Flock The Similar Users Of Twitter By Using Latent Dirichlet Allocation

B.Srinivasan, Dr.K.Mohan Kumar

Abstract: Online Social Networks(OSN) are becoming the essential needs of everyday life. The users of OSN share their interest in different areas without knowing the topic at every second. This paper forms a model to find the hidden topics on user's posts by using Latent Dirichlet Allocation (LDA) and flocks similar users whose topics are the same.

Index Terms: Flocking, Latent Dirichlet Allocation, LDA, Online Social Networks, Social Media Mining, similar users, Topic Modeling, Twitter.

1. INTRODUCTION

Online Social Networks(OSN) are becoming the essential needs of everyday life. The users of OSN share their interest in different areas without knowing the topic at every second. All social networks allow their users to share their interests by means of posts, comments or replying processes. Maximum the users are using texts to produce their own contents towards an interesting area. The OSN also allows the users to update their activities and knowledge related to the said interests[1]. The posts created by the users contain a number of attributes like feelings, opinions, sentiments related to a topic. When a conversation is started by a user, the others are joining by saying their thoughts towards the created posts. The shared contents are reflecting the inner thoughts of users and produce major impacts around the environment in which the users are involving[2]. The argument grows in proportion with time and contains a lot of hidden knowledge like the topic in which the users are discussing, the number of supports and the number of opposes. This kind of activity leads to form a hidden dynamic community by means of posts towards a topic and involves the people who are having an interest in the same topic. The formed community exists until there is a number of discussions by the users. There are different kinds of social networks like microblogging, podcasts, wikis, etc. Micro-blogging networks are suitable platforms that provide a better way for users to express their thoughts[3]. Examples of trending micro-blogging social networks are Facebook, Twitter, and LinkedIn, etc. A micro-blogged social network provides all kinds of social media like text, image, audio, and video to the users for generating their content. Moreover, the content produced by the users is in an unstructured form. The users of any such social network are also coming from different parts of the world without knowing others in advance and tend to produce the same thought towards the same topic. The important thing to be considered on any OSN is that users themselves are involving in a discussion by considering the posts of others and the social networks are not showing anything about the said topic explicitly. After the arrival of smartphones, the role of OSN is extended with location-based services[4]. All the leading OSN are allowing the users to posts their thoughts with the location. In general, a social network represents a huge group of activities containing the posts created or shared by the users, comments of users with the support of web and mobile technologies. The generation of contents regarding interest with the hidden topic by the users of OSN leads a tremendous opportunity for newer opening in the field of social network research. A newer dimension of mining known as Social Media Mining(SMM) is applied in the field of OSN to mine the unstructured data produced by the users. SMM is used for doing sentimental and opinion mining, online advertising and for doing recommender systems[5]. SMM uses a mixture of traditional data mining and machine learning techniques with statistical methods for extracting knowledge from the social network pages. The present work constructs a model for identifying the hidden topics in the user-generated tweets from Twitter by using the Latent Dirichlet Allocation (LDA) to flock the similar users who produced content with the same topic.

- *B. Srinivasan is a part-time research scholar pursuing his Ph.D. at Rajah Serfoji Government College, Thanjavur, Affiliated to Bharathidasan University, Tiruchirappalli-620024, Tamil Nadu, India. PH-+919894646669, E-mail: prof.bspapers@gmail.com*
- *Dr.K. Mohan Kumar is working as Assistant Professor & Head in the PG and Research Department of Computer Science at Rajah Serfoji Government College, Thanjavur, Affiliated to Bharathidasan University, Tiruchirappalli-620024, Tamil Nadu, India.PH-+919443805042. E-mail: tnjmohankumar@gmail.com*

2 BACKGROUNDS

Twitter is a popular microblogging network having an approximation of 313 million users and an average of 500 million posts every day[6]. The network allows the users to share their interests through a short descriptive post known as a tweet. A tweet is a maximum of 140 characters long comprising plain texts, blank spaces, URL, user names and hashtags. The users of Twitter are producing their thoughts as tweets at the speed of thought and each word provided by every user is treated as a real-world sensor[7]. Twitter is the unique platform for the rapid sharing of posts to any anonymous users in the world. Any kind of post is shared by using Twitter and can be easily identified by all users within a short period of time.

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation(LDA) is an unsupervised method used for detecting the topics across the given data. LDA is the first one, which presented a graphical representation for topic discovery by David Blei et.al in 2002[8][21]. The posts generated by the users of OSN containing unstructured data and an exact model of analyzing and finding the hidden topic is needed for efficient mining process. LDA is suitable for detecting the hidden topics and uses a generative model to mimic the writing process of humans for the generation of topics. The generative model is not present in other topic detection techniques like Latent Semantic Allocation(LSA) etc.

In general, the LDA represents each document is a mixture of topics and each topic is a discrete arrangement of words.[9]. The LDA method is diagrammatically represented in Fig.1

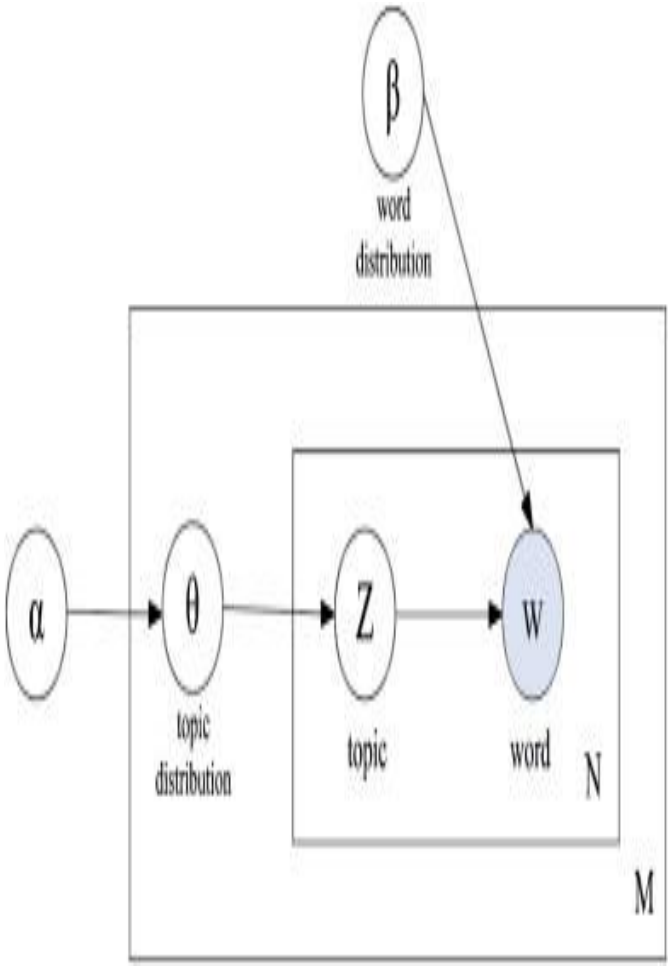

Fig.1 *Plate diagram of LDA topic detection process*

The model shown in Fig.1 is known as the Plate diagram, representing the repeated process of LDA to detect topics with M documents with N-words.

The parameters in the Fig.1 are defined as follows

- α - hyperparameter represents the topic density to be discovered.
- θ - matrix representing the topic distribution
- Z – topic assignment selected from θ
- W – found words in the selected topic.
- β – hyperparameter represents the word density
- M - M number of documents treated as plates
- N - N number of words treated as plates in M documents

The first stage of LDA is the initialization of hyperparameters α and β. Both are values used for tuning the topic detection across the given documents. When higher values are given to α and β, there are chances for finding more topics. Suppose α and β are having low values, few topics are found by LDA. After that, the topic distribution matrix θ is formed from the initialization parameters. The topic Z is selected from θ and is confirmed by using detected word W. The found word W is shown in gray color on Fig.1 and it represents, the W is the confirmed value and others are assumed variables. Today the field of OSN uses LDA for finding the roles, detecting the emotions, grading of essays, Product reviews and in sentimental mining, etc. The LDA process can also be extended by adjusting the parameters and the method of topic formation. Based on the needs, either the base model or extended version of LDA is used for topic detection. The present work uses base LDA for finding the hidden topics in the user's tweets and to flock the users who produced the same topics.

3 RELATED WORKS

A broader range of works have been carried out in the field of topic modeling by using LDA and different techniques have been adopted for flocking the users with the same interests. Tank et.al proposed a topic model that defines a semantic structure for improving the detection of relevant documents based on the terms found in queries[10]. Ramage et.al presented a technique of random projection as a way of speeding up the LSI method[11]. Vala Ali et.al developed a method for detecting topics in the domain of aviation and airport management [12]. Luca Maria Aiello et.al formulated novel algorithms for detecting topics in six areas by using LDA as the base method[13]. Zhenzong li et.al used the LDA model to classify the news from different websites[14]. Hoffman et.al constructed a new model of Probabilistic Semantic Indexing (PLSI) to deal with the domain-specific synonymy with the polysemous words[15]. Wold et.al compared four types of methods for topic modeling to show tweets as part of breaking news detection systems [16]. Tran et.al extended the LDA approach as Spatial Latent Dirichlet Allocation(SLDA) to detect region-based user posting by considering the locations and not the link. After the detection of regions, then the work detected topics based on the found regions[17]. FANG Ying et.al designed a self-adaptive topic model, utilized two entities namely, a location or a person to detect the topics modeling[18]. Shougo Kaneishi et.al developed a method to detect the author's topic for word sense disambiguation[19]. Hong, L. and B.D. Davison conducted an empirical study by using various text model detection techniques including LDA to find topics on user's tweets on Twitter[20]. In all the discussed related works, the LDA method is used as part of main work. After finding the topics by using LDA, the works apply the found topic for further processing related to a specific area.

4 METHODOLOGY

The present work used a Twitter data set containing around 10 million of trending Indian news under various hot topics. A sample of posts from the selected dataset is shown in Fig.2

	news
2684323	First coronary bypass surgery performed in gov...
1923587	Experts blame it on Raj Thackeray's ego; lack ...
326715	9;000 Turkmen prisoners to be released
2002202	Fire dept declares Empress Mall unsafe
1927845	Govt plans 1 lakh acre land bank for afforesta...
571852	Reinstatement of officials opposed
593434	Ramadugu tahsildar trapped by ACB

Fig.2 *Sample dataset*

The tweets on the selected data set are unstructured containing user's tweets in the form of texts, images, video, spaces, and special characters, etc. The present work tries to unhide the topics from the user's tweets of the above dataset without knowing the domain and then applies the topics with the following steps.

Step 1:

The first stage collects the data from the dataset and preprocesses the data by removing unwanted characters like articles, connection words, numbers, spaces, special characters. After cleaning the stage converts all the verbs in the selected sample into its root form by using lemmatization. For example, "running" is lemmatized as "run".

Step 2:

The preprocessed data is then subjected to finding the bigrams and trigramsBigrams are having two words and trigrams are having three words yielding a meaning based on the occurrence of other words. Examples of bigrams are "have to", "to be" and trigrams are " have to attend", "will have to".

Step 3:

After finding the bigrams and trigrams, the words are normalized by removing the meaningless words through normalization.

Step 4:

The last step of the work applies LDA for topic detection among the user's tweets and finds dominant topics from the result of LDA.

Step 5:

The user posts are grouped based on the found dominant topic and after that, the users are flocked towards the same topic group. The formation of a high amount of topic by the LDA is depended on the initial values set to the hyperparameters α and β in step 4.

The overall methodology is depicted in Fig.3.

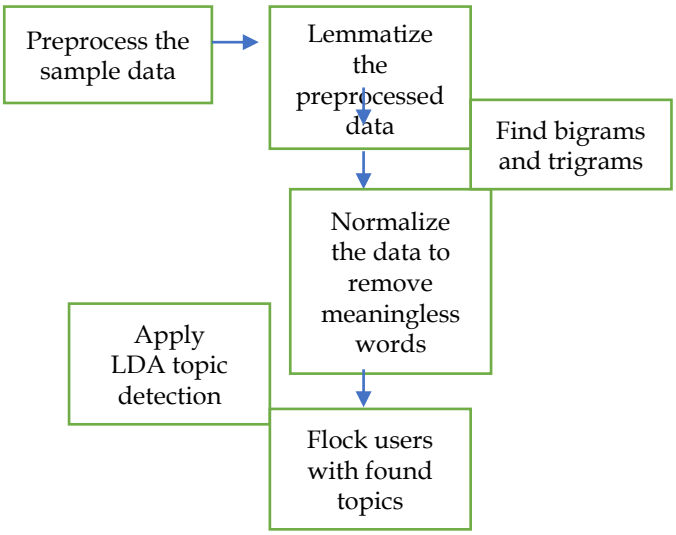

Fig. 3. *The overall methodology*

The LDA topic modeling applied in the present work considers the maximum frequency count for topic detection and allotment from the selected samples.

4 RESULT ANALYSIS

The observations of results after applying LDA for the first two topics topic 0 and topic 1 are shown in Fig.4 and Fig.5.

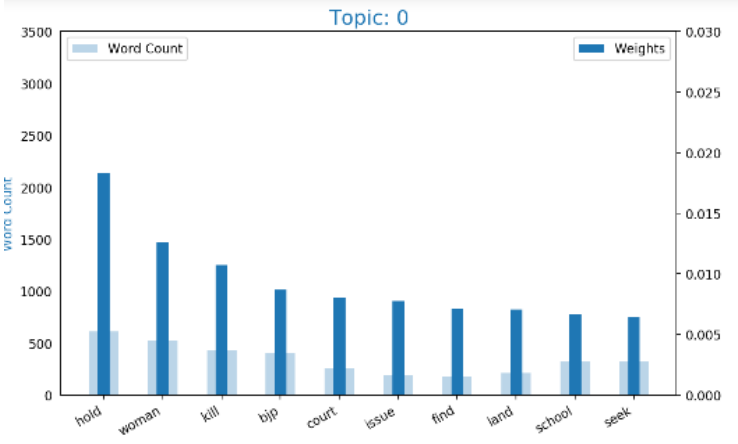

Fig.4 Word *Frequency for Topic 0*

The maximum number of words and the frequency weight of each word in the found Topic.0 is shown in Fig.4 Topic 0 is having "hold" and "woman" are the maximum used words. The left margin in Fig.4 shows the word count and the right margin displays the frequency of a word in percentage on the given topic. Such a convention is very useful for finding the low-frequency word which would not be convenient for a topic and important for another topic by refining the LDA process.

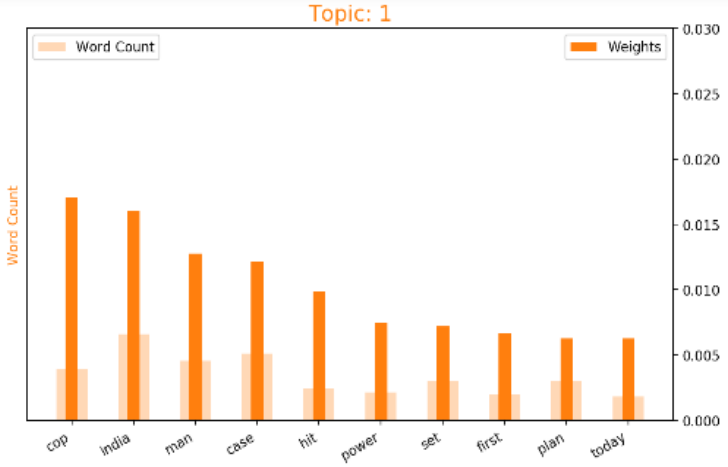

Fig.5 *Word Frequency for Topic 1*

Topic 1 shown in Fig.5 is containing "India" and "cop" are the high-frequency words. Suppose there are chances for occurring of the same word within different topics and the frequency of words gives the importance for the word in a concerned topicThe words with a maximum frequency count on found topics by using LDA is shown in Fig.6.

Fig.6 *Mostly occurred words from two different topics*

Fig.6 shows that words “India” and “Cop” are the high-frequency words in Topic 1 and “Indian”,”Day” and “Delhi” are the maximum words in Topic 2. The number of posts containing the hot topics for the selected 15000 samples is shown as a bar chart in Fig.7. The y-axis is the number of posts and the x-axis is the number of found topics with the maximum frequency of words. The bar chart in Fig.7 shows that Topic 2 is discussed by 8000 user’s tweets. The next dominant topic is Topic 0 discussed by near 7500 user’s tweets. Formation of such Topics by the LDA is giving a clear idea for the flocking of users with same posts.

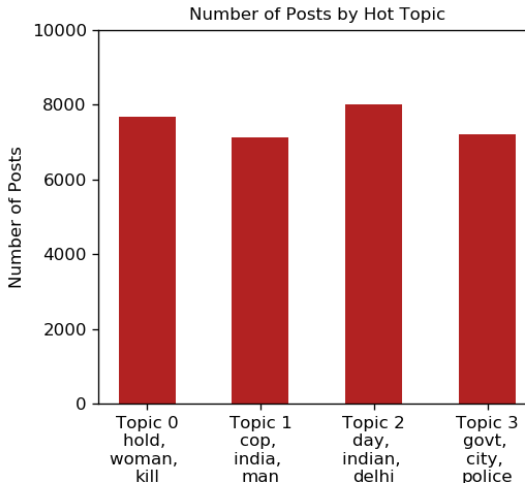

Fig.7 *Number of posts by hot topic*

Fig.7 shows that Topic 2 related to a problem in Delhi is discussed by around 8000 users and Topic 0 related to murder is discussed next by the maximum number of users. The topic-wise distribution of posts from the sample data set is shown in Fig.8 and containing the keywords used by the LDA for selecting the topic.

Keywords	Text
kill, university, give, youth, year_old, offer...	[gym, pic, varun, dhawan, give, fitness, goal]
police, panel, force, teacher, join, play, die...	[girl, play, dupatta, die, freak, accident, ta...
hospital, uk, nod, goa, date, patient, music, ...	[ntca, admit, tadoba, violate, carry, capacity]
school, indian, life, modi, brand, player, cro...	[rape, survivor, seek, life, sentence, dera, h...
new, cop, rape, victim, air, eye, charge, crim...	[nabbed, kidnapping, rape, minor, shivamogga]
delhi, centre, good, fund, parking, fail, sche...	[bride, prejudice, fail, woo, ambarsaris]
not, tell, home, cm, do, clear, due, month, ja...	[rise, byrne, have, not, leave, home, month]
school, indian, life, modi, brand, player, cro...	[forget, jab, tiny, current, deliver, anaesthe...
hold, first, pay, exam, pak, celebrate, staff,...	[bmw, unveil, first, electric, car, today]
launch, win, point, south, farmer, village, re...	[baap, beta, jodi]
say, set, party, cong, be, firm, part, wife, r...	[cbot, corn, close, firm, late, buying]

Fig.8 *Keywords used to select topics*

Sentence wise coloring of posts after the selection of topics by using LDA is shown in Fig.9. The word with the same color as belonging to the same topic and the maximum frequency of words with similar color determines the topic of a post.

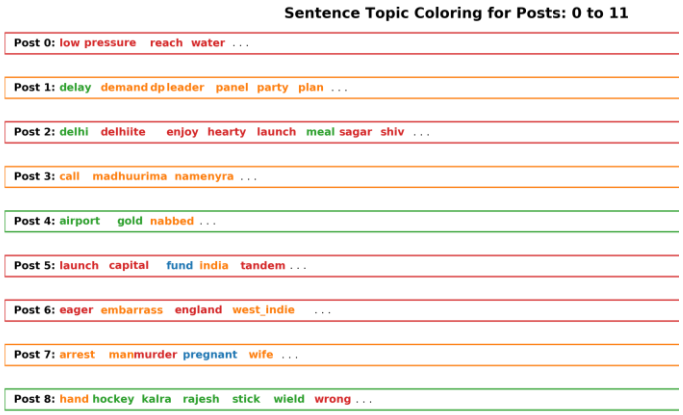

Fig.9 *Sentence wise coloring*

The detailed discussion of results shows that the topics are well modeled by using the LDA by counting the word frequency on each post of the taken twitter dataset. The LDA process also maintains the semantic meaning of posts for effective topic selection. The performance of the LDA with a different set of samples is measured with perplexity and shown in the following table.

TABLE 1

PERPLEXITY OF LDA

S.No	Samples	Perplexity value
1	500	-8.6113
2	1000	-9.1089
3	2000	-9.6677
4	3000	-10.0869
5	4000	-10.4783
6	5000	-10.7783
7	6000	-11.1203
8	7000	-11.4020
9	8000	-11.7603
10	9000	-11.7874
11	10000	-12.3267
12	15000	-12.7946

Perplexity is a measure that determines the possibilities of the collection of unseen words “W” of a post belonging to a topic. The lower the value of perplexity on each turn, the better the performance of the LDA model. It’s a log-likelihood measure calculated by using the following formula. Perplexity = exp(L(unseen words W)/count of tokens) (1) The equation (1) states that L is the likelihood and is a value divided by the total number of found tokens. Table.1 shows that perplexity values are decreasing with the increasing number of samples which clearly states that the LDA model detects the hidden topics on the user’s tweets well. The perplexity values in Table.1 are plotted in Fig.10.

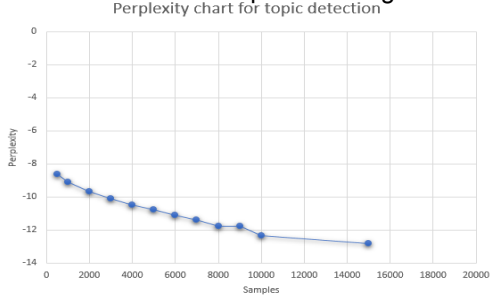

Fig.10. *Perplexity plot for the detected topics*

The drawn line curve in Fig.10 clearly shows that the possibility of detecting new topics on a given post is increasing with a maximum number of samples. When there are more posts to be subjected for topic detection, huge volume of topics is found which helps the flocking of users into correct group by considering their posts. The work also tested the user flocking with new posts as test data. Initially, the found topic from the LDA process is fed with the new input and the related words on new posts are analyzed as like the topic detection process and correctly grouped with the same topic.

5 CONCLUSIONS

The present work used Latent Dirichlet Allocation(LDA) for the detection of topics in the user's posts to the effective flocking of users with similar tastes. The LDA did the role well for the detection of topics. During the topic detection process, there are chances for the occurrence of the minimum frequency of words in the user's posts. The LDA process tries to assign a topic to the minimum frequency words by repeatedly doing the topic detection process The work used tweets with a maximum of 140 characters for topic detection. Suppose a tweet itself is having a countable number of words, the LDA method returns the minimum number of topics. Finding words with minimum frequencies in overall posts, and posts with least number of words before starting the topic detection for same user grouping is the future enhancement of this present work.

REFERENCES

[1] Bhuvaneswari Anbalagan and Dr. Valliyammai, "#ChennaiFloods: Leveraging Human and Machine Learning for Crisis Mapping during Disasters using Social Media," 2016 IEEE 23rd International Conference on High Performance Computing Workshops.

[2] K.Mohankumar and B.Srinivasan, "Formation of Similar Users group by using Support Vector Machine with Facebook Posts" International Journal of Computer Sciences and Engineering Open Access Research Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

[3] Hossein Dolatabadi et.al, " Clustering Users in Micro Blogging Social Networks using Probabilistic Topic Modeling – A Framework",2012 12th International Conference on Computational Science and Its Applications

[4] K. Mohan Kumar and B. Srinivasan, "Point-of-Interest Based Classification of Similar Users by Using Support Vector Machine and Status Homophily", International Journal of Machine Learning and Computing vol. 9, no. 5, pp. 615-620, 2019.

[5] Georgios Lappas et .al, "From Web Mining to Social Multimedia Mining," 2011 International Conference on Advances in Social Networks Analysis and Mining

[6] Doshi et .al, "TweerAnalyzer: Twitter Trend Detection and Visualization," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA).

[7] Sakaki, T., M. Okazaki, and Y. Matsuo "Earthquake shakes Twitter users: real-time event detection by social sensors." 2010. ACM

[8] Zhao, W., et al, "Comparing Twitter and traditional media using topic models" Advances in Information Retrieval, 2011: p. 338-349

[9] Rubayyi Alghamdi and Khalid Alfalqi, "A Survey of Topic Modeling in Text Mining" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 1, 2015

[10] Tang, J. et al, "Understanding the limiting factors of topic modeling via posterior contraction analysis," Proceedings of The 31st International Conference on Machine Learning. 2014.

[11] Ramage et al,"Characterizing Microblogs with Topic Models," ICWSM, 2010

[12] Vala Ali Rohani et al, "Topic Modeling for Social Media Content: A Practical Approach," 2016 3rd International Conference On Computer And Information Sciences (ICCOINS) 978-1-5090-2549-7/16/$31.00 ©2016 IEEE

[13] Luca Maria Aiello et al., "Sensing trending topics in Twitter", IEEE Transactions on Multimedia Volume: 15 , Issue: 6 , Oct. 2013.

[14] Zhenzong li et.al, "News Text Classification Model Based on Topic Model", 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS).

[15] Hofmann,"Probablistic Latent Semantic Indexing",ACM SIGIR Forum,Vol.51 No.2,July 2017.

[16] Wold et.al," Twitter Topic Modeling for Breaking News Detection",WEBIS1(2016).

[17] Tran Van Canh., et al., " A Spatial LDA Model for Discovering Regional Communities", 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

[18] FANG Ying et al., " Self-Adaptive Topic Model: A Solution to the Problem of "Rich Topics Get Richer", China Communications December 2014

[19] Shougo Kaneishi," Word Sense Disambiguation using Author Topic Model", Takuya Tajima Faculty of Information Engineering, Fukuoka Institute of Technology FIT 3-30-1

[20] Hong, L. and B.D. Davison. "Empirical study of topic modeling in twitter. in Proceedings of the First Workshop on Social Media Analytics". 2010. ACM

[21] Blei, D.M., A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation. the Journal of Machine Learning Research", 2003. 3: p. 993-1022.