunction with the Gaussian Mixture Model (GMM) to appraise *K*. Their approach begins with an expansive number of groups, and step by step consolidates the bunches if this prompts a decline in the MML basis. A related approach however utilizing the standard of Minimum Description Length (MDL) was utilized in for choosing the quantity of bunches. The other criteria for choosing the quantity of groups are the Bayes Information Criterion (BIC) and Akiake Information Criterion (AIC). Hole measurements is another regularly utilized approach for choosing the quantity of clusters. The key supposition is that while separating information into an ideal number of groups, the subsequent segment is strongest to the irregular irritations. The Dirichlet Process (DP) presents a non-parametric method for the quantity of groups. Usually utilized by probabilistic models to infer a back dispersion for the quantity of clusters, from which all likelihood number of groups can be figured. Disregarding these goal criteria, it is difficult to choose which estimation of *K* reminders more significant groups.

3.4 Data Representation

Information representation is a standout amongst the most critical elements that impact the execution of the clustering calculation. In the event that the depiction (selection of highlights) is reliable, the groups are probably going to be smaller and isolated and even a basic clustering calculation, for example, *K*-means will discover them. formally, there is no generally reliable representation; the decision of depiction must be guided by the various information.

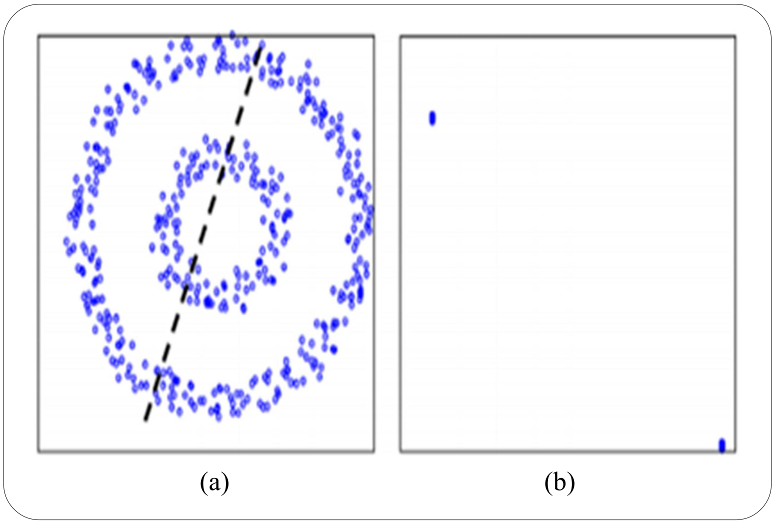


Figure 3. Data Representation

Figure 3(a) demonstrates a dataset where *K*-mean neglects to segment it into the two “regular” bunches. The parcel acquired by *K*- mean is appeared by a dashed line However, when similar information focuses in are required to utilizing the main two eigenvectors of the RBF likeness network processed from the information in Figure 3 (b) they turn out to be all around isolated, making it minor for *K* intends to group the information.

4. Performance Evaluation and Experimental Results

4.1 The Dataset

The dataset of ABS news is used in order to evaluate the performance of proposed system. This dataset covers data of the news headlines circulated over the period of about 15 years. The detail of content is given below.

- Format: Single file
- Publish_date: 02-19-2003
- Headline text: in English lowercase
- Total Records: 1,103,665

After data processing we will pass the dataset through the cleaning process of data. The major functionality of cleaning process are mentioned below.

- Removes section, chapters and subsection headings.
- Removes references and acknowledgement.
- Combines the words split by the hyphens.
- Removes the special characters from the document like, [: [>]\^% ~ # { } * % | & etc.

The cleaning procedure of the document has two primary objectives, first one of them is to filter out required features and second one is to formulate the corpus for the topic modeling stage. The cleaned form of the documents was store to the python having the option to eliminate stop words. The Stop words are the set of the words which are not as much significant their degree of significant depends on application itself. In our situation words like ‘and’, ‘the’, ‘of’, ‘or’, ‘for’ etc. will be neglected.

4.2 Experimental Results

In this section the performance of proposed model is evaluated with existing models. the model is evaluated in term of accuracy, resources consumed and time complexity. The performance of proposed model is evaluated with two existing topic extraction models e.g. LDA and PLASA. The proposed model is implemented in python language. The results are applied on ABC NEWS dataset. After the analysis it is found that proposed model perform better results in term of accuracy, system resources consumed and time complexity. Figure 4 explains the comparative results of proposed model with LDA and PLASA in term of accuracy. The accuracy of LDA and PLASA on same date set is respectively 80% and 82% however mixture model has the accuracy of 89%. Similarly, in fig 5 the comparison of LDA and PLSA model is made with Mixture Model (proposed model) in term of system resources consumed as the result proposed model consumed less system resources. In the same way the efficiency of proposed algorithm is evaluated in figure 6 in term of time complexity of algorithm with the LDA and PLASA. The experimental results of proposed method clearly shows that proposed model yields very promising results.

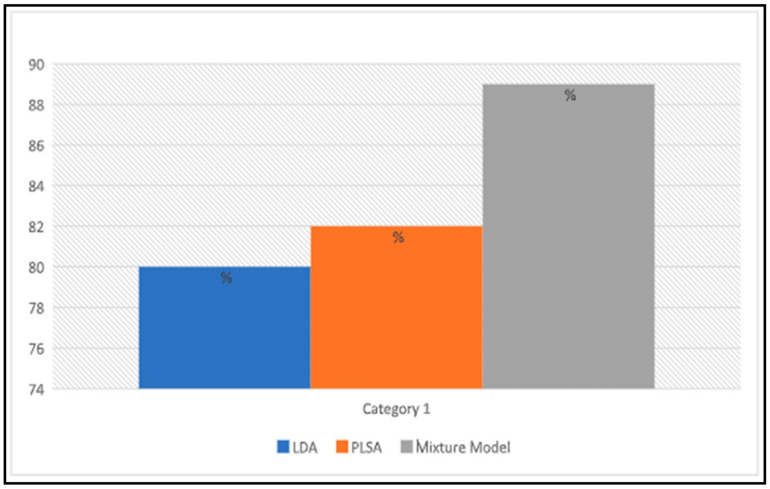


Figure 4. Comparative results of proposed model with LDA and PLASA in term of accuracy

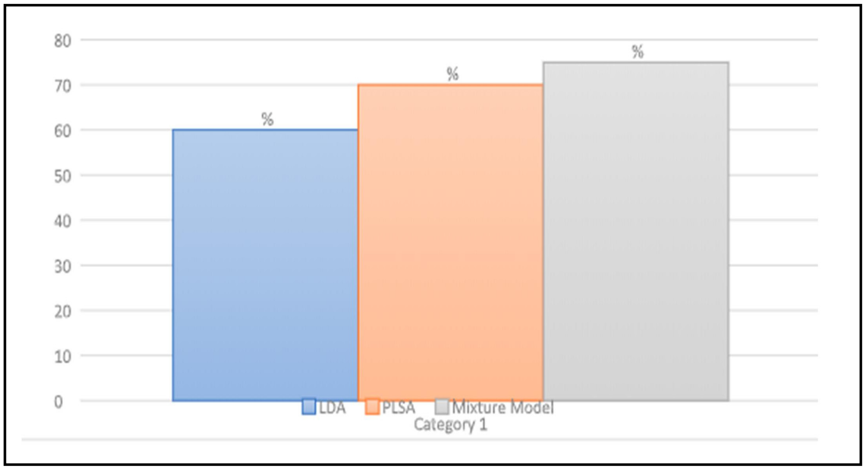


Figure 5. Comparative results of proposed model with LDA and PLASA in term of system resources consumed

5. Conclusion and Future Work

The suggested study indicates a comprehensive survey about topic mining and its recent research status. The collaboration between data mining and topic extraction techniques promotes to locate different important text patterns from retrieved reviews and articles. The proposed technique could be applicable to the variety of research documents in order to extract topics from them. One thousand articles from four hundred refereed journals are collected and analyzed textually through topic mining

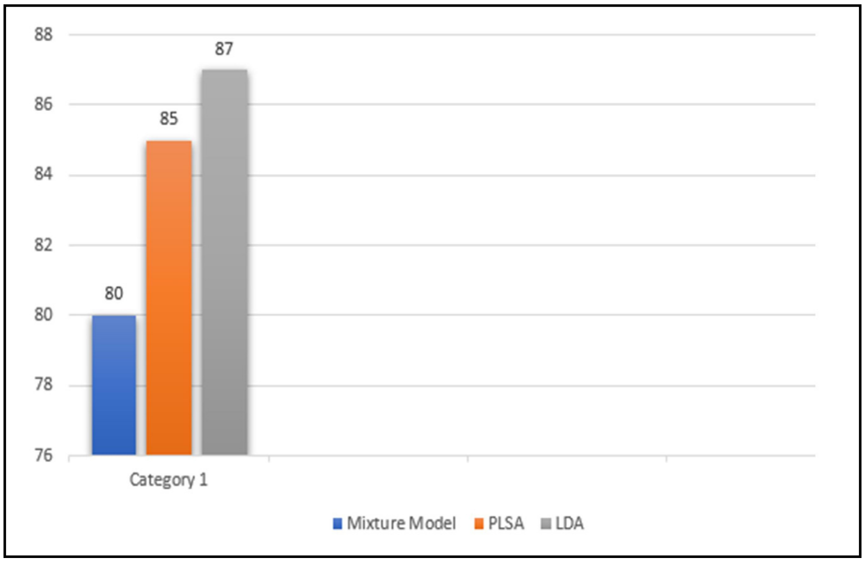


Figure 6. Comparative results of proposed model with LDA and PLASA in term of time complexity

techniques. These articles are collected from six major databases namely IEEE, Science Direct, Wiley, SAGE, Cambridge and Springer. The collection of the selected documents being based on the criteria that all selected documents incorporate the data from research articles, reviews, newspapers and different websites. In this proposed study, clustering of the text, selection of seed words from the clusters and collecting the seed words frequency is the major task applied to analysis the text. From the simulation results, it was clearly observed that the proposed solution outperforms the already existing model.

As the future work, our interest is to collect articles from the various research topics, i.e. not only focus on one or two research areas. It will help us to find out more compelling patterns from these articles and to find how such documents are dispersed among the various targeted databases. In addition, it will enable the similarity operators to work appropriately and to present a logical relationship among research articles.

References

[1] Padmaja, C. H ., Narayana, S. L. (2018). Probabilistic Topic Modeling And Its Variants—A Survey, *International Journal of Advanced Research in Computer Science,* p. 35-43, (May 1).

[2] Jia, Hailong., Fang, Lina. (2016). Design of Web Crawler Based on Improved Hidden Markov Model, *International Journal of u-and e-Service, Science and Technology,* p. 227-36, (August 30).

[3] Cao, Ziqiang Li., Sujian, Liu., Yang, Li., Wenjie, Heng, J. A Novel Neural Topic Model and Its Supervised Extension, *In:* Twenty-Ninth AAI Conference on Artificial Intelligence NLP and Machine Learning:

[4] Teja Santosha, D., Sudheer Babua, K., Prasada, S.D.V., Vivekananda, A (2016). Opinion mining of online product reviews from traditional LDA Topic Clusters using Feature Ontology Tree and Sentiwordnet, *International Journal of Education and Management Engineering (IJEME),* p. 34.

[5] Yang, Guangbing., Wen, Dunwe, Kinshu, Chen, Nian-Shing., Sutinen, Erkki . (2015). A novel contextual topic model for multi-document summarization, *Expert Systems with Applications,* p. 1340-52, (February 15).

[6] Alvarez-Melis, David., Saveski, Martin (2015). Topic Modeling in Twitter: Aggregating Tweets by Conversations. ICWSM, p. 519-22, (March 31).

[7] Chen, Mo., Yang, Xiao-Ping (2016). Research on model of network information extraction based on improved topic-focused web crawler key technology, *Tehnicki vjesnik/Technical Gazette,* (July 1).

[8] Chen, Xilun., Candan, Kasim., Sapino, Luisa, Maria. (2018). Incremental Multi-Scale Dynamic Topic Models. *In*: AAAI, p. 5078-5085.

[9] Qian, Shengsheng., Zhang, Tianzhu., Xu, Changsheng., Shao, Jie. (2016). Multi-modal event topic model for social event analysis, *IEEE transactions on multimedia,* p. 233-46., (February).

[10] Zhang, Yongjun., Ma, Jialin., Wang, Zijian., Chen, Bolun., Yu, Yongtao. (2018). Collective topical PageRank: a model to evaluate the topic-dependent academic impact of scientific papers. *Scientometrics*, p. 1345-72. (March 1).

[11] Nguyen, Thien Hai., Shirai, Kiyoaki Velcin, Julien (2015). Sentiment analysis on social media for stock movement prediction, Expert Systems with Applications, p. 9603-11, (December 30).

[12] Cai, Zhiqiang., Hu, Xiangen., Li, Haiying., Graesser, Art (2016). Can Word Probabilities from LDA be Simply Added up to Represent Documents, p. 577-578.

[13] Prabhudesai, Kedar., Mainsah, B., Collins, Leslie., Throckmorton, Chandra., S. (2018). Augmented Latent Dirichlet Allocation (Lda) Topic Model with Gaussian Mixture Topics., *In*: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 2, (April 15).