International Journal of Research in Advent Technology, Vol.6, No.9, September 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

Topic Modeling for Online Social Network

Reena Pagare[1], Akhil Khare [2]
Research Scholar, PAHER [1],Udaipur India[1]Akhil Khare, MVSR COE, Hyderabad [2]
Email: reenawp5@gmail.com[1], khare_cse@mvsr.edu.in[2]

Abstract- Huge amount of data is generated due to use of social network. The need to study and analyze this data to generate information to find which topic is discussed heavily in the network is topic modelling. Existing study show that LDA is an effective method for topic modelling. LDA has shown to produce good results over many domains. Consideration of social attributes along with the content on the network will increase the accuracy and efficiency of the hot topics identified. We propose a modified LDA model by considering different attributes of social network

Index Terms- Topic modelling; social network; latent dirichlet allocation; topic detection; trendy topics.

1. INTRODUCTION

The recognition of the well-liked topics in the flow associated with communications made by the members of the OSN depends on the id associated with bursts. There are primarily two methods to identify this kind of models, through examining (i) phrase rate of recurrence or even (ii) social interaction rate of recurrence.
Topic recognition entails discovering the event of the new hot topics like a plane crash, the murder, the political news, or even scandal information news through several resources. Topic tracking may be the process for checking the flow associated with information tales to locate the ones that monitor (or discuss) exactly the same occasion as you specific with a person. There are lots of current investigation functions focusing on discovering the topics by way of the topic modeling in the social press information. These types of investigation are modeled to find out the fundamental crucial topics which happen within some online articles that aid businesses to keep track of as well as sum up information that individuals tend to be talking about the social network. An additional number of investigations happen to be concentrating on determining the individual viewpoint by way of the sentiment evaluation in the social news information. This type of investigation is modeled to recognize user's sentiments depending on exactly what they've discussed online to comprehend individuals fulfillment upon every topic. Nevertheless, many of these functions don't classify the information based on their groups; however evaluate the information as entire.
With regard to social news information, it's quite common that individuals make use of emoticons or even acronyms or even each expressing their own feelings. With this particular unique function associated with social news information, numerous researchers by using this among the elements to enhance the precision from the sentiment evaluation outcome. A few of the earlier outcomes including a

good emoticon to their sentiment evaluation tend to be created. Some other models are Multinomial Naive Baye (MNB) as well as Support Vector Machines (SVM). Rather than utilizing emoticon to classify twitter, very first produced emoticon and acronym literature needs to pre-process the information. Utilizing prepared information, this classifies twitter in line with the consequence of earlier polarity associated with phrases.
Related Topic Recognition would be to identify models or even developments upon some topics which happen often collectively inside a twitter with time models and the relationship in between information as well as twitter posts. Current investigation utilizes studies upon politics viewpoint as well as twitters posts which was submitted within the exact same time period to investigate the relationship in between sentiment key phrases as well as caused by the poll. The sentiment of every topic is used to calculate the Pearson relationship.
Topic modeling is really a crucial topic within textual content exploration. Probably the most popular device with regard to topic modeling is Latent Dirichlet Allocation (LDA), which has already been prolonged with regard to social networking information evaluation. Existing studies suggested a many of strategies to apply LDA as well as evaluate their high quality usage. Study also suggested adding LDA in to community recognition. Other few investigations suggested the models to find out groups of the organizations as well as topics.

2. RELATED WORK

Latent Dirichlet Allocation (LDA) [1] is a classic model for topic modelling of documents. LDA detect underlying topics in text documents. LDA is an unsupervised, probabilistic model. LDA sets that word which have high meaning correlation and documents which talk about same topics will utilize words or group of words with similar meaning. Latent topics are subsequently found by recognizing a bunch of words in the dataset that are repeatedly found together within the document. In this way LDA models, document distribution over latent topics, where every topic is differentiated on the basis of its own correlation over words. Every topic is, in turn, demonstrated as an infinite mixture over a hidden set of topic probabilities. With regards to text modeling, the topic

probabilities give an correct representation of an document. Using topic models for identifying trendy topics in social network has gained attention. LDA has been appeared to be powerful in some text-related tasks for example, document classification, yet the feasibility and viability of utilizing LDA as a part of IR tasks remains mostly unknown. Having completely generative semantics, LDA is having more improved model than past point models, for example, pLSI. Language modeling, which is a standout amongst the most prominent measurably principled ways to deal with IR, is likewise a generative demonstrate, inspiring us to analyze LDA-based document representations in the language modeling structure. [2] propose non-Markov online LDA topic model which uses Gibbs sampling called OLDA. Online LDA is a incremental topic model. The topic model is updated after every time slice using model generated in the previous time slice. The OLDA generates an evolutionary matrix to store the generative process of every topic over time and thus store the evaluation of each topic over time and thus permits to detect bursty topic. Multi-Document Summarization manages [3] presents a method where LDA and SVD are combined. The aim is that the topic should be represented in form of a sentence. The sentences should be as few as possible and they should cover different events in the document. Latent Dirichlet Allocation finds the different topics in these documents. However to reduce the similar type of data available in the documents , the documents have to satisfy the orthogonal property . Been orthogonal minimized the possibility of documents having similar content or correlation. LDA is used to find the different topics in the documents and using SVD the sentences that best represent these topics are found. In this paper [4], they propose a method based on the word-occurrence frequency. A topic is controlled by recognizing words that show up with high frequency in the topic and low frequency in other topics. The method not only identify words which are frequent but also identify the hierarchy of words. It demonstrates patterns of word co-occurrence furthermore, co-occurrences of those patterns utilizing a hierarchy of discrete latent variables. The conditions of the latent variables represent clusters of documents and, they are interpreted as topics. The words that best recognize a cluster from different clusters are chosen to describe the topic. It handles words as binary variable. Online social network generates huge data which is dynamic in nature and keeps changing as per the time interval[5] . Finding popular topics or events thus gives an insight into which topics are trendy and are generated from which source. Vast amount of research is done for topic detection and thus there are variety of techniques available for finding trendy topics.The authors give comparison of six topics identification method on three different datasets. The topic detection problem is dynamic in nature and varies with time interval. The study is based on the nature of the event, the activity, the sampling methods used and preprocessing done one the dataset. In [6] author uses wavelet signal shows a lightweight event detection utilizing wavelet signal for detection of event. The author uses hashtags available in messages in any microblog. It uses LDA based on Gibbs sampling. The use of hashtag proved to give better results for event detection. Also combination of wavelength signal and LDA along with hashtag increases accuracy. [7] propose a method which is a combination of temporal and social properties of the stream of messages. The hot topic is related to terms frequently occurring in a specific time interval. To increase the accuracy of the result they have considered the social characteristics of the users in the network. The LDA model is used to work on sparse data and short messages like tweets, post and thus discover topics underlying in microblog [8]. The author proposes a modified version of LDA for microblogs and also presents a distributed version of the same algorithm. [9] uses random walk to find the similarity between events. A parallel approach is used to enhance the efficiency of the method.

Most of the work on topic consider the content of tweets. We believe that for online social network along with content, social parameters also play a substantial role in identifying popular topics. Thus along with tweets content we also consider retweet count, hashtag for identifying the popular topics. Traditional topic modeling method (eg LDA and PLSA) fail to produce required accuracy when applied to microblogs. Thus there is need to consider the social attributes while modelling for microblogs. In this paper [10] author proposes to use social attributes related to the messages like retweet or repost count, comment count, likes which contain important data and using these parameters can improve the accuracy of topic modelling for online social network. Learning significant topics models with huge document collections which contain a huge number of documents and billions of tokens is challenging in light of two reasons. Initial, one needs to manage an extensive number of topics (normally on the request of thousands). Second, one needs an adaptable and effective method for distributing the calculation over various machines. Clearly, the greater part of the recent efforts is towards increasing the accuracy of topic models. The proposed work thus work in the same direction of increasing the accuracy by making use of social attributes.

3. FRAMEWORK

Latent Dirichlet Allocation (LDA) [1] is a classic model for topic modelling of documents. LDA detect underlying topics in text documents. LDA is based on the assumption that documents with similar topics will use similar groups of words. Documents are probability distributions over latent topics. Topics are probability distribution over words. Thus LDA is an unsupervised, probabilistic model.

Generative Process

LDA assumes that new documents are created in the following way

1. Determine number of words in document
2. Choose a topic mixture for the document over a fixed set of topics

3. Generate the words in the document by
   i. First pick a topic based on documents multinomial distribution above.
   ii. Next pick a word based on topics multinomial distribution

Working Backwards

Suppose there is corpus of documents LDA is used to learn the topic representation of K topics in each document and the word distribution of each topic. LDA backtracks from the document level to identify topics that are likely to have generated the corpus. The algorithm is presented below:

1. Randomly assign each word in each document to one of the K topics.
2. For each document d:
   a. Assume that all topic assignments except for the current one are correct
   b. Calculate two properties
      i. Proportion of words in document d that are currently assigned to topic t = p(topic t | document d)
      ii. Proportion of assignments to topic t over all documents that come from that come from this word w = p(word w 1| topic t)
   c. Multiply these two proportions and assign w a new topic based on that probability , p(topic t | document d) * p(word w| topic t)
3. Eventually we'll reach a steady state where assignments make sense

To summarize LDA takes a number of documents. It assumes that the words in each document are related. It then rises to figure out the method for how each document could have been created. We just need to tell the model how many topics to construct and its uses that generative process to generate topic and word distribution over a corpus. Based on that output, we can identify similar documents within the corpus.

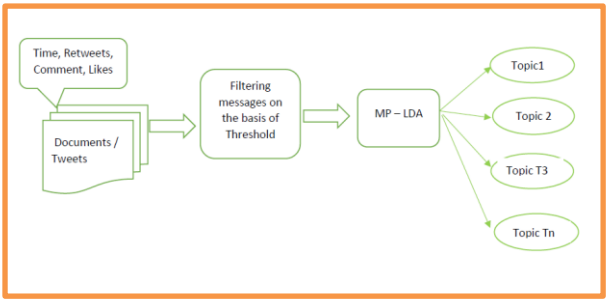

Fig 1: Framework for Finding Trendy topics in Microblog

We consider the hashtag, comment count, retweet count and likes for a particular post to identify if the topic is trendy or not.

The posts are filtered on the basis of retweet count, comment and likes. Topics which are popular will have a higher retweet and comment count and followers count. The post whose repost count and comment count is above a particular threshold are considered for topic detection. Discarding these low count messages will therefore increase the efficiency since the unwanted information will not be considered during modelling. Identification of these post which are useful is therefore important. We perform an empirical analysis to decide the threshold value. When the experiment was done to observe the distribution of messages as compared to retweet count and comment count and likes of the post, we found that most of the messages have less than 100 retweet and comment count. As the count reaches 500 the distribution in number of messages is almost constant The data used for this study consist of 15,20,000 messages. Figure 2 gives the details of experimentation done. Therefore we choose the threshold as 500.

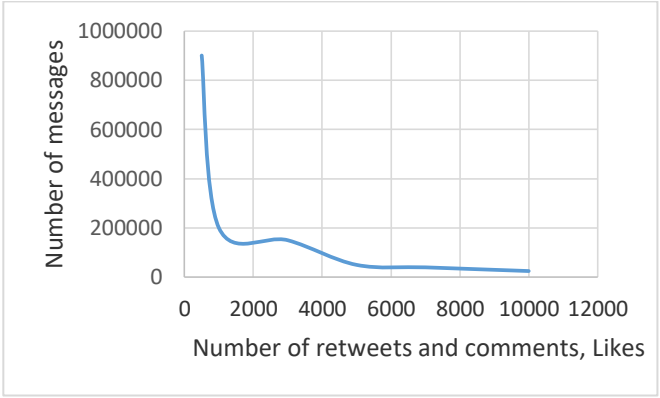

Fig 2: Messages Vs Retweet, Comments count and likes

The different twitter attributes that are captures are time, retweet count, comment count, followers count, hashtag, tweet content. Among all the attributes time is one of the most important parameter while finding trendy topics. Thus incorporating time is challenging. Also not all gathered post have hashtag and there consideration of hashtag is also challenging.

3.1. *Multiple Parameters LDA (MP –LDA):*

MP – LDA is a probabilistic graphical model based on Twitter LDA [Zaho et al. 2011]. Fig 1 gives a layout of MP-LDA. MP-LDA models each document as one topic and generate each word from one topic. Unlike Twitter LDA it generates two topics: hot topics and general topics. The hot topics or general topics are decided on the basis of time distribution. The hashtags are used will affect the

generation of topic and word both. The graphical representation of MP-LDA
model is shown as below in Fig 3.

Table 1. Notations

Symbol	Meaning
Nu	No of tweets of user U
TN	The number of topics
V	Voculabary size
α, β	Dirichlet Parameter
$\vartheta g, \vartheta h$	General , hot topics
$\emptyset\, t$	Word distribution for topic t
Nm	The length of tweet t
Yu,s,n	Topic of nth word in tweet t for user u
Wu,s,n	The nth word in tweet t for user u
P	Binary variable for time distribution
Pw, Pd	Time distribution about word and document
δM	Vector of hashtag for documents
δV	Vector of hashtags for words
θu	Topic distribution of users u

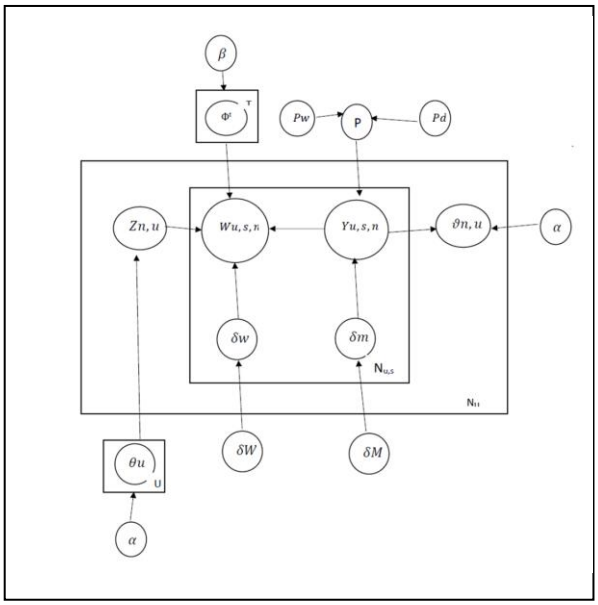

Fig 3: Graphical Representation of MP-LDA

Let $\emptyset\, t$ be word distribution for topic t. Parameter P decides whether the current word inherits topic from general topics or hot topics. The value of p is affected by two variables Pw and Pd. Pw is calculated as follows: At random select any tweet from existing dataset. In second step the dataset is splitted by the time interval and the value of A_t is the number of recurrence of word w in (t+1)th time interval. The time interval selected is one day. The eigenvalue of time distribution for w is obtained as follows ($A_{w,o}$,A_{w1} , A_{w2}, A_{w3} ,.....A_{wt}).

$$B_w = \frac{\sum_t (Awt - Aavg)^2}{T\ X\ Aavg^2}$$

In the equation above $Aavg$ is the mean value of Awt. The value of B_w for a hot term is likely larger than a general one. Pw = 1 when $B_w > 0.5$ and Pw = 0 when $B_w < 0.5$.In a document Pd = 1 if Pw =1 else Pd = 0 . The thought that a document if has word which is related to hot topics then the document is associated with hot topic. The final value of P is obtained by 'oring' the value of Pw and Pd. The hashtag in the model is associated with a topic and word. Since here is document is one tweet. The hashtag is associated with every tweet. The vector of hashtag is defined as follows $\delta M = \{\delta 1, \delta 2, \delta 3, \ldots \delta nu\}$ for hashtag – documents. The element δM can take value 0 or is either 0 or the occurance count of hashtag on overall dataset. Similarly other vector δV **f**or hashtag –words. The vector values will control the sampling of hot topic and word together with the vector ϑh and $\emptyset h$. Learning of latent variable is done by using Gibbs Sampling.

4. RESULTS

4.1 Dataset

The twitter dataset was considered for experimentation. The twitter data was collected using Twitter API. The twitter data was collected from March 2018 to August 2018. Due to restriction by twitter only 15 lacs of tweets were collected. The data collected included fields like userid, username, tweet, retweet count, mention count.

Table 2. Dataset Statistics

Dataset	Users	Tweets
Twitter	95,283	15,20,000

Table 2 gives statistics of users and the retweet count and messages. The graph was plotted for 10,000 users and it was observed that there are around 4000 users who did not have any retweet or comment on their posted tweet. And around two thousand users had a retweet count of more than 500, around 2000 users who have retweet and comment count in the range of 2000 to 7000.

The figure 5.4 shows the distribution of messages over different domains. The diagram below shows that floods and health and politics topics were mostly covered in the tweets captured.

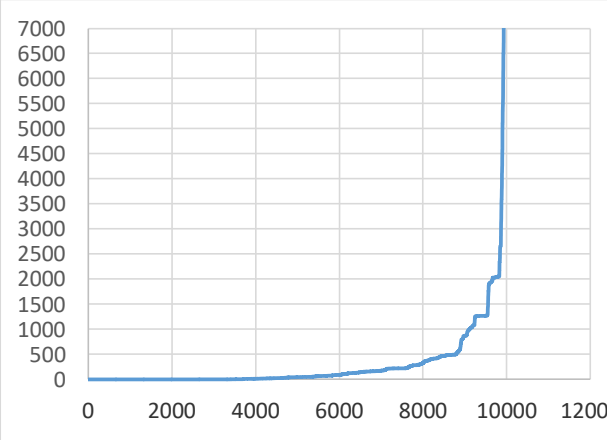

Figure 4: No of users Vs Retweet count and comment count

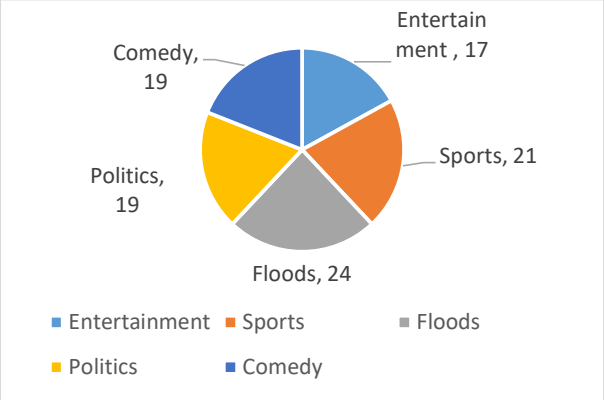

Fig 5 Distribution of Messages in the dataset over different domain

4.2 Metrics

Accuracy: The accuracy metrics is used to show how accurately the topics are identified by the proposed system. Coverage rate is to calculate the accuracy of the system. Greater the value of coverage rate greater the accuracy of the system. .

Coverage Rate = $\frac{Extracted\ hot\ topics}{Actual\ hot\ topics}$ X 100%

We have compared the result of our method with the baseline LDA method and Twitter LDA. It was observed that the result obtained by our method which considered social attributes gave better accuracy as compared to traditional LDA method.

Table 3. Coverage Rate

	Top 10	**Top 20**	**Top 30**
MP-LDA	0.66	0.75	0.78
T-LDA	0.45	0.50	0.53
LDA	0.23	0.35	0.40

Perplexity:
Perplexity is a measure of how well a sample fits the model. It is used to evaluate the probability model. Lower the value of perplexity the better is the probability distribution and predicting the sample. The graph shows the perplexity value for different values of topics T = {10, 20, 30, 40}. It can be observed from the graph that MP-LDA performs much better in terms of finding trendy topics since we consider the retweet information and user information along with hashtag for finding trending topics.

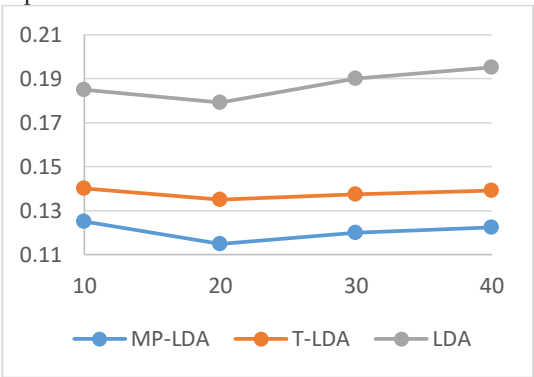

Figure 6: Perplexity Vs No of Topics for Twitter Dataset

When LDA was executed on the given data with 400 iterations, the below result was achieved. The top 3 trendy topics along with the top word are listed

Table 4. TOPIC and Associated TOP WORD

Topic Id	**Top Words...**
0	flood feeling dirty stressed relief kerala
1	donaldson josh indians american game football cleveland
2	health care happy mental love repeal _ happiness

Twitter LDA was applied to the same dataset. Below is snapshot of topic distribution over users. Topic count over users is shown in the below fig 8. We can observe that user related to the first file is tweeting more about topic 11 and topic 17 along with most prominent topic 0.

1.txt	16	0	3	8	1	1	1	0	61	7	2
	2	4	0	0	17	0	4	80	0	0	8
	0	2	56	5	1	0	7	1	17	0	9
	10	1	0	17	0	1	2	11	17	1	11
	7	41	0	1	0	0	1	4	0	1	0
	11	11	1	2	6	1	0	6	0	0	17
	1	0	11	0	2	1	0	3	17	40	4
	50	2	80	3	1	12	4	3	8	5	32
	45	3	8	0	11	0	29	1	1	1	0
	1										
2.txt	0	9	1	0	9	0	0	1	0	2	216
	0	0	3	0	0	37	1	0	7	326	0
	0	0	1	0	0	0	0	0	0	0	0
	0	3	114	0	8	120	0	0	10	36	0
	0	0	1	0	0	38	0	0	1	0	0
	8	7	0	37	0	2	0	0	0	34	0
	1	2	0	63	0	0	103	81	0	1	12
	0	1	0	2	8	0	0	0	0	27	1
	0	0	0	0	0	0	1	70	0	0	16
	0										

Fig 7 : Topic Count over users

MP-LDA when applied to the dataset the following result were observed. The dataset given as input to MP-LDA was filtered on the basis of retweet count, comment count and likes. The list of hashtag was prepared on the basis of the above filter. From the total 15 lacs messages 5 lacs messages were filtered out, so there were 10 lacs messages which had the retweet count, comment count and likes above the threshold. The users count after filtration was 50,000. There were around 30,000 users which were in passive mode in the network and tweeted or retweeted rarely. So we applied MP-LDA on this dataset of 50,000 users and 10 lacs messages. Hashtag related to top 3 topics are were as shown below in Table 5

Table 5. TOP hashtags related to top 3 topics

# flood	**# health**	**# sports**
Flashfloodwarning	Womenshealth	Asiangames
Kerala	Rohingya	USopen
Gordan	Behavior	bluejays
Relief	Organic	Berksgameday
Rescue	Mentalhealth	Football
uttarpradesh	Yoga	Cleveland
Sriabhiyan	medicine	Donald
Builttoserve	Weightloss	Baseball
Japan	Addiction	Serena
Disaster	Exercise	manjitshingh

If we compare the result obtained from LDA we observe that words like Japan and Gordon do not occur top words whereas MP-LDA shows those words in the list of words for hashtag #Flood. Also for #health we observe that words like yoga do not occur in the top words by LDA as compared to MP-LDA. The result obtained were as per the events that occurred during the period when tweets were captured. In India Kerala faced severe floods and the top words captured were in relation to the event like floods, relief, and rescue. There were floods in Japan also during the same period. These words were not seen as top words when applying LDA and T-LDA. Thus MP-LDA gives a more accurate top words list as compared to LDA and Twitter LDA.

5. CONCLUSION

The experimentation shows that our proposed approach outperforms the existing methods. The consideration of social attributes of the users give better accuracy as compared to traditional LDA and twitter –LDA model.
The perplexity values of MP-LDA is much lower as compared to other two models. Thus MP-LDA models better than the traditional methods.

Future Scope: This work can be extended by consideration of location of the users in the network. Also if the follower's data is considered it may lead to more accurate and personalized results.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation, The Journal of Machine Learning Res. 3 (March 2003), 993-1022.

[2] L. AlSumait, D. Barbará and C. Domeniconi, "On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking," *2008 Eighth IEEE International Conference on Data Mining*, Pisa, 2008, pp. 3-12.

[3] Rachit Arora, Balaraman Ravindran, Latent Dirichlet Allocation and Singular Value Decomposition based Multi-Document Summarization 2011.

[4] Liu T., Zhang N.L., Chen P. (2014) Hierarchical Latent Tree Analysis for Topic Detection. In: Calders T., Esposito F., Hüllermeier E., Meo R. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2014. Lecture Notes in Computer Science, vol 8725. Springer, Berlin, Heidelberg

[5] G. Liu, X. Xu, Y. Zhu and L. Li, "An Improved Latent Dirichlet Allocation Model for Hot Topic Extraction," *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*, Sydney, NSW, 2014, pp. 470-476.

[6] Cordeiro, Mario, Twitter event detection: Combining wavelet analysis and topic inference summarization. In Doctoral Symposium on Informatics Engineering, DSIE, 2012.

[7] M. Cataldi, L. Di Caro, and C. Schifanella,," Emerging topic detection on twitter based on temporal and social terms evaluation", AI in Proceedings of the Tenth International

Workshop on Multimedia Data Mining. ACM, 2010, .

[8] Chenyi Zhang and Jianling Sun. 2012. Large scale microblog mining using distributed MB-LDA. In Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion). ACM, New York, NY, USA, 1035-1042. DOI: https://doi.org/10.1145/2187980.2188238

[9] Wu C., Wu B., Wang B. (2016) Event Evolution Model Based on Random Walk Model with Hot Topic Extraction. In: Li J., Li X., Wang S., Li J., Sheng Q. (eds) Advanced Data Mining and Applications. ADMA 2016. Lecture Notes in Computer Science, vol 10086. Springer, Cham

[10] Shiva Prasad Kasiviswanathan, Prem Melville, Arindam Banerjee, and Vikas Sindhwani. 2011. Emerging topic detection using dictionary learning. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (CIKM '11), Bettina Berendt, Arjen de Vries, Wenfei Fan, Craig Macdonald, Iadh Ounis, and Ian Ruthven (Eds.). ACM, New York, NY, USA, 745-754.

[11] Brown Peter F. et al "An Estimate of an upper bound for the entropy of English" Computational Linguistics, March 1992