Micro-blog message based interesting key terms Identification

[1]R.Sateesh, [2] B.Rajesh, [3]V.Rajya Lakshmi, [4]M.S. Vani, [5]D.Sowjanya,

[1]Assistant Professor, [2]Assistant.Professor, [3]Assistant.Professor, [4]Assistant.Professor, [5]Assistant.Professor
Computer Science and Engineering,
Mother Theresa Institute of Engineering and Technology, Palamaner, India

Abstract: **Micro-blogging platforms like Twitter and Wikipedia are increasing main streams which provide user-made information for publishing and sharing messages. Identifying interesting and useful key terms from E-Learning documents are a critical issue in social media because users can suffer with data overload. To resolve the problem we suggest a novel topic model called Latent Dirichlit Allocation (LDA), LDA is an arithmetical and statistical model to discovering the interesting topics that occur in micro-blog. Interesting key words can be found based on the frequency of each word in the input; topics are categorized by using topic modeling technique. LDA works by, first partitioning each document into paragraphs and split into topic of words, Word frequency is a forwarding function acting as an important role in identifying key terms, the interesting key word can be based on frequency. Thus E-learning documents can be simply retrieved and classified using these methods which is also proven by experimental verifications. From the experimental result real world data from twitter shows that our proposed system out performs more than a few other baseline methods.**

Index Terms: **Micro-blog twitter, topic modeling, LDA, classification and clustering.**

I. INTRODUCTION

Micro-blog services like twitter, Facebook and MySpace are becoming an important communication tools for many online users to web communicate with each other [4]. Twitter (http://twitter.com) is one of the popular social medias It can provide user to share his own ideas and trending news in the form of tweets and messages, posts containing up to 140 characters, there are more than 500 million registers in twitter and they can produce 400 million tweets per a day [5] in micro blogging nature the large amount of tweets containing use-full data but hardly to identify the key contents. Amazon Mechanical Turk (AMT) is a one type of platform for predicting the large number of tweets in micro-blog, the twitter data set must be containing more than one lakh tweets to classify whenever there are a popular tweet's also existing un-popular tweets, we also find easily based on the frequency [8].E-learning is also called as electronic-learning which is used to learn the user generated data from the internet, and here we can use the e-learning data from social media and micro blog. Here we can analyze the large number of twitter content for finding the interesting key words and classifying by using topic modeling, also use k-means clustering algorithm for grouping the related content from twitter data. In previous paper they can find top key words based on ranking of each tweet, but here we use word frequency to list the top keywords. High frequency words are said as interesting key words and the low frequency words are un-interesting contents. Basically the large amount of data is not an easy task to analyze because of this problem we adopt a statistical and popular unsupervised approach called Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). It is a one type of text clustering model used mostly for data mining problems. The functioning of LDA is, first it adopts a large twitter data set and then predict that data, split into similar size of documents and, again split each document into a number of words after that we assign a frequency to each word and classify it based on frequency.

Topic modeling is a one type of classifier, it can classify the input data into a number of topics, and each topic containing similar words for each label. Labels are areas like politics, sports and movies, etc.., the LDA model contains two types: the word probability of words under the probability of topics, it's under the probability of documents, and test and train LDA data can be used in same labels.

In this study, we mainly focus on statistical method to identify interesting tweets to a wide audience. Here we perform LDA-based topic model analysis and find latent topics and assign word frequency for each latent word. Our contributions are summarized as follows:

1. We proposed a novel unsupervised approach that identifies interesting key terms in twitter and filters low frequency words said to be un-interesting key words from the twitter data set.

2. The text classification method called Latent Dirichlet Allocation (LDA) is used to analyze the data mining documents and propose an LDA based topic modeling to trending topic analysis.
3. Based on topic identification, we assign frequency to each word and topics as per its relative importance.

Then the remaining paper is organized as section 2, section 3, and section 4, in section 5 we examine related work, proposed work, and experimental result and finally conclude the paper.

Identifying interesting key terms on Twitter:

We treat the task of finding interesting keywords based on the frequency of the words; here we study mainly two concepts in this paper. Definition1. Interesting is a type of area in micro-blogging. It means the twitter content may be containing some potential interest to not only the user and their followers but also having high frequency words in twitter data, in other hand uninteresting means low frequency content present in the twitter data.

II. RELATED WORK

Social media can be proved by several studies covered about the twitter and micro-blogs. Rowan Nairn, Jilin Chen, Michael Bernstein, Les Nelson (2010) were the first to study the structure of the tweeter by investigating a number of Twitter features. Now the process is going on to maintain and analyze the twitter messages generated by the users, from large twitter data set. The mainly existing approaches (Hong, Dan, & Davison, 2011, Duan, Jiang, Qin, Zhou, & Shum, 2010; Liangjie Hong Ovidiu Dan Brian D. Davison, 2011;) proposed to view influence, twitter counts as fame measure, interestingness and present classifiers used to how we can deal with those in the future. They oppressed different features of twitter, such as user Meta data, textual data, and propagation information, social media or twitter containing mostly textual data and linkage data. The linkage data can be described as a re-tweet data in the social media, Gerster, Ji, Alonso, Carson, and Nabar (2010) used a crowd sourcing mechanism to categorize the twitter short messages to group interesting and un-interesting key words. The data may be occurring in link is a single, highly effective data, Ntoulas, Law, and Kenthapadi (2010) used a recommendation system to identify interesting tweets, but he does not prove experimentally. Godoy, Armentano, and Amandi (2012) defines topology of followers and followees, also identify the users using micro-blogging related factors. In a micro-blogging area, topic modeling and semantic analysis are widely used to analyze textual data and help in many users related modelings. (Content filtering (Duan & Zeng, 2013; Pennacchiotti & Gurumurthy, 2011), sentiment analysis (Lin & He, 2009), Martinez-Romo & Araujo, 2013), to classify and analyze the topical differences between traditional media and twitter using twitter- LDA for investigating short messages. LDA based topic modeling regards the combination of trending topics and times to distribute them, the time distribution is based on the word frequencies. (Ramage, Dumais, & Liebling, 2010) deal with Labeled-LDA model here a tweet uses its labeled information, and then built the probability distribution for latent topics to represent the tweet's content. Ramage et al. (2010) study the similarities between the topic modeling and interestingness, hear interesting contents are based on the word frequencies and trending topics also collect from those. In proposing system Min-Chul Yang, Hae-Chang Rim (2013) invent a topical analysis for identifying interesting twitter contents from the large amount of the micro - blog dataset.

III. PROPOSED SOLUATION

Due to the disadvantages described in the existing system, we propose a new method to overcome its drawbacks. The new system implements Latent Dirichlit Allocation (LDA) based topic modeling along with a text mining concept to prove that the proposed system outperforms the existing system.

The Process involved in the proposed system can be described as follows: A twitter dataset containing much number of tweets or messages in the form of unstructured format. In that dataset we can perform a pre-processing mechanism, where we can remove stop words, numbers, punctuations, mailing constants and finally stemming words this process is said to be a data cleaning. Then we can get a structured data and create a term document matrix (TDM) and Document term matrix (DTM), both can generate a number of tweets and containing terms as well as words. After creating a document matrix we must calculate the word frequency for all terms containing in the twitter data set. There are two types of matrixes head frequency matrix represent an interesting keywords and low frequency specifies un-interesting key terms. Now, we must compute the term weighting factor by using the product of the frequency of words. This weight is plotted on the matrix which in turn becomes the word cloud or a topic model graph. The graph values can be processed by using LDA to represent high frequency words. Thus the proposed system outperforms the existing system in accuracy, performance and efficient retrieval and classification of documents.

LDA Algorithm

Latent Dirichlit Allocation is the most widely used classifier in topic modeling. LDA assumes "M" number of documents in that document containing "N" number of words and each word "W" contain assigned frequencies.

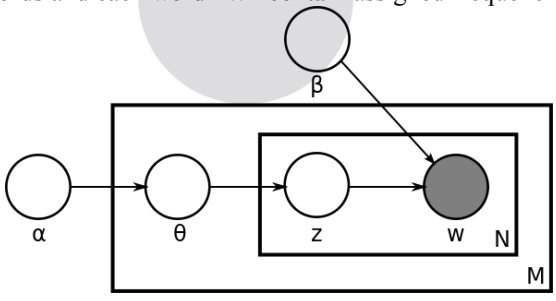

Fig.1. Plot notation for LDA algorithm

One of the main advantages of LDA is that it could easily split the documents in to words and cluster top key terms; LDA is a type of un-labeled topic modeling classifier. Working of LDA generative algorithm

The working of LDA has the following steps

1. Collecting “M” twitter documents
2. Determine term distribution
3. Determine topic distribution for documents
4. Split documents in to number of topics
5. Split topics into number of words
6. Determine word frequency’s
7. Classify based on frequency

Above plot representation for LDA can represent a sequential flow of the model, here “M” represents a document (collection of tweets) N represents a number of topics present in the document and the W is a number of words in the each topic, then the parameters represent probability distributions.

IV. SYSTEM DESIGN

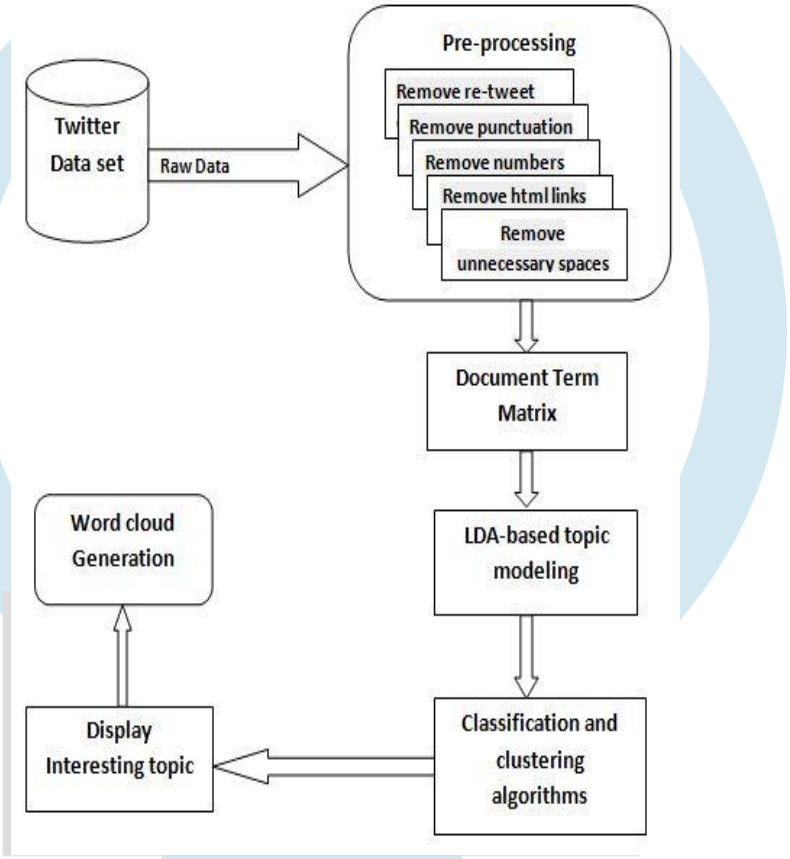

Fig.2. System Architecture

The structural evaluation of the project may be described as, the twitter dataset is a collection of millions of tweets in the form of unstructured format or said to be raw data. Every tweet contains valuable data and also useless data; from here we can gather useful and interesting key terms. Here we can go for interestingness of the twitter data so we focus on use-full data and may or may not remove useless data.

A. LOADING THE CORPUS

A corpus might be a collection of user generated data in twitter and micro-blog; each corpus has a number of topics and tweets. Data may be authenticated or freely available in twitter.

B. PRE-PROCESSING THE DATA

Preprocessing is a type of data analysis technique it can do mainly in two types Data Cleaning and Stemming, in Data cleaning we can perform some transformations including converting the text to lower case, removing stop words and punctuation, , removing re-tweet entities, removing numbers, removing html links, removing unwanted spaces. In Data Stemming we can remove the same words and identifying synonyms. Stemming use an algorithm that removes frequent word endings for English words, such as \es", \ed" and \'s". All this transformation is done under the text mining.

C. DOCUMENT TERM MATRIX

Document term matrix (DTM) is a one type of matrix having rows as terms and columns as topics, term document matrix is an inverse of DTM and it contains frequency of each word in the dataset. Head frequency denotes top key words in twitter dataset. In classification algorithm also use this to classify topics based on word frequency.

D. CLISSIFICATION ALGORITHM

Classification is a process of supporting the area of topics, Here we can use an LDA based topic modeling classifier, used mostly in large data mining areas where the trending and interesting key words are found to classify the similar words and topics.

E. CLUSTERING ALGORITHMS

Clustering means grouping of similar things or topics to one, In this paper we can use k-means clustering algorithm, it is a one of most used partitioning methods. K-means clustering handle large data sets, the process is first select k random rows, then assign every data point to closest centroids, recalculate the centers and again assign centroids.

F. WORD CLOUD GENERATION

A word cloud is a type of plot, it can represent the number of words in cloud forms the high frequency word can be displayed in bold and big size. The words mostly follow colors in random order the use of word cloud is that we easy to find the most popular and interesting topic in the dataset.

V. EXPERIMANTAL RESULTS

A primary thing is to find the interestingness of the twitter data and key words. Generally every tweet must have some message and that can deal with the particular area. In this paper, we can deal with the large amount of tweets grouped to gather together to form a corpus or data set, we can use the LDA based topic modeling to apply classification and k-means algorithm for clustering mechanisms on the input data. Before we apply the LDA algorithm do the data cleaning process, it contains two main techniques Data pre-processing and stemming. Pre-processing is a type of data cleaning process here we can remove unwanted data from input data like removing numbers, punctuation, spaces and http links. Data stemming is a process of removing synonyms. After that data cleaning process creates a Term Document Matrix and Document Term Matrix. These two matrices contain words as columns and rows containing frequency's of all key words in twitter data. Then we find the head frequencies of key terms, using topic modeling with an LDA algorithm to classify the interesting topics based on frequency.
Clustering is a grouping of interesting key words and topics from twitter data; in this paper K-means clustering to group similar topics and key words.

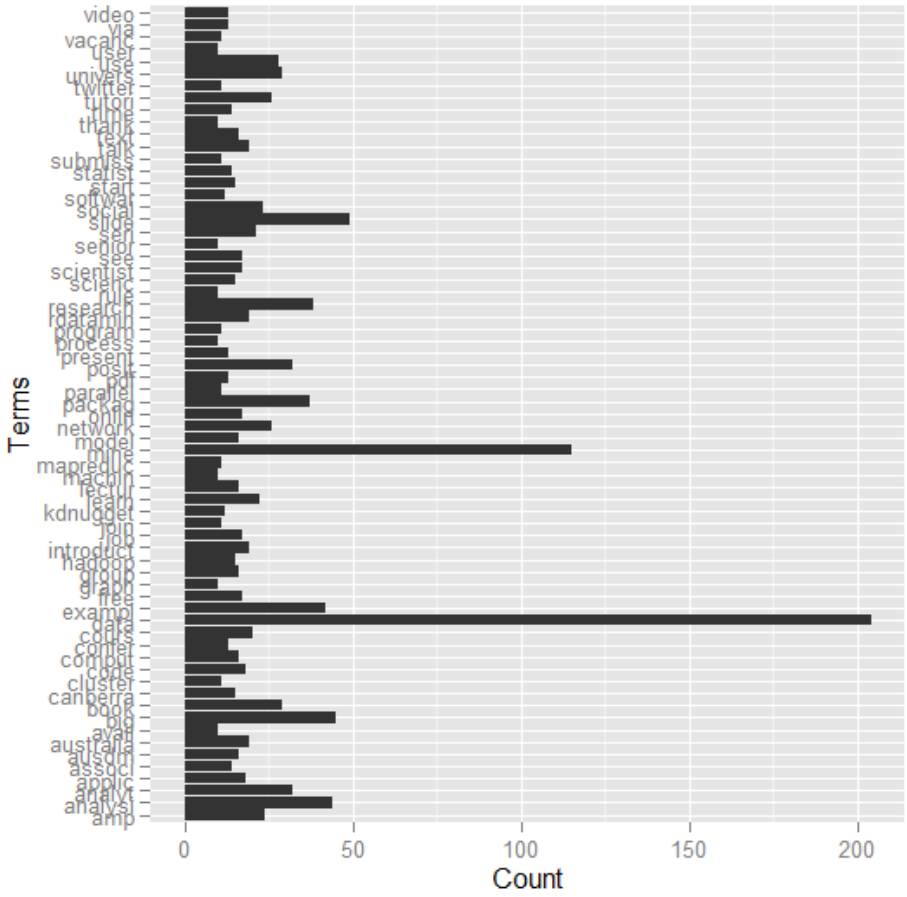

Fig.3. Top key words and their frequencies

The histogram represents the high frequency terms and their frequencies, in that the high frequency word is data and it presents more than 180 times in the twitter data set.

```
> dtm <- as.DocumentTermMatrix(myTdm)
> library(topicmodels)
> lda <- LDA(dtm, k = 10)
> term <- terms(lda, 10)# first 4 terms of every topic
> term
      Topic 1     Topic 2    Topic 3    Topic 4    Topic 5     Topic 6
 [1,] "data"      "talk"     "data"     "slide"    "posit"     "tutori"
 [2,] "scientist" "packag"   "ausdm"    "data"     "analysi"   "exampl"
 [3,] "big"       "user"     "submiss"  "big"      "research"  "packag"
 [4,] "analyt"    "code"     "cfp"      "machin"   "univers"   "use"
 [5,] "vacanc"    "rstudio"  "mine"     "join"     "social"    "hadoop"
 [6,] "melbourn"  "exampl"   "confer"   "canberra" "network"   "comput"
 [7,] "tool"      "program"  "due"      "learn"    "introduct" "parallel"
 [8,] "australia" "canberra" "amp"      "ieee"     "lectur"    "present"
 [9,] "googl"     "webinar"  "juli"     "engin"    "data"      "mapreduc"
[10,] "technolog" "give"     "workshop" "miner"    "video"     "larg"
      Topic 7   Topic 8   Topic 9    Topic 10
 [1,] "mine"    "mine"    "cours"    "data"
 [2,] "job"     "data"    "free"     "mine"
 [3,] "data"    "book"    "onlin"    "rdatamin"
 [4,] "rule"    "applic"  "data"     "seri"
 [5,] "statist" "exampl"  "start"    "analysi"
 [6,] "associ"  "pdf"     "mine"     "analyt"
 [7,] "thank"   "see"     "packag"   "kdnugget"
 [8,] "senior"  "case"    "stanford" "scienc"
 [9,] "exampl"  "chapter" "learn"    "time"
[10,] "analyst" "studi"   "forecast" "text"
```

Fig.4. LDA Topic modeling

Above screenshot represents topic modeling of 10 topics and each topic, contain related top key words, the words are linked from the term document matrix. TDM contains the words and their frequent rights, every topic containing the interesting key word the top word on every topic are the most interesting word in twitter data set.

Fig.5. Word cloud for topic modeling

Word cloud represents an easy understanding and visualizing of the top key words, this may be useful in understanding of topics. In that the bold word is top key word in giving twitter data set.

```
> k <- 8
> kmeansResult <- kmeans(m3, k)
> # cluster centers
> km <- round(kmeansResult$centers, digits=3)
> km
    amp analysi analyt   big  book  data exampl  mine
1 0.019   0.000  0.014 0.053 0.048 0.183  0.062 0.067
2 0.037   0.037  0.000 0.000 0.000 0.074  0.148 0.037
3 0.077   0.077  0.154 1.385 0.000 2.385  0.000 0.385
4 0.000   0.951  0.024 0.049 0.000 0.293  0.146 0.146
5 0.000   0.020  0.157 0.039 0.020 1.196  0.000 1.059
6 0.000   0.000  0.000 0.000 0.773 0.773  0.682 1.136
7 0.044   0.044  0.378 0.200 0.000 0.689  0.000 0.067
8 1.143   0.000  0.071 0.214 0.071 0.857  0.286 0.500
  research slide social tutori univers   use
1    0.010 0.135  0.005  0.053   0.034 0.058
2    0.037 0.111  0.000  0.185   0.000 0.148
3    0.077 0.231  0.000  0.077   0.000 0.231
4    0.073 0.098  0.390  0.122   0.049 0.098
5    0.039 0.118  0.000  0.059   0.039 0.039
6    0.000 0.045  0.000  0.000   0.000 0.091
7    0.644 0.000  0.133  0.022   0.400 0.000
8    0.000 0.286  0.000  0.000   0.000 0.071
```

Fig.6. K-means clustering for top tweets

K-means is a one type of greedy, computationally efficient technique and also most popular representative based clustering algorithm. Initially we can give an input as a set of entities to be clustered at last it gives a set of cluster labels. The procedure is as follows:

- First, define initial group centroids, here we select

Random values of centroids and define K entities.

- Then assign each entity to the cluster that has been closest to the centroid.
- Next recalculate the values of centroids.

```
cluster 1: data slide mine exampl use big tutori
cluster 2: packag tutori exampl use slide data am
cluster 3: data big mine slide use analyt packag
cluster 4: analysi network social data exampl min
cluster 5: data mine analyt slide tutori big pack
cluster 6: mine book data exampl use packag slide
cluster 7: data research posit univers analyt big
cluster 8: amp data mine exampl slide big analyt
```

Fig.7. 8 clusters divided by using K-means algorithm

VI. CONCLUSION

We proposed a novel un-supervised approach to discover interesting key terms from the user generated data in Twitter. To analyze textual data more effectively, we developed a new LDA based topic modeling technique which is used to determine the interestingness of an individual tweet, we first extracted micro-blog topics based on our proposed functions. From the latent variables of LDA, we calculate word frequency of a topic by utilizing its representative words. Topic modeling is a one type of classifier; it can use the Term Document Matrix (TDM) in text mining to construct a number of topics. Also, for each topic we calculated two types of frequencies. One is high and another is low; here high frequency denotes interesting topics and low frequency denotes un-interesting key terms in twitter data. The clustering algorithm is making the interesting topics into groups and assigning a topical identification for each cluster. From our observations, each latent topic with a high weight value covers a specific topical theme. We then investigated how important topics spread with a target tweet. In a series of experiments, we demonstrated the ways in which our model can be naturally applied to recommend, filter, and understand textual posts in micro-blog. In terms of understanding a large number of tweets, weighing and analyzing latent topics using the LDA model can reduce cost and complexity because the clusters from each topic can be viewed as a set of significant contents.

REFERENCES

[1] Martinez-Romo, J., & Araujo, L. (2013). Detecting Malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications, 40, 2992–3000.
[2] Armentano, M., Godoy, D., & Amandi, A. (2013). "Follo Wee recommendation in twitter based on text analysis of micro-blogging activity". Information systems, 1116-1127.
[3] Lee, C.-H. (2012). "Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams". Expert Systems with Applications, 39, 13338–13356.
[4] Hong, L., Dan, O., & Davison, B. D. (2011). "Predicting popular messages in twitter" In Proceedings of the 20th international conference companion on World Wide Web WWW '11 (pp. 57–58).
[5] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or news media? In Proceedings of the 19th international conference on World Wide Web WWW '10 (pp. 591–600). New York, NY, USA: ACM.
[6] Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. (2010). Short and tweet: Experiments on recommending Con tent from information streams. In Proceedings of the SIGCHI conference on human factors in computing systems CHI 10 (pp. 1185–1194). New York, NY, USA: ACM.
[7] Lauw, H. W., Ntoulas, A., & Kenthapadi, K. (2010).Estimating the quality of postings in the real-time web. In Proce-edings of the WSDM 2010 workshop on search in social media SSM '10.
[8] Alonso, O., Carson, C., Gerster, D., Ji, X., & Nabar, S. U. (2010). Detecting uninteresting content in text streams. In Proceedings of the SIGIR 2010 workshop on crowd sourcing for search evaluation CSE '10 (pp. 39–42).
[9] Ramage, D., Dumais, S., & Liebling, D. (2010). Cairo -cterizing microblogs with topic models. In Proceedings of the Fourth international AAAI conference on weblogs and Social Media. AAAI.
[10] Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit Attribution in multi-labeled corpora. In Proceedings of the 2009 conference on empirical methods in natural language Processing: Vol. 1 – Vol. 1 EMNLP '09 (pp. 248–256). Stroudsburg, PA, USA: Association for Computational Linguistics.
[11] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
[12] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., et al. (2011). Comparing Twitter and traditional media Using topic models. In Proceedings of the 33rd European Conference on advances in information retrieval ECIR'11 (pp. 338–349). Berlin, Heidelberg: Springer-Verlag.

[13] Geng, L., & Hamilton, H. J. (2006). Interestingness Measures for data mining: A survey. ACM Computing Surveys, 38.
[14] Armentano, M., Godoy, D., & Amandi, A. (2013). Followee recommendation in twitter based on text analysis micro-blogging activity. Information systems, 38, 1116–1127.
[15] Duan, J., & Zeng, J. (2013). Web objectionable text Content detection using topic modeling technique. Expert Systems with Applications, 40, 6094–6104.