DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5271

ISSN No. 0976-5697

Volume 9, No. 1, January-February 2018

International Journal of Advanced Research in Computer Science

CASE STUDY

Available Online at www.ijarcs.info

A PRAGMATIC TEXT MINING ANALYSIS OF DEMONETISATION MOVE BASED ON TOPIC MODEL AND TWITTER COMMUNICATIONS

Doddi Srilatha
Department of CSE
Sreenidhi Institute of Science and Technology
Hyderabad, India

Shirisha Kakarla
Department of CSE
Sreenidhi Institute of Science and Technology
Hyderabad, India

Abstract: Demonetisation move has been the most trending topic on the twitter during the recent months. Indian government had taken a sudden decision to phase out the two most high valued denominations of Rupees 500 and 1000 currency notes. This decision was observed with different sentiments among the various sects of the people in the society. It caused sensation in the whole country. In this paper, we considered tweets related to demonetisation move and the speech contents that were delivered by the premier of the country. The significant number of tweets are downloaded, preprocessed to remove the noise like re-tweets, analyzed using frequency analysis and a topic model using Latent Dirichlet Allocation (LDA) is constructed. The LDA model automatically discovered the most relevant topics. Further clustering is done to project the clear view of the citizens' and especially the public leaders' opinions on this move. The most of the topics were related to criticize the Prime Minister's speech addressed to the nation on new year event, although few opined the benefits of demonetisation.

Keywords: Demonetisation; Topic Model; Twitter; Text mining; Latent Dirichlet Allocation

1. INTRODUCTION

The currency note of a nation that was used as a legal tender and is declared as ceased in lieu of its substitute is termed as demonetisation. The need of the demonetization usually appears to counter the ill-effects faced by a nation in using its counterfeits. The Indian Government has earlier ceased the currency notes 1,000 and 10,000 rupees in 1946. Later, India on 16 January 1978 also witnessed the second demonetization phase of 1,000, 5,000 and 10,000 rupees. It was the decision imposed by the then ruling government of Janata Party and its coalition. Indian government had again demonetized banknotes of 500 and 1000 rupees right from 8 November 2016. The premier of the Indian Government along with the Governor, RBI, has taken the decision to curb the 500 and 1,000 rupee notes as legal tender and the people were given a period of fifty days to either exchange their cancelled currency notes to new Rs.500 and Rs.2000 notes or deposit them into their accounts in banks with the stipulated norms.

The decision of demonetization caused a furore among the citizens in India and elsewhere as it caught the unawareness of the people. It was largely reflected among the media, both print and television. The internet also saw an upsurge in expressing the public opinions about the hardships faced by them and the advantages of the post demonetization phase.

Twitter – a popular social media giant, gave the platform to the public where the topics of significance are discussed in forums. Apart from many other topics like Rio Olympics, U.S. Election, PokemonGo, the demonetization was also emerged as widely expressed global topic. On the evening of 08th November, 2016, the Prime Minister of India Mr. Narendra Modi announced the daring decision to ban the currency notes of rupees 500 and 1000 which were the highest denominations in India, till then. Post speech, the excited twitter users took to social media to opine their views. Most of the opinions expressed were of the Indians who faced the challenges and severe discomfort in exchanging their cancelled notes to new notes. The move impacted billions of Indians, in the harsh ways as the people were stranded cashless and the usual economic ecosystem going berserk. Many others, who were directly affected like the normal public and the banking sector or the indirectly affected sects like foreign nationals and organizations spoken their views about the demonetization effect on the Twitter. As such, Indians tweeted 6.5 lakhs tweets within next 24 hours, and millions more in the following days. Twitter continues to reverberate with their beliefs on the pros and cons of the currency ban. The scale of these debates by the experts as well as the common people driving a global conversation on Indian Demonetization effects on Twitter.

In this paper, the section II presents the Literature Survey. The section III is focused on data collection and the preprocessing. In this section the significant tweets are collected and are readied for the investigation. The section IV discusses the analysis and the findings derived thereafter. In section V, conclusions are made from the findings of the twitter data.

2. LITERATURE SURVEY

Twitter, is a micro-blogging online platform and social networking service, for communicating and discussing on various subjects. The users can tweet about any subject within the 140 characters. Users use and create hashtags by placing the # in front of a word either in the main text of a message or at the end [1]. In Twitter or any social media service, hashtag (#) is a type of metadata tag or label which makes it easier for users to find messages with a specific content or theme or event discussed. Searching for a hashtag will defer each message that has been tagged with it. The tweets collected are to be preprocessed and analyzed further to derive a semantic conclusion on the trending topic. The analysis is performed by using any of the topic models.

Topic models [2][3] are a type of statistical model used in text mining to discover the abstract "topics" that may appear in a collection of documents, called corpus. This also facilitates the identification of the patterns in a corpus. Topic models [3] are primary tools to identify latent text patterns in the unstructured content. As the social media content, largely an unstructured data, differs from some standard text domain having web pages, citation network, etc, the standard text mining techniques does not provide appropriate outcomes. A number of topic modeling techniques that were recently worked out and elaborately developed are namely, Mixture of Unigrams (MU) model, Latent Semantic Indexing (LSI), Pachinko Allocation Model (PAM), Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation (LDA), which assists in uncovering the hidden semantic structures on the contextual information. Given a corpus of any given topic to be analyzed, one may expect particular words or its synonyms/substitutes to be repeated often.

Topic model techniques are therefore helpful in a wide range of areas including recent work on Twitter, done by D. Ramage et al. [4]. The outcomes are represented in the form of the mathematical models and the clusters of similar words that allow the user to draw conclusions of the trending sub-topics in the collected information.

A. *Latent Dirichlet Allocation*

In this paper, we use the Latent Dirichlet Allocation (LDA) as one of the topic models. This was first proposed by Blei et al. [2][3]. The LDA model [3] is considered to be a non-supervised technique with probabilistic and the statistical approach. The LDA is used for modeling large unstructured text corpus by discovering latent hidden semantic topics. Every document conceals few latent semantic structures in the form of groups, whereas each group is a collection of inter-related words. The key approach used in the LDA is to extract the latent semantically structured information in the form of words of the corpus. This is performed by identifying the frequently occurring group of words in the corpus. LDA is particularly useful in the given situation where the corpus is unstructured and no idea how the data are interrelated. One of the first things that one can do is to generate a near idea by classifying the words in the corpus and to generate the topics are the subjects. The pictorial representation of LDA [5] is shown in Fig. 1.

B. *Notation and Terminology*

i). A word w $\in$ {1, 2, 3,. . . , V} is the most basic unit of discrete data. For cleaner notation, w is a V-dimensional unit-based vector. If w takes on the i th element in the vocabulary, then w ^{i}=1 and w^{j} =0 for all j $\neq$i.

ii). A document is a sequence of N words denoted by W = (w_1, w_2, w_3, . . . , w_N), where w_n is the nth word in the sequence.

iii). A corpus is a collection of M documents denoted by D = {w_1, w_2 ,w_3, . . . , w_M}.

iv). A topic z $\in${1, 2, 3, . . . , K} is a probability distribution over the vocabulary of V words.

Topic model is a particular groups of words that commonly occur collectively in text documents in a corpus, and thus can be understand as “topics” and z = (z_1 , z_2 , . . . , z_N) denotes the sequence of topics across all words in a document.

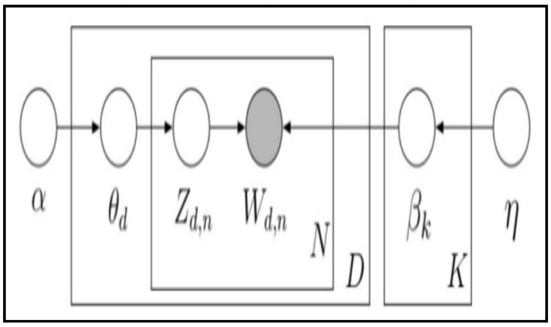

Figure 1. Graphical Representation of LDA

Where
K - total number of topics
$_k$ - topic distribution over the vocabulary
D - total number of documents
$_d$ - per document proportions
N - total number of words in document (in fact, it should be N_d)
$Z_{d,n}$ - per-word topic assignment
$W_{d,n}$- observed word
- dirichlet parameters

C. *Generative Process of LDA*

In order to derive topics from a corpus one has to have a certain model of how documents are generated. LDA is basically inputted with the collection of documents containing the words. The order of the words is insignificant and remained exchangeable with contextual synonyms. This leads to documents’ representation as the bag-of-words, where each term is indexed with respective document within a collection, with its corresponding frequency. This is termed as term-document matrix (and sometimes also referred as document-term matrix). LDA’s generative model [5] can be analyzed as follows:

I. Decide what words are likely appear for each topic.
II. For each document in the corpus,
a) Decide what proportions of topics should be in the document,
b) For each word,
i. Choose a topic,
ii. Given this topic, choose a likely word (generated in step I).

3. DATA AND METHODOLOGY

The input data for this study have been pooled from the twitter social network using the handler: #demonetisation. A total of ten thousand tweets have been collected that were posted in the twitter communication, soon after the prime minister addressed the nation on the demonetisation move. The civilians and the public figures had to accept this move with the mixed emotions and were reflected well in the virtual social communications. With this unexpected and immediate decision by the beaurocrats, came the newly framed regulations which have taken the nation by surprise in understanding these new guidelines and following them in the nook and corner of the Indian landscapes.

For the purpose of gathering the tweets, the API Authentication of R studio tool [6] is used. R Studio is a tool for statistical data modeling and widely used by data miners to study and analyze the large scale datasets. In this study of tweets related to demonetisation move, text mining is done using the following libraries like “tm”, “twitter”, “topic models”, wordcloud etc. After collection of the tweets, the cleansing is done for removing nonesse-ntial characters such as web addresses, punctuation marks, hashtags, retweets, user handles, html links, time stamps, numbers and special characters, etc. Post cleanup a corpus of tweets devoid of unneeded data is obtained.

The efficacy of cleansing (pre-processing) procedure used is observed by plotting the word cloud. The frequently appearing terms in the tweets in the preprocessed corpus can be viewed in the wordcloud shown in the Fig. 2.

Post this, a term-document matrix is created depicting the most tweeted words upfront. Here, some of the top most frequent key terms are shown in the Table 1 in a row wise manner.

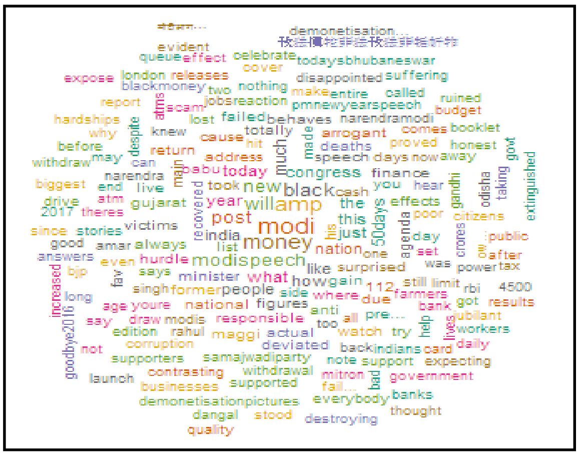

Figure 2. Plot of Word Cloud for the Demonetisation Move Tweets

Table 1. Frequent Key Terms Appearing in Tweets Related to Demonetisation Move

"112"	"50days"	"actual"	"address"	"agenda"
"always"	"amp"	"anti"	"arrogant"	"babu"
"behaves"	"black"	"cash"	"cause"	"congress"
"day"	"days"	"deaths"	"demonetisation"	"demonetisationspeech"
"due"	"effects"	"failed"	"farmers"	"figures"
"finance"	"former"	"gain"	"gujarat"	"his"
"how"	"hurdle"	"india"	"just"	"like"
"list"	"live"	"made"	"maggi"	"main"
"minister"	"modi"	"modis"	"main"	"money"
"much"	"nation"	"national"	"new"	"now"
"people"	"pmnewyearspeech”	"poor"	"post"	"pre…"
"recovered"	"responsible"	"return"	"says"	"speech"
"still"	"surprised"	"the"	"this"	"today"
"took"	"totally"	"victims"	"was"	"watch"
"what"	"where"	"will"	"year"	"you"

4. ANALYSIS AND FINDINGS

Term Document Matrix creation helped in finding term’s frequency and association between terms using correlation. This had been helpful in conducting hierarchical clustering analysis using Ward method.

The widely popular technique used in the text corpus mining is the cluster analysis. Out of the numerous techniques used for clustering, the hierarchical clustering is the most used one and is done here using the Ward’s method. Although a dendrogram is outputted and shown in the Fig. 3, which is suitable for browsing, it usually suffers from efficiency problems. This is due to the fact that it does not provide complete analysis as some of indexed words do not help in further analysis and no major value addition is guaranteed towards topic modeling. In the dendrogram the right most cluster having the words (modi, amp, new, year, modispeech, and will) is considered insignificant and ignored. The remaining clusters with the bag-of-words in each make sense and the same resulted in the frequency analysis.

Latent Dirichlet Allocation (LDA) as taken as one of the topic modeling techniques. For proceeding, a document-term matrix is created by removing the sparse terms with the threshold value set at 0.98, and parameters: row sums greater than zero and selection of ten topics.

Identification of frequent terms in the tweets, highlights the some of the keywords based on frequency and a look at the some of the top words (shown in the Table 1) will be related to Prime Minister Narendra Modi speech addressed the nation on new year event.

To further enhance the analysis and understanding, topic models using Latent Dirichlet Allocation (Gibbs Method) was conducted on the corpus of documents. The LDA topic model was run with 10 topics as criteria and the output was highlighted in the Table 2.

In this paper, we considered ten topics and discovered as follows:

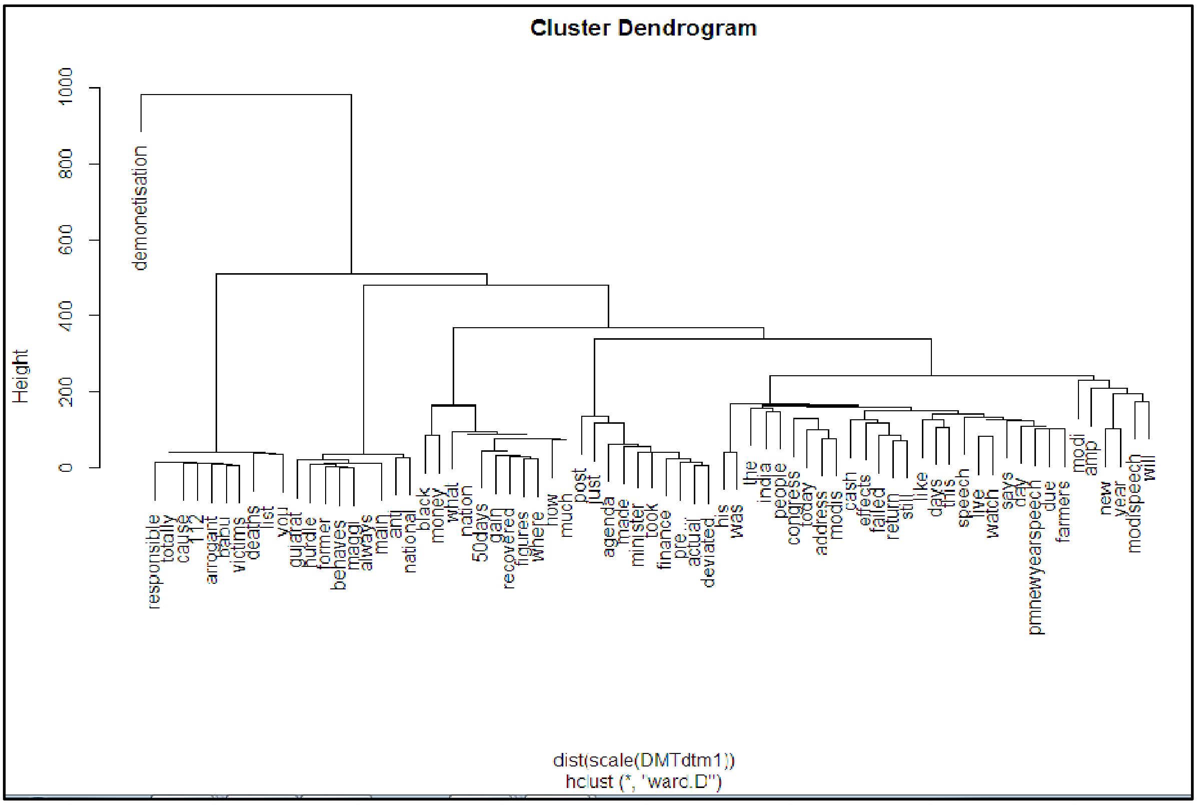

Figure 3. Data Visualization using Hierarchical Clustering and Ward Method

Table 2. Results of Topic models using Latent Dirichlet Allocation (Gibbs Method)

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
"demonetisation"	"nation"	"demonetisation"	"post"	"like"	"demonetisation"	"will"	"demonetisation"	"modi"	"demonetisation"
"india"	"money"	"cash"	"just"	"this"	"modi"	"modispeech"	"the"	"you"	"amp"
"his"	"what"	"congress"	"money"	"anti"	"says"	"today"	"speech"	"deaths"	"new"
"people"	"how"	"effects"	"black"	"national"	"people"	"address"	"day"	"list"	"year"
"was"	"much"	"failed"	"made"	"gujarat"	"india"	"watch"	"pmnewyearspeech	"responsible"	"due"
"modi"	"black"	"return"	"agenda"	"always"	"live"	"demonetisation"	"farmers"	"112"	"farmers"
"speech"	"50days"	"days"	"took"	"main"	"pmnewyearspeech"	"modis"	"modi"	"totally"	"post"
"says"	"recovered"	"still"	"minister"	"hurdle"	"farmers"	"live"	"post"	"arrogant"	"will"
"modispeech"	"figures"	"modispeech"	"finance"	"former"	"effects"	"congress"	"what"	"cause"	"list"
"amp"	"gain"	"much"	"deviated"	"maggi"	"speech"	"days"	"amp"	"babu"	"today"
"pmnewyearspeech	"where"	"watch"	"actual"	"behaves"	"congress"	"nation"	"just"	"victims"	"deaths"
"actual"	"effects"	"figures"	"amp"	"people"	"day"	"black"	"recovered"	"india"	"says"
"agenda"	"day"	"new"	"pre…"	"speech"	"will"	"says"	"says"	"congress"	"agenda"
"the"	"list"	"anti"	"will"	"new"	"days"	"due"	"return"	"demonetisation"	"days"
"cause"	"speech"	"day"	"demonetisation"	"gain"	"year"	"former"	"took"	"took"	"how"

Topic 1 can be classified as related to PM Modi's new year speech had actual agenda of demonetisation in India.

Topic 2 relates to PM Modi's new year speech where are the figures of demonetization, how much of black money was recovered ,what did the nation gain after demonetisation.

Topic 3 relates to demonetisation effects failed to watch cash figures return congress days that like this.

Topic 4 can be related to PM Modi deviated from actual agenda of black money and demonetisation, just took over the post of finance minister and made pre-budget speech.

Topic 5 opines that the people of Gujarat in their speech are anti-national and always behaves as the main hurdle.

Topic 6 related to demonetization in India effects people and formers in the coming days of the year.

Topic 7 related to Modi will address the nation today on demonetisation and black money, watch speech live.

Topic 8 related to post Modi's speech on demonetisation what was recovered and what farmers took to return.

Topic 9 related to Modi, you are totally arrogant and responsible for the victims list of 112 deaths in India that demonetisation took, says congress.

Topic 10 did not provide any better contextual meaning and hence ignored.

The correct and relevant topics from the document corpus are generated from tweets' data and are automatically enlisted using Gibbs method under LDA Modeling.

5. CONCLUSION

Tweets under the demonetisation twitter handler are downloaded using API authentication and the corpus is preprocessed. A word cloud is plotted, and visualization using hierarchical clustering is performed for discovering the most relevant topics. The topic modeling involving Latent Dirichlet Allocation using Gibbs method is done to uncover the most relevant topics.

The most of the topics were related to PM Modi's new year speech had actual agenda of demonetisation in India, where are the figures of demonetization, how much of black money was recovered ,what did the nation gain after demonetisation, demonetisation effects failed to watch cash figures return congress days that like this, PM Modi deviated from actual agenda of black money and demonetisation, just took over the post of finance minister and made pre-budget speech, opines that the people of gujarat in their speech are anti-national and always behaves as the main hurdle, demonetization in India effects people and formers in the coming days of the year, Modi will address the nation today on demonetisation and black money, watch speech live, post Modi's speech on demonetisation what was recovered and what farmers took to return, Modi, you are totally arrogant and responsible for the victims list of 112 deaths in India that demonetisation took, says congress.

From the above discussion and analysis, the LDA can be seen as the best method for topic modeling in order to automatically discover topics and trends in such large scale twitter datasets.

REFERENCES

[1] Hashtag: https://en.wikipedia.org/wiki/Hashtag

[2] Blei, David, "Probabilistic Topic Models". Communications of the ACM. 55 (4): PP:77–84. doi:10.1145/2133806.2133826 , 2012.

[3] D. Blei and J. Lafferty. Topic Models. Text Mining: Theory and Applications, 2009.

[4] D. Ramage, S. Dumais, and D. Liebling. Characterizing Microblogs with Topic Models. In International AAAI Conference on Weblogs and Social Media, 2010.

[5] David M. Ble, Andrew Y. Ng, and Michael I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3, 993-1022, 2003.

[6] R Studio: https://www.rstudio.com/