# An Event detection method in social media based on LDA and K-mean algorithm

**[1]Zarana Patel, [2]Rachana Modi**

Assistant Professor
Department of Computer Engineering
Ganpat University
Mehsana, Gujarat, India

***Abstract*- As an Event constituting a new stage in a changing situation like social media today the short text generated at large scale through those medium for convenient communication between people all over the world. Topic model like LDA have huge success to find latent pattern from large amount of documents, but it is challenging to apply it directly on short text like tweets on twitter micro blog. Due to uncertainty of term distributed over topic the cross topics discovered through LDA can be further cluster using K-mean algorithm to extract accurate event from twitter. Further discovery of topics k-mean is integrated with LDA topic model. At the end we conclude that LDA with K-mean was discovered good topics than LDA alone.**

***Keywords:*** **Topic modeling, Event detection, Twitter, LDA, Data mining, K-mean.**

## I. INTRODUCTION

Twitter is one of the most popular micro blog services in the world. There are 500 million tweets generated per day and around 200 billion per year as of 2018.Those millions of tweets generated everyday tells important stories have been taken place. Those stories abbreviated as Event. An Event is an activity or action occurred with a clear finite duration in which the target entity play a key role [1].

In contrast to conventional media, event detection from Twitter streams poses new challenges. Twitter streams contain large amounts of meaningless messages and polluted content, which negatively affect the extraction performance. In addition, traditional text mining techniques are not suitable, because of the short length of tweets, the large number of spelling and grammatical errors, abbreviated words, and the frequent use of informal, unstructured and mixed language.

Topic Modeling is one of the most powerful techniques in text mining for data mining, latent pattern discovery and finding coherence relation between data and text documents. Topic modeling methods are Latent Semantic Analysis(LSA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA).In which LDA is one of the most popular method in this field [2].

As millions of tweets created every day become more demanding to extract meaningful or major events like disaster, Entertainment, sports, festivals. For human being it's not challenging to figure out heading of news article but for computers we need to teach them to better understand the same topics. Here extracted event has large application like tracking major events, detailed information of event can be found using topic modeling techniques, those rare information not only use to track particular event but also useful to make application for news agency. As there are multiple languages used for event by different user natural language processing on text content like twitter become a taxing task.

## II. Related work

### Event detection from Twitter alleviating data sparsity

Recently probabilistic topic model like PLSA and LDA gain considerable attention in machine learning[3,5].As many variants of Topic Model proposed the basic idea behind virtually model multinomial distribution of words i.e. a unigram language model. The continues growth of information technology increase, Organize and analysis large collection of data become challenging. Topic model have great success over text classification in large document like corpus of newsgroup[6].As the way user gain information is changed through web and social media like twitter.

Topic model can now use mostly over mining topics from Twitter. There are variants of topic model methods. Jianxin Li et al. proposed busty event detection (BEE) Model i.e. an incremental topic model to detect busty events online. They proposed this method to overcome problem of traditional topic model method of LDA and PLSA for short text[7].Xiang Sun et al. used plsa topic model to cluster similar post in twitter and extract hot topics and claim that BEE model doesn't specify relationship between user and their post[8] Lei-lei Shi et al. focus on key post related to event and automatically discover no. of topics and related key post from large no. of post. Restriction over their proposed method is only those user who published, retweet or comment upon post are included in dataset[9] Liang Jiang et al. proposed HEE model that not only consider user interest but also solve data sparsity problem due to short length of post .Topic clustering are done to cluster similar short text post then topic of each document are model by LDA algorithm[10].

### LDA method and its variants

Based on the study of the researchers LDA is the best topic model for event extraction [3,14].But due to problem of short text and data sparseness as discussed in above section other variants are also introduced. Rishabh Mehrotra et al. improved LDA topic

model without modifying underlying mechanism of LDA by hash tag pooling and temporal Pooling to make large document of tweets contain similar hash tag and merge tweets come in short period of time[11].They shows that hash tag based pooling outperform all other pooling Schema to aggregate tweets. So these approach tend to improve upon removing topic model on unpooled technique but there underlying assumption about topic consistency within user and hash tag are frequently violated.

David Alvarez-Melis et al. proposed new pooling technique in which they group tweet according in same user to user Conversation and show that this pooling technique outperforms hash tag based and temporal Based pooling[12].Pranav Surietet al. compares NMF with LDA and concludes that semantically generated by LDA is more meaningful than NMF[13]. Wayne Xin Zhaoet al. proposed TwitterLDA to overcome short text problem of tweets by consider one topic related to one post[14]. Muthukkaruppan Annamalaiet al. proposed clusLDA which combine clustering with LDA but their experiment indicates clustering not necessarily improve content quality of short text[15].

Another pooling process to overcome sparsity problem of twitter using LDA proposed by Malek Hajjem et al. combining information retrieval techniques to LDA. They expand original tweet(i.e query)in order to enhance the effectiveness of IR task [16]. xing et al. proposed MGe-LDA method which add hash tag layer between document and topic layer. Hashtag associated with multinomial distribution over topic and topic represent multinomial distribution over word[17].Marina Sokolova et al. apply LDA topic model on large twitter dataset and show the performance of LDA was affected by changing in distribution parameters α and β. They conclude that LDA perform very well in large dataset and also detect the rare events[18].The another recent work done for extract major life event from twitter using LDA and naive byes classification to give score to the tweets. They apply scoring function to extract time and location information of tweets[19].A. Fathan et al. apply topic modeling like LDA on tweet to get information about football match topic like pre-match information and updated information about match in Indonesia[20].

## III. LDA topic model and K-mean algorithm

### LDA topic model

LDA, an unsupervised generative probabilistic method for modeling a large amount of text data, is used most commonly in topic modeling method, represented as graphical model for finding topic was proposed by David Blie, Andrew Ng and Michael I. Jordan[3].

LDA work on large text documents D. LDA models *D* according to the following generative process :

(i)Choose a multinomial distribution β for topic t (t ∈{1,..., T}) from a dirichlet distribution with parameter $\eta$.

(ii)Choose a multinomial distribution θ for document d (d ∈{1,..., M}) from dirichlet distribution with parameter α.

(iii)For a word $w_n$ (n ∈{1,..., $N_d$ }) in document d,

(a)Select a topic $z_n$ from $\theta_d$ .

(b)Select a word $w_n$ from $\eta_{zn}$ .

This joint distribution between observed and latent variable defines a posterior $(\theta, z, /w)$[3].

$$p(\beta,\theta,z/w) = \left(\prod_{i=1}^{k} p(\beta_i|\,\eta)\right)\left(\prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_{d,n}|\theta_d)\,p(w_{d,n}|\beta_{1:k}, z_{d,n})\right) \quad (1)$$

### LDA interfaced by gibbs sampling

In generative process of LDA, words in documents are the only observed variables while others are latent variables ($\varphi$ and $\theta$) and hyper parameters ($\alpha$ and $\beta$).

Here in statistical method all those nice structure we are assumed and not observed so interface required to infer those latent variable, here we are using gibbs sampling with LDA topic model, It is a Monte Carlo Markov-chain algorithm, powerful technique in statistical interface, and a method of generating a sample from a joint distribution when only conditional distribution of each variable can be efficiently computed.

For every document d go through each word w, Reassign a new topic to w, where we choose topic t with the probability of word w given topic t multiply by probability of topic t given document d, denoted by the following mathematical notation:

$$P(Z_i = j|z_{-i}, w_i, d_i) = \frac{C_{w_i j}^{WT}+\eta}{\sum_{w=1}^{W} C_{wj}^{WT}+W\eta} \times \frac{C_{d_i j}^{DT}+\alpha}{\sum_{t=1}^{T} C_{d_i t}^{DT}+T\alpha} \quad (2)$$

### Clustering algorithm

After mining topic from LDA topic model we used K-mean algorithm for allivating cross topic discovery and finding Hot topics from Twitter. K-mean clustering is fastest and simplest among all other clustering algorithm. It required prior knowledge of topics and then divide documents into no. of cluster prior to topics. The basic idea of k-mean on LDA topic model is as follow:

**Algorithm 1 : K-mean algorithm based on LDA**

**Input :** The no. of cluster K, Topic set obtained by LDA T.

**Output** : Cluster set D.

**Steps :**

1) Chose K data point from as cluster center from D.
2) Distribute every data point to the nearest center x.
3) Calculate mean value of topics to update center of cluster.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2};$$ (3)

4) Until the topic reach convergence.

$$\lim \sum (x_i - x)^2 = 0.$$ (4)

When the cluster is neatly packed together and easily distingue between cluster the k-mean algorithm obtain better cluster result, And as we got the topic from LDA is obvious compact the integration of k-mean with LDA should discover good quality of topics.

## IV. Experiments

### Dataset

we collected tweets of all the major events happened in India in year of 2018 through twitter rest api call from date 04/08/2018 to date 19/12/2018. Here we used topic filter to get more relevant tweets of event. For that we used search term for ex. Kerala, flood, keralaflood, #keralaflood. After collecting tweets the duplicate tweets are removed and Finally collect total 1022 tweets ,the no. of tweet collected per event shows in following fig 5.1.

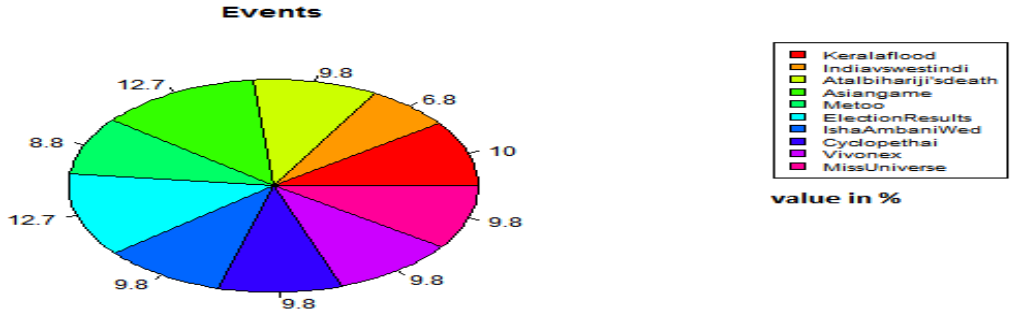


fig 1. Dataset collection

### Result of LDA and k-mean

After preprocessing of raw tweets including tokenization, removal of stop word and stemming the LDA topic discovered 10 topics. Here topic no. above 10 was founded duplication of topic so we took top 10 topic which is unique. Here our benchmark is the top 10 word obtain by LDA topic model.

We can easily compare the top5 word of top 3 topic discovered by LDA in Table1 and LDA+k-mean in Table 2.

**Table 1.**Top 5 word drawn from top 3 topic through LDA

| Top 5 | Topic1 | Topic2 | Topic3 |
|---|---|---|---|
| 1 | Vivonex | Ishaambaniw | Metoo |
| 2 | Vivo | Indian | Today |
| 3 | Dual | Time | Done |
| 4 | Display | Won | Record |
| 5 | Nex | second | Serv |

**Table 2.**Top 5 word drawn from top 3 topic through LDA+k-mean

| Top 5 | Topic1 | Topic2 | Topic3 |
|---|---|---|---|
| 1 | Elect | India | Asiangame |
| 2 | Result | Win | Women |
| 3 | Bjp | Indiavswestindi | Medal |
| 4 | Electionresult | Team | Live |
| 5 | Congress | gold | Update |

### Evaluation of clustering method

Clustering with LDA topic model ensure good quality of topics and avoid crossed topics. After applying LDA topic model on dataset we cluster the words given by term-topic matrix with the help of WEKA Tool. We know precision and recall can reflect the accuracy, degree of check. If we want the related topic, we may pursue precision. On the contrary, we request recall when we expect discovered topic be more. In this paper, we adopt *F*(*F*-measure) to measure the result of cluster, which can combine precision with recall. We can measure it through following equation.

$$P(i,j) = \frac{\text{the number of class } i \text{ in cluster } j}{\text{the number of documents in cluster } j}, \quad (5)$$

$$R(i,j) = \frac{\text{the number of class } i \text{ in cluster } j}{\text{the whole number of documents with class } i}, \quad (6)$$

$$F(i) = \frac{2PR}{P+R}, \quad (7)$$

**Table 3.** The Precision, Recall and F-measure of k-mean clustering with different no. of cluster when no. of topics is 10.

| cluster | Precision | Recall | F- measure |
|---|---|---|---|
| 1 | 0.015 | 0.015 | 0.154 |
| 2 | 0.320 | 0.103 | 0.488 |
| 3 | 0.924 | 0.839 | 0.879 |
| 4 | 0.074 | 0.1 | 0.085 |
| 5 | 0.765 | 0.780 | 0.772 |

When we get 10 topics from LDA, we set 1 to 10 cluster to get P,R and F. From fig. 2 we knew that the F for less than 10 topic is perfect and we got highest F-value in 3,4,6 and 10 in decreasing order. Through analysis, we consider the ambiguity of topic may affect the *F*-measure of clustering. The ambiguity makes the topic belong more clusters.

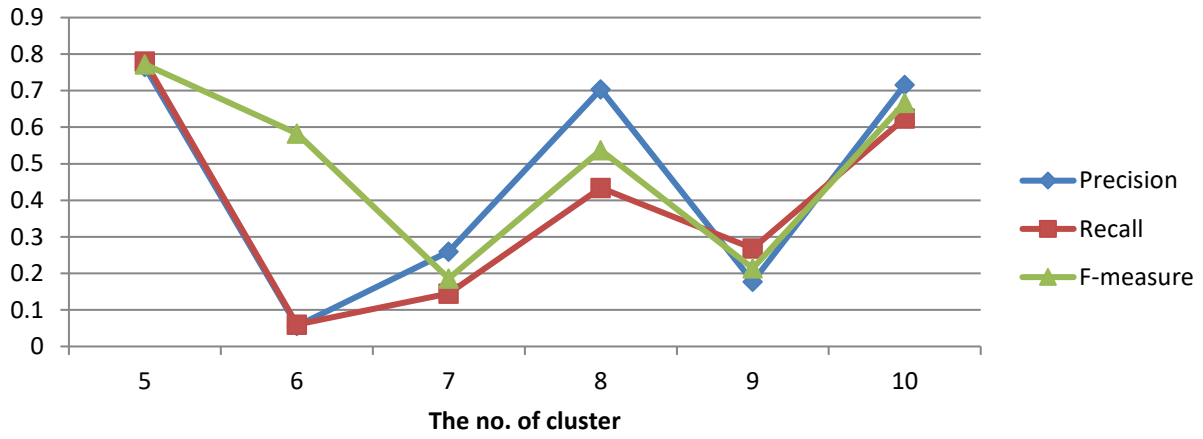


Fig 2. Effect of cluster no. for 10 topics

**V. Conclusion and Future scope**

Topic discovery from micro blog are hot topic because of the topic discovered are too great to calculate. Compare to normal text mining, mining topics from social media are suffered from data sparsity problem. On the other hand selection of proper algorithm is considered and compared. In this paper we proposed combination of LDA and k-mean algorithm to get good quality of cluster and to alleviate the word scarcity problem of topics in micro blogs.

**REFERENCES:**

[1] Dharini Ramachandran, Parvathi R, "Event detection from twitter - a survey Article information : About Emerald www.emeraldinsight.com Event detection from twitter - a survey," 2018.

[2] Jelodar, Hamed, et al. "Latent Dirchlet Allocation (LDA) and Topic modeling: models, applications, a survey." *arXiv preprint arXiv:1711.04305* (2017).

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 993–1022, 2003.

[4] A. Karandikar, "Clustering short status messages : A topic model based approach," *Work*, p. 55, 2010.

[5] T. Hofmann, "Probabilistic latent semantic analysis," *UAI'99 Proc. Fifteenth Conf. Uncertain. Artif. Intell.*, pp. 289–296, 1999.

[6] K. Krishnamurthi, "Impact of Topic Modelling Methods and Text Classification Techniques in Text Mining : a Survey," no. 3, pp. 72–77, 2017.

[7] J. Li, Z. Tai, R. Zhang, W. Yu, and L. Liu, "Online Bursty Event Detection from Microblog," 2014.

[8] X. Sun, Y. Wu, L. Liu, and J. Panneerselvam, "Efficient Event Detection in Social Media Data Streams," *2015 IEEE Int. Conf. Comput. Inf. Technol. Ubiquitous Comput. Commun. Dependable, Auton. Secur. Comput. Pervasive Intell. Comput.*, pp. 1711–1717, 2015.

[9] L. Shi, L. Liu, X. Wu, L. Jiang, and Y. Sun, "Event Detection and Key Posts Discovering in Social Media Data Streams," *IEEE Access*, vol. 3536, no. c, pp. 1–1, 2017.
[10] L. Shi, L. Liu, Y. Wu, L. Jiang, and J. Hardy, "Event Detection and User Interest Discovering in Social Media Data Streams," *IEEE Access*, vol. 3536, no. c, pp. 1–1, 2017.
[11] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA topic models for microblogs via tweet pooling and automatic labeling," *Proc. 36th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '13*, p. 889, 2013.
[12] D. Alvarez-Melis and M. Saveski, "Topic Modeling in Twitter: Aggregating Tweets by Conversations," *$Icwsm16*, no. Icwsm, pp. 519–522, 2016.
[13] P. Suri and N. R. Roy, "Comparison between LDA & NMF for event-detection from large text stream data," *3rd IEEE Int. Conf.*, pp. 1–5, 2017.
[14] W. X. Zhao, J. Jiang, J. Weng, J. He, and E. Lim, "Comparing Twitter and Traditional Media Using," pp. 338–349, 2011.
[15] M. Annamalai and S. Farah Nasehah Mukhlis, "Content quality of clustered latent dirichlet allocation short summaries," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8870, pp. 494–504, 2014.
[16] M. Hajjem and C. Latiri, "ScienceDirect Combining IR and LDA Topic Modeling for Filtering Microblogs Microblogs," *Procedia Comput. Sci.*, vol. 112, pp. 761–770, 2017.
[17] C. Xing, Y. Wang, J. Liu, Y. Huang, and W. Ma, "Hashtag-Based Sub-Event Discovery Using Mutually Generative LDA in Twitter," *Aaai*, pp. 2666–2672, 2016.
[18] M. Sokolova *et al.*, "Topic Modelling and Event Identification from Twitter Textual Data," 2016.
[19] M. Gupta and P. Gupta, "Research and implementation of event extraction from twitter using LDA and scoring function," *Int. J. Inf. Technol.*, 2018.
[20] Hidayatullah, Ahmad Fathan, et al. "Twitter Topic Modeling on Football News." *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2018.