International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 7 | ISSN : 2456-3307

Review on an Enhanced LDA Topic Model Approach for Event Extraction from Twitter

Zarana Patel*[1], Jitendra Dhobi[2]

*[1] PG Student, Department of Computer Engineering, Government Engineering College Gandhinagar, Gandhinagar, Gujarat, India

[2]Associate Professor, Department of Computer Engineering, Government Engineering College Gandhinagar, Gandhinagar, Gujarat, India

ABSTRACT

Topic models are powerful tools to identify latent text patterns in the content. They are applied in a wide range of areas including event extraction from Twitter. Twitter, as a popular micro blogging service, has become a new information channel for users to receive and exchange the important information on current events. Tweets recently gain a lot of importance due to its ability of produce information rapidly. Tweets are commonly related to some events. In this paper provide you a review on event extraction from twitter using topic modeling. Based on the study of the researchers LDA is the best topic model for event extraction. Though applying traditional LDA topic model directly on tweets posses two challenges: 1) Data scarceness problem due to the nature of short text length of the tweets. 2) Generated summaries contain words that are somewhat general and independent to the topic that is failed to understand the semantic of twitter data. Event Extraction methods present in this literature address this problem and classify different approach and discuss commonly used features.

Keywords : Topic modeling, Event Extraction, Twitter, LDA, Data mining

I. INTRODUCTION

Twitter is one of the most famous 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 per day tell new stories about real life and real world. Those stories abbreviated as Event. An Event is an activity or action 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 unimportant messages and polluted content, which negatively affect the extraction performance. Also the traditional text mining techniques are not suitable, because of the short length of tweets, the large number of spelling and grammatical errors, and the frequent use of informal, unstructure and mixed language.

Tweets can be extracted from twitter Stream Api or Rest Api in json format which then converted in any structural format.NLP required when work with tweets to remove common and unrelated words. Human created tweets are very noisy included punctuation, emoticons and Unicode characters, removing those definitely increase performance of topic model. Those preprocessing steps on tweets must be done in ordering include remove emoticons ,url ,punctuation then strip white space, convert tweets into lowercase after stem document and remove stop words.

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 methods in this field [2].

As millions of tweets created every day it become challenging 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. This is where Topic Model came into picture .Topic Model is unsupervised classification algorithm of machine learning.

LDA, an unsupervised generative probabilistic statistical method for modeling a corpus, is most commonly used topic modeling method, was first presented as a graphical model for topic discovery by David Blei,Andrew Ng and Michale Jorden in 2003[3].It provide thematic Summary of documents, useful for large dataset and in text mining.

The organization of document is as follow. The section2(Literature Survey),The literature review done on Twitter As a Source of Information, Topic Model Role for Event Extraction from Twitter , LDA method and its variants and Evolution Metrics for LDA .In section 3(Conclusion and future scope) conclusion derived from literature survey.

II. LITERATURE SURVEY

A. Twitter as a Source of information

Twitter has several properties that differentiate it from other information channel like news websites and television. First tweets are created in real time. For example, a tweet related to a tornado might be written one minute after a user witnessed a tornado was formed. This information could be spread even faster than TV broadcasts [4].From so many years the task of event detection from twitter is carried out by researchers.

Events are two type specific and unspecific, Specified event detection are those aims at identifying known social events which are known previously with fully or partially content or metadata information such as location and time. The real time nature of social posts of twitter reflect event that happened as emerging event, breaking news and general topic that attract attention of large number of users are called unspecified events[5].

Several study have aimed at analyzing social data from twitter through performing data mining technique such as supervised classification[6].This technique are more suitable when events are specified and trained on label data.Hovever these techniques could be considered to have limited capabilities because of unpredictable nature of the dataset. Cluster analysis of tweets have been reported particularly suitable for those kind of data for two reason(Go et al., 2009):(1)The amount of training data is too large for manual labeling(2)The nature of data implies unnoticed group of text that carries important information that can only be exposed by unsupervised learning[11].

As K-mean algorithm is fastest than other partitioning based algorithm many researcher used this algorithm for summarization of events in twitter[7,8].Though directly apply those algorithm on categorical data posses challenging due to high dimensional twitter data and tweets contain many misspelling, abbreviation of words which is difficult to accurately classify events, so the performance of k-mean algorithm is decrease when data are not normally distributed or when cluster do not have equal variance[9].Other aggregation filtering and hierarchical clustering use for event detection suffer from fragmentation of topic over cluster[10].

In new growing digital era of world basic text mining technique replace with topic modeling technique in machine learning as a part of data mining. Now in next section we explore the survey about topic modeling role for event extraction from twitter.

B. Topic Model Role for Event Extraction from Twitter

With the passage of time, the importance of Topic modeling in different disciplines will be increase. Topic model is a popular method for modeling term frequency occurrences for documents in a given corpus. A topic basically consists of set of words that co-occur frequently. Unsupervised nature allows topic models to be trained easily on datasets meant for specific domains[12] .

Recently probabilistic topic model like PLSA and LDA gain considerable attention in machine learning[3][13].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 [14].As the way user gain information is change through web and mostly through social media like twitter Topic model can now use mostly over mining topics in twitter .

There are variants of topic model methods. Jianxin Li et al praposed 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 of short text. They assumed one twitter post related to one event[15].Xiang Sun et al. used plsa topic model to cluster similar post in twitter and extract hot topics. HITS (hyper induced topic search) scoring method incorporate to distill high quality post by showing relation between user and its tweets. After EM (expectation Maximization) algorithm is employed to train parameter and obtain estimator to describe hot events. They claim that BEE model doesn't specify relationship between user and their post[16] 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[17] 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[18].

The fundamental reason lies in that conventional topic models implicitly capture the document-level word co-occurrence patterns to reveal topics, and thus suffer from the severe data sparsity in short documents to solve this problem another variant proposed by Xiaohui Yanet al a biterm topic model for short text[19].K-mean clustering algorithm can make topic discriminative when dataset is intensive and the different among topic document is distinct so Weijiang Li*a* et al. integrated k-mean and HC to BTM for further topic discovery[20] Ahmad Hany Hossny et al use SVD with clustering to group related word as enhanced signal for textual feature in tweet in order to improve correlation with tweet[21].

Another novel way proposed by Xiaohui Yan et al. to tackle short text problem of social media hinders existing topic model to find latent topic by using nmf(non negative matrix factorization)[22]. Liangjie

Hong et al proposed extended AT (author topic) model where topic Learn from aggregated message by same user may lead to superior performance in classification problem[23].

C. 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 introduce which we have discuss in this paper. Rishabh Mehrotraet 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[24].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 not topically homogeneous, because 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[25].

As we discussed nmf for solve short text problem in topic model in above section Pranav Surietet al. compares nmf with lda and concludes that semantically generated by lda is more meaningful than nmf[26]. Wayne Xin Zhaoet al. proposed TwitterLDA to overcome short text problem of tweets by consider one topic related to one post[27]. Muthukkaruppan Annamalaiet al. praposed clusLDA which combine clustering with LDA but experiment indicates clustering not necessarily improve content quality of LDA topic model.[28].

Another pooling process to overcome sparsity problem of twitter in 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 .Tweet Expansion add additional terms from external data source. They compare this extension of enriched tweet with Hashtag based aggregation and there proposed method increase the accuracy of basic lda algorithm[29].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[30].Marina Sokolovaet al. apply lda topic model on large twitter dataset and show the performance of lda affect by change in distribution parameters α and β.They also show the assignment of topic to document is poor due to the short length of tweet. They include the Gikomba twitter data include bombing incident, The Mandera Twitter data contains tweets mainly talking about so-called “tribal clashes”, The Makaburi dataset contain information about the violent death of Sheikh Makaburi and The Mpeketoni data set mainly discuss an attack that happened in Mpeketoni town in the coastal region of Kenya. They conclude that LDA perform very well in large dataset and also detect the rare events[31].The another recent work done for extract major life event from twitter using lda and used naive byes classification to give score to the tweets. They apply scoring function to extract time and location information of tweets.[32].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[34].The measurement of proper assignment of event to tweet is challenging task and discussed in next section.

D. Evaluation Metrics

Performance evaluation of different approaches and features is a major issue facing event detection in Twitter[10]. The quantitative and qualitative performance of event detection technique for twitter data is itself the challenging research question[33].In Fig.1 we summarize different evaluation metrics used corresponding to different topic modeling methods for event extraction in twitter.

TABLE 1: COMPARISSION OF TOPIC MODELING METHOD AND ITS VARIANTS WITH EVALUTIONMETRICS

<table>
<tr><th>Author</th><th>Techniques</th><th>Collection</th><th>Evaluation</th></tr>
<tr><td>Rishabh Mehrotra et al.[24].</td><td>Hashtag based pooling+LDA</td><td>20-30%sample of public twitter post ,use term query to retrieve tweet collection.</td><td>Purity=+8.69%
NMI score=+12.8%
PMI score=+127%</td></tr>
<tr><td>Malek Hajjem et al.[29].</td><td>IR + LDA</td><td>Twitter API in French and Arabic language with specific keyword</td><td>Recall=0.790
Precision=0.789
F-measure=0.789
Purity=0.8
NMI=0.2641
PMI=1.68</td></tr>
<tr><td>David Alvarez-Melis et al.[25].</td><td>User to user coversation+LDA</td><td>Select top 25 most influance user for each topic + GNIP Historical API to retrieve public post of this user.</td><td>NMI=7.819
Adj.RandIndex=4.514</td></tr>
<tr><td>Weijiang Li et al.[20].</td><td>BTM+K-mean
BTM+HC</td><td>Sino Microblog API</td><td>F(F-measure)
HC-F=0.837
k-mean-F=0.849</td></tr>
<tr><td>Xiang Sun et al.[16].</td><td>HITS+PLSA+EM</td><td>Sino Microblog API</td><td>precision@K where K=Top post related to topic</td></tr>
<tr><td>Marina Sokolova et al.[31].</td><td>LDA</td><td>4 sets of data collected from Twitter Streaming Api</td><td>NMI=0.606
Coherence= 3.976 ± 0.696</td></tr>
<tr><td>Xiaohui Yan et al.[22].</td><td>NMF</td><td>Twitter data,TREC2011
Title data, contain news title
Question Data,que. crawl from Chinese microblog</td><td><table><tr><th></th><th>NMI</th><th>Purity</th><th>ARI</th></tr><tr><td>Titledata</td><td>0.4-0.5</td><td>0.6-0.8</td><td>0.25-0.3</td></tr><tr><td>Question Data</td><td>0.6</td><td>0.6</td><td>0.5</td></tr></table></td></tr>
</table>

III. CONCLUSTION

In this paper a review on different topic modeling techniques in twitter for event detection has been discussed. First the Twitter as information source is reviewed then the role of topic model in twitter and the lda method and its variants has been also discussed. At last different evaluation metrics for lda topic model is summarize, which help to reader of this paper to select proper topic model method and use of suitable evaluation metric of it. We observed from this survey that most of the work done to improve LDA topic model relay on meta data information like user, time and other feature like hash tag. We conclude from this survey is by setting different value for α and β hyper parameter in lda topic model and applying aggregated tweets on lda improve performance than directly applying basic lda topic model for events extraction from twitter.

IV. 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,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] Yang, Shih-Feng, and Julia Taylor Rayz. "An event detection approach based on Twitter hashtags." *arXiv preprint arXiv:1804.11243*,2018.

[5] F. Zarrinkalam and E. Bagheri, “Event identification in social networks,” Encycl. with Semant. Comput. Robot. Intell., vol. 01, no. 01, p. 1630002, 2017.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors,” Proc. 19th Int. Conf. World Wide Web, pp. 851–860, 2010.

[7] S. B. Kaleel and A. Abhari, “Cluster-discovery of Twitter messages for event detection and trending,” J. Comput. Sci., vol. 6, pp. 47–57, 2015.

[8] D. Godfrey, C. Johns, C. Sadek, and M. L. Aug, “A Case Study in Text Mining : Interpreting Twitter Data From World Cup Tweets,” vol. 5, pp. 1–11.

[9] R. Musters, “Topic detection in microblogs using big data and neural networks,” Univ. Groningen, no. June, 2017.

[10] G. Ifrim, B. Shi, and I. Brigadir, “Event detection in Twitter using aggressive filtering and hierarchical tweet clustering,” CEUR Workshop Proc., vol. 1150, pp. 33–40, 2014.

[11] F. Atefeh and W. Khreich, “A survey of techniques for event detection in Twitter,” Comput. Intell., vol. 31, no. 1, pp. 133–164, 2015.

[12] A. Karandikar, “Clustering short status messages : A topic model based approach,” Work, p. 55, 2010.

[13] T. Hofmann, “Probabilistic latent semantic analysis,” UAI’99 Proc. Fifteenth Conf. Uncertain. Artif. Intell., pp. 289–296, 1999.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] X. Yan, J. Guo, Y. Lan, and X. Cheng, “A

biterm topic model for short texts," WWW '13 Proc. 22nd Int. Conf. World Wide Web, pp. 1445–1456, 2013.

[20] W. Li, Y. Feng, D. Li, and Z. Yu, "Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm," Autom. Control Comput. Sci., vol. 50, no. 4, pp. 271–277, 2016.

[21] A. H. Hossny, T. Moschuo, G. Osborne, L. Mitchell, and N. Lothian, "Enhancing keyword correlation for event detection in social networks using SVD and k-means: Twitter case study," Soc. Netw. Anal. Min., vol. 8, no. 1, p. 0, 2018.

[22] X. Yan, J. Guo, S. Liu, X. Cheng, and Y. Wang, "Learning Topics in Short Texts by Non-negative Matrix Factorization on Term Correlation Matrix," Proc. 2013 SIAM Int. Conf. Data Min., pp. 749–757, 2013.

[23] L. Hong and B. D. Davison, "Empirical Study of Topic Modeling in Twitter," pp. 80–88,2010.

[24] 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.

[25] D. Alvarez-Melis and M. Saveski, "Topic Modeling in Twitter: Aggregating Tweets by Conversations," $Icwsm16, no. Icwsm, pp. 519–522, 2016.

[26] 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.

[27] W. X. Zhao, J. Jiang, J. Weng, J. He, and E. Lim, "Comparing Twitter and Traditional Media Using," pp. 338–349, 2011.

[28] 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.

[29] M. Hajjem and C. Latiri, "ScienceDirect Combining IR IR and and LDA LDA Topic Topic Modeling Modeling for for Filtering Filtering Microblogs Microblogs," Procedia Comput. Sci., vol. 112, pp. 761–770, 2017.

[30] 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.

[31] M. Sokolova et al., "Topic Modelling and Event Identification from Twitter Textual Data," 2016.

[32] M. Gupta and P. Gupta, "Research and implementation of event extraction from twitter using LDA and scoring function," Int. J. Inf. Technol., 2018.

[33] A. Weiler, M. Grossniklaus, and M. H. Scholl, "Evaluation Measures for Event Detection Techniques on Twitter Data Streams," pp. 108–119, 2015.

[34] Hidayatullah, Ahmad Fathan, et al. "Twitter Topic Modeling on Football News." 3rd International Conference on Computer and Communication Systems (ICCCS). IEEE, 2018.