Comparison between LDA & NMF for Event-Detection from Large Text Stream Data

Pranav Suri[1], Nihar Ranjan Roy[2]
[1,2] School of Engineering, G D Goenka University, Gurgaon, India
[1]suripranav1995@gmail.com, [2]nihar.ranjanroy@gdgoenka.ac.in

***Abstract*—Usage of social network for topic identification has become essential when dealing with event detection, especially when the events impact the society. In order to address this task, machine learning algorithms and natural language processing techniques have been extensively used. In this paper, an approach to obtain meaningful data from Twitter has been discussed. Further, Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) have been used in order to detect topics from this textual data obtained from Twitter along with RSS feed of news headlines. The observed results show that both the algorithms perform well in detecting topics from text streams, the results of LDA being more semantically interpretable while NMF being faster of the two.**

Keywords—Topic Modeling; Latent Dirchlet Allocation; Non-Negative Matrix Facorization; Twitter; Natural Language Processing

I. INTRODUCTION

SOCIAL NETWORKS are platform for people to connect with each other on the internet. Social Media platforms are also a huge source of information, reflecting people's opinions, and serve nearly as an instantaneous channel for spreading information [1].

Twitter is a widely used microblogging service. It contains millions of accounts, and its tweets form a huge source of unstructured textual data, which include discussions and opinions on several matters from its users [2]. The popularity and high on-site activity of Twitter has made it a valuable source of information of events of various types and scales [3].

The high connectivity and almost instantaneous responses entails that this social network platform makes this platform as a reliable source of gathering information in real time [2]. However, to extract useful information from the texts that appear on the social networking sites is not an easy task, due to their huge size, speed of their creation & unstructured form [4]. This problem can be tackled by employing automatic procedures implementing advanced Text Processing (TP) algorithms [5, 6] and Machine Learning (ML) [7] techniques.

Event detection from textual data streams generally involves semantic analysis of the text such as topic modeling to find the context of the text under study. The detection of event can be then extended to finding its location, time of occurrence [15].

In natural language processing, a topic is a text mining tool used for discovering semantic structures in a text body. In the age of data, the amount of text to be dealt with is beyond the manual processing limits of humans. Topic models help offer insights to apprehend large streams of unstructured text bodies.

This paper is organized as follows. In the next section, we present the literature reviews of similar works done by other scholars followed by a brief description of Latent Dirichlet Allocation and Non-Negative Matrix Factorization techniques in Section III. Section IV discusses the methodology for the preparation of datasets and topic detection from the collected streams. Section V focuses on the experimental results and analysis followed by the conclusion in Section VI.

II. LITERATURE REVIEW

In the past, several studies related to topic modeling and event detection have been performed. In 2016, Marjori N. M. Klinczak and Celso A. A. Kaestner [16] compared four clustering algorithms – k-means, k-medoids, DBSCAN and NMF, so as to observe the context of tweets obtained from Twitter. The tweets with relevant hashtags were extracted using the Twitter API. The text was cleaned using text pre-processing techniques viz. case folding, stop-words removal & stemming. The mentioned algorithms were applied to a database composed of tweets having hashtags associated with Nepal Earthquake as initial context. To form a contrast between the results obtained by using the mentioned algorithms, quantitative comparison was done using cohesion and separation measures. For empirical comparison, word clouds were used which algorithm provided more human interpretable results. They conclude their work by suggesting that the NMF clustering algorithm provides less complicated clusters that are easier to interpret.

In 2010, Sakaki, Okazaki and Matsuo [15] investigated if a real-time event (such as an earthquake) can be detected by monitoring Twitter activity. For the study, data was extracted from Twitter using query words 'earthquake' & 'shaking'. Classification of tweets was done using a Support Vector Machine (SVM) classifier in which the size of tweet, textual attributes and their context were used. The paper assumes Twitter to be a network of social sensors where tweets are considered to be sensory information. Subsequently, a probabilistic spatiotemporal model was prepared to assess an event occurrence and its location. The model included a quantitative analysis of reliability of tweets and an improved

978-1-5090-6218-8/17/$31.00 ©2017 IEEE

algorithm for location estimation. Using the proposed methodology, they proposed a system that could detect earthquakes with seismic intensity of scale 3 or more with accuracy and send e-mail alerts to registered users. It is claimed that the systems alerts were delivered faster than the announcements that are broadcast by the Japan Meteorological Agency (JMA).

The above research works prove the significance of Twitter in the detection and analysis of real-time events. For an event detection, it is important to find the context of the text stream, hence, we address this challenge of event-detection through our comparison study of the said algorithms. One of the research agendas of our paper is to extend this work for comparison of NMF clustering algorithm to LDA (which was not included in the former study). The primary research question of investigation is: *"Can topic modelers be used for semantic analysis for detection of an arbitrary event in text data-streams? If yes, what are the factors to consider for the choice of the algorithm"*.

III. Algorithm Descriptions

A. *Non-Negative Matrix Factorization*

Non-Negative Matrix factorization or NMF, is a linear-algebraic optimization algorithm. One of its properties is that it can extract meaningful information about topics without prior knowledge of the underlying meaning in the data. The mathematical objective of NMF is to decompose a single '$n \times m$' input matrix into two matrices such that their product is a close estimate to the input matrix. For topic modeling, the input matrix of choice is the document-term matrix. This matrix is factorized into the document-topic matrix, of dimensions '$n \times t$', and a topic-term matrix, of dimensions '$t \times m$', where 't' is the number of topics that are to be produced.

The NMF clustering algorithm has been employed successfully, mainly because it can be adapted to specific application such as natural language processing [13].

B. *Latent Dirichlet Allocation*

Latent Dirichlet Allocation or LDA, is a generative probabilistic model largely used for topic modeling. LDA is a three-level hierarchical Bayesian model, within which every item of a corpus is modeled as a finite mixture over an underlying set of topics. Each topic is then modeled as an infinite mixture over an underlying set of topic probabilities. There topic possibilities offer a specific illustration of a document [14].

LDA represents documents as a mixture of topics that contain words with certain probabilities of occurrence. Given a collection of documents, some fixed number of topics to find, LDA learns the topic representation of each document and therefore the words associated to each topic via an iterative procedure. LDA then tries to backtrack from the documents to seek out a collection of topics that are likely to have generated the collection [12].

IV. Event Detection Procedure

A. *Preparation of the Dataset*

As the first step towards event detection, an approach close to a real time event detection system was develop so as to compare the algorithm in the same scenario as for semantic analysis in an event detection system. To prepare a dataset, a script that connects to the Twitter streaming API in order to extract tweets in JSON format was developed.

To reduce the data cleaning process, initially, the tweets were filtered using only a location filter viz. India (and neighboring regions). Although, it provides a satisfactory performance, without a topic filter it adds a lot of irrelevant tweets which do not represent any real event. Another limitation of adding a location filter lies in the fact that not all tweets have location metadata available, which excludes a lot of relevant tweets.

To tackle the former problem, a topic filter was added. For this, tweets with hashtags (and words) '#news', '#breaking', '#emergency' are selected through the filter. Even though the tweets were a lot more relevant after adding a topic-filter, it significantly reduced the number of tweets being extracted per unit time. This poses a setback to the topic detection models in use because of inadequate data.

To address this issue, a straightforward solution implemented was to add more terms to the topic filter. To find the relevant keywords, a script that returns the trending topics for India was written. The script makes a call to the Twitter API and returns location based trends. These trends are then added to the topic filter. Fig. 1. explains this procedure with the help of a flowchart.

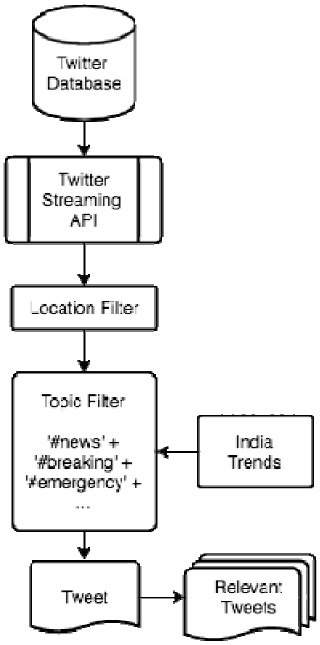


Fig. 1. Procedure to extract and store the relevant tweets from Twitter.

Due to the unreliability of Twitter data, to test the performance of the topic detection models, namely, Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), a cleaner data source was required. To address this problem, RSS news feed from popular news websites was collected over a time period. News RSS feed is a clean source of textual information and also contains text who context are real events. Hence, making it suitable to test the algorithms.

B. *Topic Modeling Procedure*

The topic modeling process is a multi-step procedure for both LDA and NMF. Before applying the algorithms, text processing techniques – tokenization, URL removal, stop-word removal were done to clean the data. The process is then followed by the vectorization of the data to document-term frequency matrix.

In case of LDA, the model is trained over a corpus which in this case is the collection of words in the documents of our textual data. This training of the model can be avoided by using an already trained model. This however results in some reduction in the accuracy of the model. This model is used to backtrack to seek out the topics to be detected.

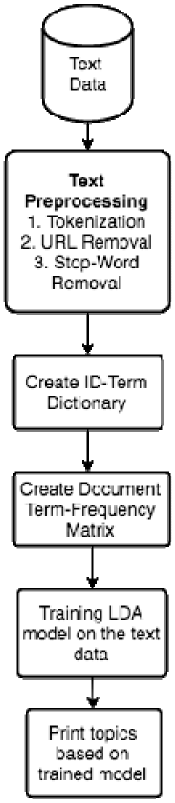

Text Data
Text Preprocessing
1. Tokenization
2. URL Removal
3. Stop-Word Removal
Create ID-Term Dictionary
Create Document Term-Frequency Matrix
Training LDA model on the text data
Print topics based on trained model

Fig. 2. Flowchart representing the steps taken to detect topics using Latent Dirichlet Allocation (LDA).

For NMF, the created document-term frequency is factorized as per the algorithm before printing of the obtained topics.

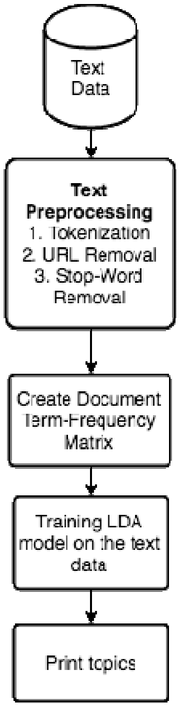


Fig. 3. Flowchart representing the steps taken to detect topics using Non-Negative Matrix Factorization (NMF).

A detailed and in-sequence picture of these processes are depicted in Fig. 2. & Fig. 3.

V. Experimental Analysis & Results

The experiment was carried out on a hardware configuration of – 2.7 GHz Intel Core i5 processor, 8GB 1867 MHz DDR3 RAM. The codes were written in Python and R programming languages. The efficiency comparison of the algorithms is done on the basis of their runtime as depicted in Fig. 4.

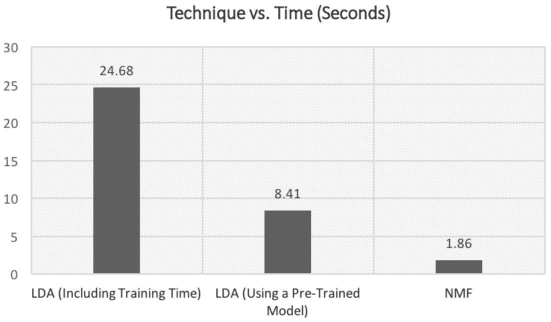
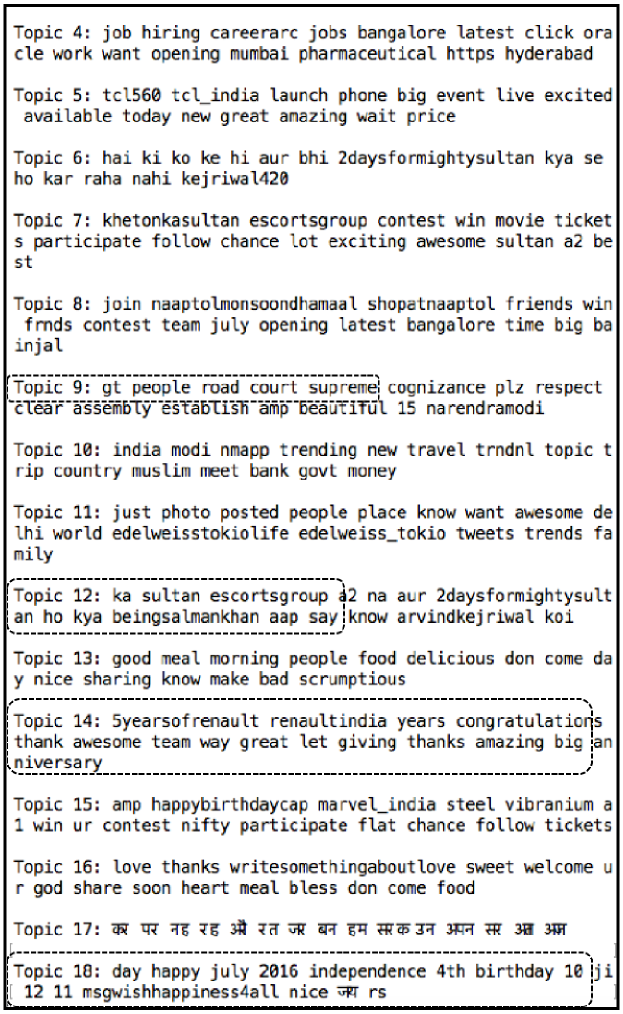


Fig. 4. Comparison of LDA & NMF based on their runtime.

During the experiment, it was observed that the time taken by LDA (including training time) was 24.68 seconds; LDA (using a pre-trained model) was 8.41 seconds; NMF was 1.86 seconds.



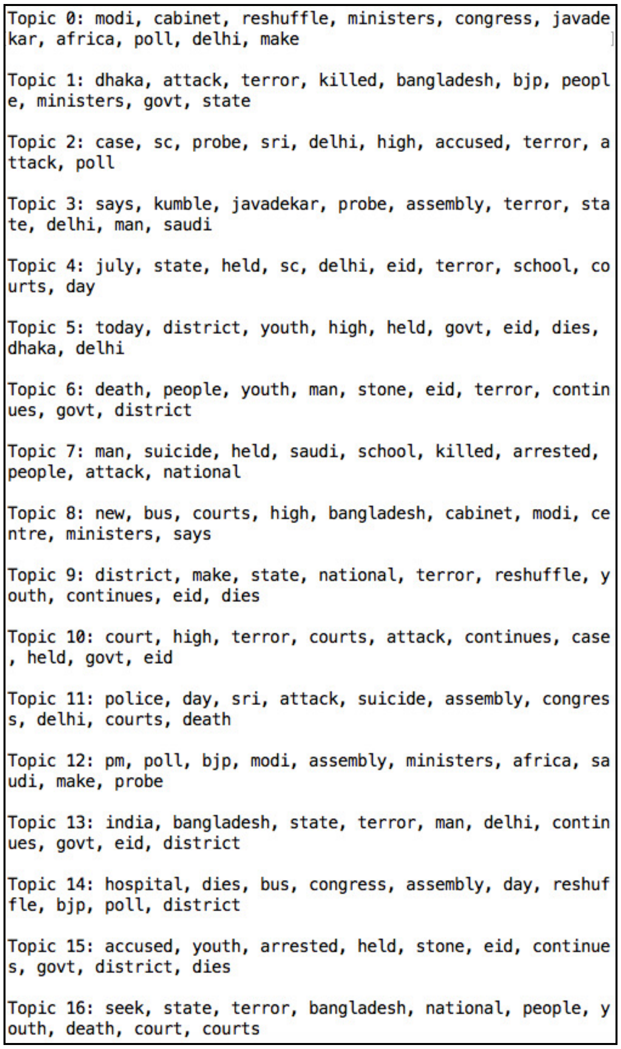
```
Topic 4: job hiring careerarc jobs bangalore latest click ora
cle work want opening mumbai pharmaceutical https hyderabad

Topic 5: tcl560 tcl_india launch phone big event live excited
 available today new great amazing wait price

Topic 6: hai ki ko ke hi aur bhi 2daysformightysultan kya se
ho kar raha nahi kejriwal420

Topic 7: khetonkasultan escortsgroup contest win movie ticket
s participate follow chance lot exciting awesome sultan a2 be
st

Topic 8: join naaptolmonsoondhamaal shopatnaaptol friends win
 frnds contest team july opening latest bangalore time big ba
injal

Topic 9: gt people road court supreme cognizance plz respect
clear assembly establish amp beautiful 15 narendramodi

Topic 10: india modi nmapp trending new travel trndnl topic t
rip country muslim meet bank govt money

Topic 11: just photo posted people place know want awesome de
lhi world edelweisstokiolife edelweiss_tokio tweets trends fa
mily

Topic 12: ka sultan escortsgroup a2 na aur 2daysformightysult
an ho kya beingsalmankhan aap say know arvindkejriwal koi

Topic 13: good meal morning people food delicious don come da
y nice sharing know make bad scrumptious

Topic 14: 5yearsofrenault renaultindia years congratulations
thank awesome team way great let giving thanks amazing big an
niversary

Topic 15: amp happybirthdaycap marvel_india steel vibranium a
1 win ur contest nifty participate flat chance follow tickets

Topic 16: love thanks writesomethingaboutlove sweet welcome u
r god share soon heart meal bless don come food

Topic 17: कर पर नह रह औ रत जर बन हम सकउन अपन सर अत अम

Topic 18: day happy july 2016 independence 4th birthday 10 ji
 12 11 msgwishhappiness4all nice जय rs
```

Fig. 5. Topics detected on application of NMF on twitter data (with 16152 tweets) obtained over a period of 2 hours (4th July 2016, 12 Noon to 2:00PM).

On applying a topic modeler on the corpus obtained from Twitter, it was noted that most of the topics detected do not represent any real event. However, some events such as – release of Bollywood film 'Sultan'[1], USA's Independence Day[2] were discovered, which can be confirmed through human inspection in Fig. 5. Hence, proving the possibility of detecting an arbitrary event from textual data streams obtained from social media.

The procedure to model topics was repeated on a dataset consisting of news headlines obtained from RSS feed of popular news websites. The output has been shown in Fig. 6. and Fig. 7. for NMF and LDA algorithms respectively.

On inspection, it was noted that semantically results generated by LDA are more meaningful than the ones modeled by NMF.

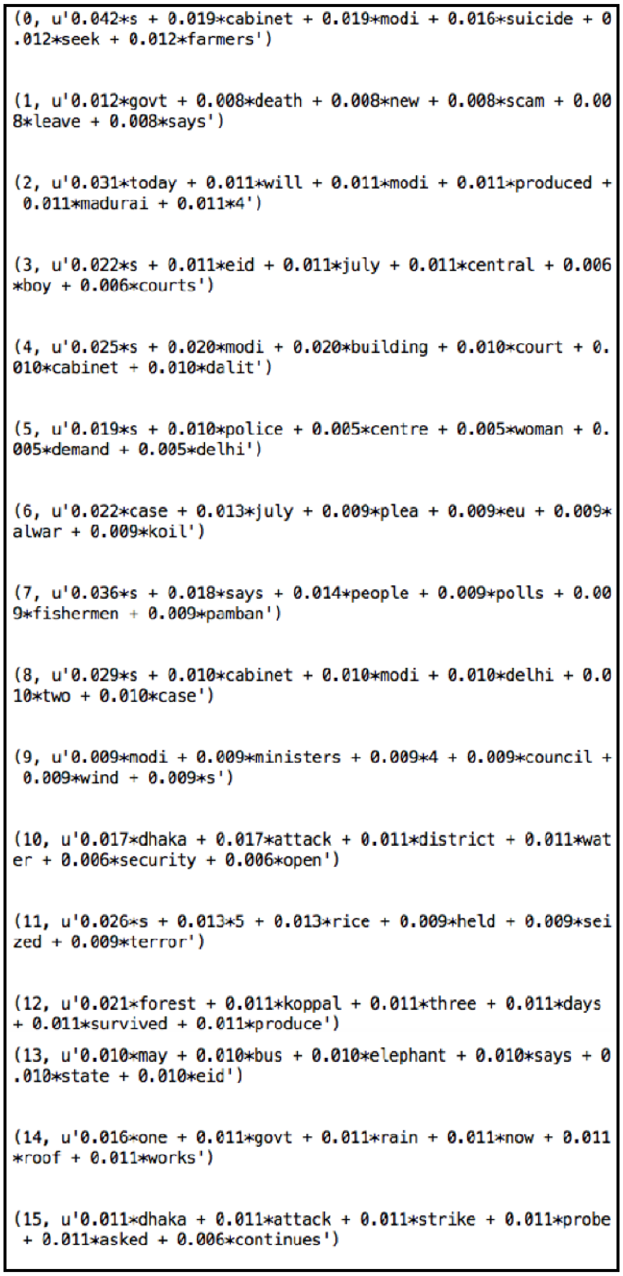


```
Topic 0: modi, cabinet, reshuffle, ministers, congress, javade
kar, africa, poll, delhi, make

Topic 1: dhaka, attack, terror, killed, bangladesh, bjp, peopl
e, ministers, govt, state

Topic 2: case, sc, probe, sri, delhi, high, accused, terror, a
ttack, poll

Topic 3: says, kumble, javadekar, probe, assembly, terror, sta
te, delhi, man, saudi

Topic 4: july, state, held, sc, delhi, eid, terror, school, co
urts, day

Topic 5: today, district, youth, high, held, govt, eid, dies,
dhaka, delhi

Topic 6: death, people, youth, man, stone, eid, terror, contin
ues, govt, district

Topic 7: man, suicide, held, saudi, school, killed, arrested,
people, attack, national

Topic 8: new, bus, courts, high, bangladesh, cabinet, modi, ce
ntre, ministers, says

Topic 9: district, make, state, national, terror, reshuffle, y
outh, continues, eid, dies

Topic 10: court, high, terror, courts, attack, continues, case
, held, govt, eid

Topic 11: police, day, sri, attack, suicide, assembly, congres
s, delhi, courts, death

Topic 12: pm, poll, bjp, modi, assembly, ministers, africa, sa
udi, make, probe

Topic 13: india, bangladesh, state, terror, man, delhi, contin
ues, govt, eid, district

Topic 14: hospital, dies, bus, congress, assembly, day, reshuf
fle, bjp, poll, district

Topic 15: accused, youth, arrested, held, stone, eid, continue
s, govt, district, dies

Topic 16: seek, state, terror, bangladesh, national, people, y
outh, death, court, courts
```

Fig. 6. Topics detected on applying NMF on news headlines obtained from RSS News Feed collected over a period of 24 hours (from 4th July 2016 to 5th July 2016, 6:34 PM).

[1] http://www.imdb.com/title/tt4832640/

[2] https://www.cia.gov/library/publications/the-world-factbook/fields/2109.htm[1]



```
(0, u'0.042*s + 0.019*cabinet + 0.019*modi + 0.016*suicide + 0
.012*seek + 0.012*farmers')


(1, u'0.012*govt + 0.008*death + 0.008*new + 0.008*scam + 0.00
8*leave + 0.008*says')


(2, u'0.031*today + 0.011*will + 0.011*modi + 0.011*produced +
 0.011*madurai + 0.011*4')


(3, u'0.022*s + 0.011*eid + 0.011*july + 0.011*central + 0.006
*boy + 0.006*courts')


(4, u'0.025*s + 0.020*modi + 0.020*building + 0.010*court + 0.
010*cabinet + 0.010*dalit')


(5, u'0.019*s + 0.010*police + 0.005*centre + 0.005*woman + 0.
005*demand + 0.005*delhi')


(6, u'0.022*case + 0.013*july + 0.009*plea + 0.009*eu + 0.009*
alwar + 0.009*koil')


(7, u'0.036*s + 0.018*says + 0.014*people + 0.009*polls + 0.00
9*fishermen + 0.009*pamban')


(8, u'0.029*s + 0.010*cabinet + 0.010*modi + 0.010*delhi + 0.0
10*two + 0.010*case')


(9, u'0.009*modi + 0.009*ministers + 0.009*4 + 0.009*council +
 0.009*wind + 0.009*s')


(10, u'0.017*dhaka + 0.017*attack + 0.011*district + 0.011*wat
er + 0.006*security + 0.006*open')


(11, u'0.026*s + 0.013*5 + 0.013*rice + 0.009*held + 0.009*sei
zed + 0.009*terror')


(12, u'0.021*forest + 0.011*koppal + 0.011*three + 0.011*days
+ 0.011*survived + 0.011*produce')
(13, u'0.010*may + 0.010*bus + 0.010*elephant + 0.010*says + 0
.010*state + 0.010*eid')


(14, u'0.016*one + 0.011*govt + 0.011*rain + 0.011*now + 0.011
*roof + 0.011*works')


(15, u'0.011*dhaka + 0.011*attack + 0.011*strike + 0.011*probe
 + 0.011*asked + 0.006*continues')
```

Fig. 7. Topics detected on applying LDA on news headlines obtained from RSS News Feed collected over a period of 24 hours (from 4th July 2016 to 5th July 2016, 6:34 PM).

VI. Conclusion and Future Work

In this paper, analysis of topic identification algorithms has been done in the context of text messages obtained from the social microblog Twitter and RSS feed of popular media channels. It can be concluded for semantic analysis for event detection systems, the proposal for topic detection is feasible for human analytical purposes – to obtain the main topics related to an arbitrary event.

Although the topic detection system works fairly well, it has it shortcoming when it comes to detecting real events. This is because of the fact that there is a lot of noise in data obtained from Twitter.

On the basis of the experimental analysis, it can also be noted that for a real-time event detection system, LDA is a slow algorithm and it'd be a better choice to use NMF. However, if runtime is not a constraint, LDA outperforms NMF.

This work can be extended by adding a location estimation method for events detected by the topics modelled. Moreover, the source of data can be expanded to multiple sources like Wikipedia & Facebook. Moreover, visualization of results and addition of a GUI can make the system ready for a layman to use and interpret the topics.

References

[1] Naaman, Boase, and Lai. Is it really about me? Message content in social awareness streams. CSCW10. 2010.

[2] Twitter Team, 2012. Twitter turns six. http://blog.twitter.com/2012/03/twitter-turns-six.html accessed at July 08, 2016.

[3] Hila Becker, Mor Naaman and Luis Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 438–441. 2011.

[4] Ryaboy,Dmitriy & Lin, Jimmy. Scaling Big Data Mining Infrastructure: The Twitter Experience. ACM SIGKDD. 2012.

[5] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. ACM Press, New York. 1999.

[6] C. D. Manning, P. Raghavan and H. Schütze. Introduction to Information Retrieval, Cambridge University Press. 2008.

[7] T. M. Mitchell. 1997. Machine Learning, McGraw-Hill.

[8] Twitter API Announcements, "Update on Whitelisting, http://groups.google.com/group/twitter-api-anounce/browse_thread/thread/1acd954f8a04fa84/9321c609cd8f7751?lnk=gst& q=whitelist#9321c609cd8f7751," 2011

[9] Borasky Research Journal. (2010). The Twitter Streaming API -- How It Works and Why It's A Big Deal, http://borasky-research.net/2010/01/06/the-twitter- streaming-api-how-it-works-and-why-its-a-big-deal/

[10] J. Huang, K. M. Thornton, and E. N. Efthimiadis, "Conversational tagging in twitter," in ACM Hypertext and Hypermedia, Toronto, Ontario, Canada, 2010, pp. 173-178

[11] Zhao, Siqi. Detecting Events from Twitter in Real-Time. Diss. Rice University, 2013.

[12] Chen, Edwin. "Introduction to Latent Dirichlet Allocation". Blog.echen.me. N.p., 2016. Web. 12 July 2016.

[13] Moody Chu and Robert Plemmons. Non-negative matrix factorization and applications. Bulletin of the International Linear Algebra Society, 34, 2– 7. 2005.

[14] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

[15] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[16] Klinczak, Marjori NM, and Celso AA Kaestner. "Comparison of Clustering Algorithms for the Identification of Topics on Twitter." Latin American Journal of Computing Faculty of Systems Engineering National Polytechnic School Quito-Ecuador 3.1 (2016): 19-26.