ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Sentiment Analysis of News Articles using Probabilistic Topic Modeling

[1] Gaurav M Pai, [2] Paramesha K, [3] K C Ravishankar
[1][2] Vidyavardhaka College of Engineering, Mysore, 570002, India
[3] Government Engineering College, Hassan, 573202, India

Abstract: **- In this age, information is available in abundance on the internet, there are many platforms where news articles are published. It is nearly impossible to manually read articles from all the sources to understand the opinions of the authors. We require sentiment analysis systems to automatically detect sentiments of topics discussed by authors of these articles. Sentiment Analysis is the process of evaluating a piece of text to determine whether the expression is positive, negative or neutral in nature. In this paper, we present a system that performs Sentiment Analysis of topics which are discovered from a collection news articles using probabilistic topic modeling technique called Latent Dirichlet Allocation. We found that there was coherence between our results and the real-world scenarios that existed at the time of publication of the articles.**

Index Terms— Sentiment Analysis, News Articles, LDA, Topic Modeling, Computational Linguistics, Machine Learning

I. INTRODUCTION

News articles and blogs, along with reporting recent events, express opinions about news entities. We can use machines as an aid to analyze these opinions expressed by the authors. Even though machines cannot understand the exact and complete opinions that are conveyed by the natural language, the statistical analysis of simple cues that characterize the sentiments [1] can lead to meaningful results as to how the topics discussed in the news articles and real-world events are correlated In this paper, we present a system for modeling a large collection of news articles from a domain of news, into topics and compute the sentiment scores of these topics and determine whether the discussion of those topics had a positive or negative tone.

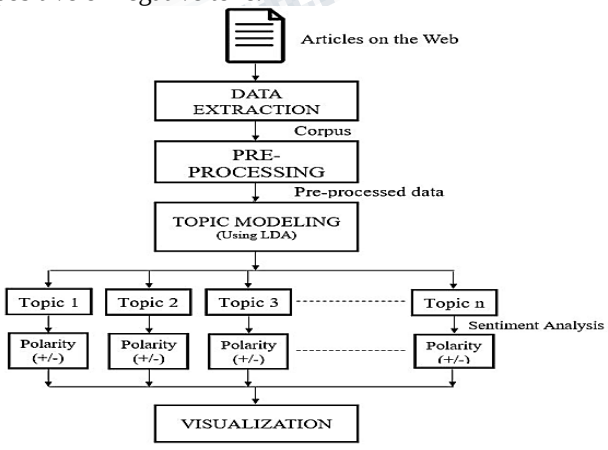

Figure 1: Flow Diagram

II. LITERATURE SURVEY

Most of the previous work falls under two main categories. The first pertains to the techniques that automatically generate sentiment lexicons [2]. The second category pertains to the systems that perform analysis of sentiments of the entire document. Our work falls into the second category where we perform the analysis of the entire document. Using topic modeling to summarize the entire collection of articles into topics and analyze subjectivity of these topics.

In various fields such as Hotel, Restaurants, Tourism, Products etc. the sentiment analysis of the reviews of customers has been handled as a Natural Language Processing task but due to fact that sentiments clues could be captured by presence of words and expressions and their interactions. So, in Machine Learning Techniques are employed, lot of supervised learning methods such as Naive Bayes, Maximum Entropy, SVM have been incorporated for good results only in a feature based implementation.

Sentiment analysis to investigate topics related to public opinion of race, gender and religion due to the biases in news articles in [3]. Pang et al [4] perform sentiment analysis on review of movies. They successfully conclude that machine learning techniques perform the sentiment analysis better than simple methods such as term counting.

In recent times, one of popular Probabilistic Topic Modeling techniques that has more plausible and robust

assumptions on the generation of text compared to other methods is Latent Dirichlet Allocation [5] used for Topic modelling in opinionated text. Tian et al. [6] used LDA to index and analyze the source code documents as a mixture of topics. Tasci et al. [7] evaluated the performance of latent Dirichlet allocation based representation in text categorization. Since the results in this paper were promising we use the same approach to analyze the entire corpus as a mixture of topics. Putri et al. [8] have used Latent Dirichlet Allocation to discover general tendency from tourist review into certain topics that can't be classified toward positive and negative sentiment. In this paper, we use a similar method to categorize the corpus into topics and then analyze sentiment of these topics. Another work [9] proposes a semantic parser that extracts concepts from sentences. This is done by subdividing the sentences and addresses the fine-grained sentiment analysis. But as it is said earlier, we aim to evaluate the sentiments that are conveyed by simple statistical cues, hence we opted to evaluate sentiment scores of the words using SentiWordNet 3.0 that gave us meaningful results.

III. DATA EXTRACTION

Data extraction is the process of extracting relevant data from the relevant resources which is later used for processing or storage. In the proposed system, the news articles from websites like www.hindu.com www.indiatimes.timesofindia.com were chosen. A web crawler was used to crawl the website and retrieve all the Uniform Resource Locators (URLs) of the articles. Once the URLs were obtained, they were categorized based on domains such as National, Politics, Sports, and Business. Goose-Extractor [10] was used to extract articles from the website. It is an open source article extraction library in Python. It is strongly based on NLTK (Natural Language Tool Kit) and Beautiful Soup, which are leading libraries in the text processing and HTML parsing. The reason why goose-extractor was chosen is because it could not only extract the main body of the article but also other elements of the article like main image of an article, videos in an article, meta description which contains the date of publishing the articles which was used to extract articles published in a particular time frame only. Here, we used the articles extracted that belonged to National News category as our dataset.

IV. DATASET

The dataset that we have used was prepared on our own. The news articles are taken from websites like www.hindu.com, www.timesofindia.inditatimes.com etc. The dataset comprises of articles from the month of January 2017 to the month of March 2017. The distribution of articles is shown in the table below. All the articles belong to the National News category.

Month	No. of Articles
January 2017	1050
February 2017	1035
March 2017	1115
Total	**3200**

Table I Article Distribution

This dataset was then pre-processed. The pre-processing consisted of steps to remove stop words such as 'a', 'and', 'the' etc. which have no inherent meaning, and stemming of the words. This prepared our dataset for topic modeling using LDA.

V. TOPIC MODELING USING LATENT DIRICHLET ALLOCATION

LDA [5] represents corpus as a cluster of topics which split out into words having some probability.
LDA assumes the following generative process (as shown in the plate notation Figure 2) for each document w in a corpus D:
1) Choose N ~ Poisson(ξ)
2) Choose θ ~ Dir(α)
3) For each of the N words wn:
a) Choose a topic zn ~ Multinomial(θ).
b) Choose a word wn from p(wn|zn, β), a multinomial probability conditioned on the topic zn

As proposed in [5] we assume that k dimensionality of Dirichlet distribution, is fixed. Probability of the words are parameterized by a (k x X) matrix β where βij = p(wj = 1 | zi = 1), this is assumed as a fixed this is assumed as a fixed quantity which needs to be estimated. Poisson assumption is lenient to anything that follows and the realistic documents length distributions will be used as needed. We assume that N, number of words, is independent of the generating variables (θ and z)

A k-dimensional Dirichlet random variable θ can take values in the (k-1)-simplex (a k-vector θ lies in the (k-1)-simplex if θ¬j ≥ 0, , θi = 1), and has the following probability density on this simplex, where the parameter α is a k-vector with components αi > 0, and where γ(x) is the is the Gamma function. The Dirichlet is a convenient distribution on the simplex - it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Given the parameters α and β, a set of N topics z, and a set of N words w is given by:

$p(\theta,z,w|\alpha,\beta) = p(\theta|\alpha)\ p(zn \mid \theta)p(w\neg n|zn,\beta)$ (1)

where $p(zn \mid \theta)$ is simply θi for unique I such that zni = 1

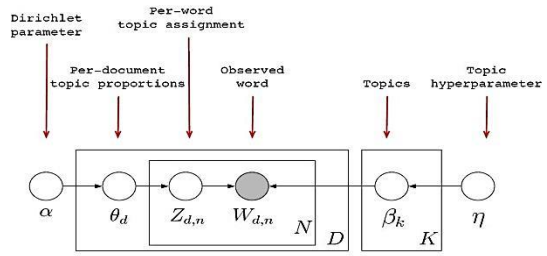

Figure 2 Latent Dirichlet Allocation Plate Diagram

Goods and Service Tax(GST)	Jammu-Kashmir Issue	Demonetization
Government	Pakistan	Sad
Crore	Kashmir	Budget
Said	Two	Government
tax	Indian	Budget
GST	Along	Minister
Would	Pakistani	Lakh
Jaitley	Came	Union
Farmers	Jammu	Year
Finance	Army	Crore
Act	Terror	Demonetization
Per	Ministry	Finance
Minister	March	Tax
Centre	Said	Capital
State	Srinagar	Farmers
Union	Friday	This
Budget	Recorded	Sector
Tax	Uri	disappointment
Council	Hours	Jaitley
Amendments	LoC	Chief

Table II Example words generated for topics

IV. SENTIMENT ANALYSIS

Sentiment Analysis also known as opinion mining aims to determine the attitude of the speaker or a writer with respect to some topic or the overall contextual polarity of a piece of text. The polarity of the given word can be positive, negative or neutral.

In this paper, we present a system that assigns positive and negative scores each distinct word in the topics generated We use the dictionary called SentiWordNet 3.0 [6] to find the positive score as well as negative score of each word in the documents to find out polarity of the topic. SentiWordNet is a sentiment lexicon associating sentiment information to each wordnet synset i.e., SentiWordNet is formed by Wordnet and Sentiment Information. A typical use of SentiWordNet is to enrich the text representation in opinion mining applications, adding information on the sentiment-related properties of the terms in text. SentiWordNet 3.0 has positive and negative score for each synID. i.e. the unique identification given to each synset. Each word may be a noun, verb, adverb or adjective. The main aim is to find the polarity of the topic, as to whether it is positive or negative by comparing its positive score and negative score.

The scores are computed using the following steps:

1) Find all the positive scores for each word a topic and compute the average score. This gives the positive score of the word

$$word\_pos\_polarity = \frac{pos\_sentiment\_references}{total\_sentiment\_references} \quad (2)$$

2) Find all the negative scores for each word a topic and compute the average score. This gives the negative score of the word

$$word\_neg\_polarity = \frac{neg\_sentiment\_references}{total\_sentiment\_references} \quad (3)$$

3) Repeat 2 and 3 for all the words in a topic

4) Add the positive scores of all the words associated to the topic

$$topic\_pos\_polarity = \sum_{1}^{n_t} word\_pos\_polarity \quad (4)$$

5) Add the negative scores of all the words associated to the topic

$$topic\_neg\_polarity = \sum_{1}^{n_t} word - neg\_polarity \quad (5)$$

6) Repeat 4 and 5 for each topic.

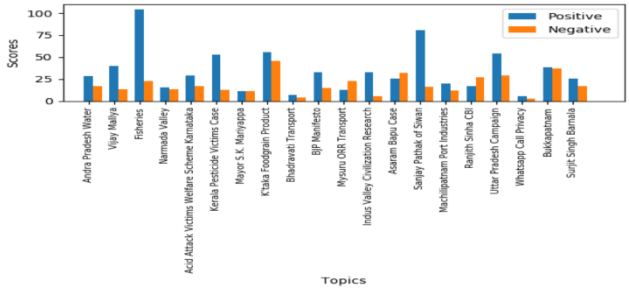

Figure 3: Sentiment scores of topics in the month of January 2017

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

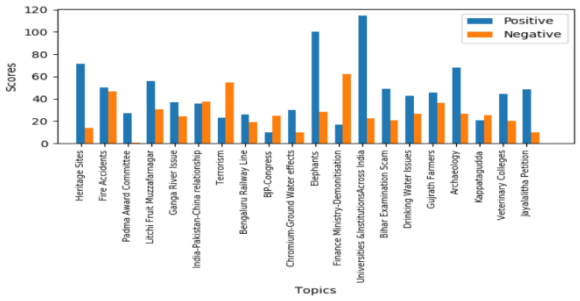

Figure 4: Sentiment scores of topics in the month of February 2017

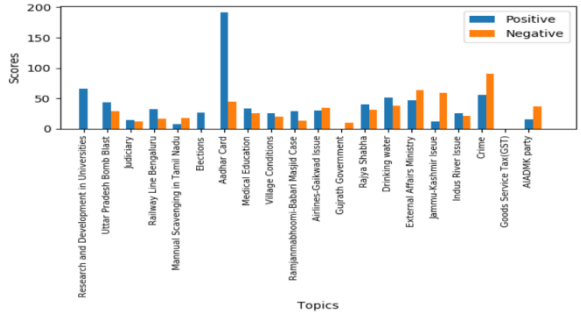

Figure 5: Sentiment scores of topics in the month of March 2017

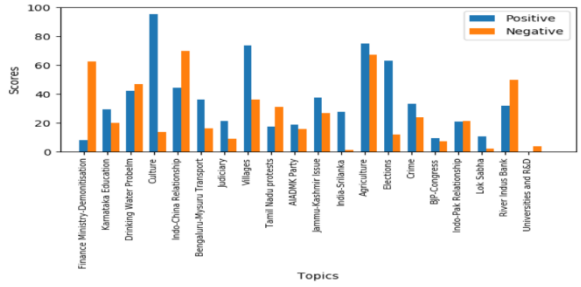

Figure 6: Sentiment scores of topics from January-March 2017

VII. DISCUSSION

The sentiment analysis of topics and the results of sentiment scores were represented in the form of graphs are as shown in Figures 3, 4, 5 depict sentiment scores of topics from the month of January, February and March of 2017 respectively. Figure 6 shows the sentiment score of topics discovered when all the articles from January, February and March was combined.

We found that the insights of topics and sentiment got from the analysis had correlation with the real-world scenario for afore mentioned time period. For instance, in the graph for the month of January (Figure 3), the Asaram Bapu case had received a high negative sentiment score, accordingly the news articles had negativity on the topic and also the interim bail plea of Asaram Bapu was rejected.

Similarly, in the graphs for the month of February (Figure 4) the demonetization of 500 and 1000-rupee notes by the finance ministry, received a high negative sentiment, which was due to the difficulties faced by the public during the early days of demonetization in India the same topic also resurfaced in Figure 6, when all the articles from January, February and March 2017 were combined

VIII. Conclusion and Future Work

In this preliminary work, the proposed model and implementation of the system in the analysis of news article showed some correlation the with real world scenarios and with opinion of general public. Of course, the results are limited by the size of the dataset. The main purpose of the work is to analyze the bipartisanship of various news publishers on government decision and policies. With comparative study of new articles from different outlets, we could estimate their bigotry and fairness stand across various current affairs and issues. However, the most challenging part is to validate the results. In our future work, we would like extend the model for huge dataset with broad range of topics and leverage the model with efficient sentiment capturing techniques, and provide a framework for validation of sentiments.

References

[1] K. Paramesha and K. Ravishankar, “A perspective on sentiment analysis,” arXiv preprint arXiv:1607.06221, 2016.

[2] K. Paramesha and K. C. Ravishankar, “Exploiting dependency relations for sentence level sentiment classification using svm,” in 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), March 2015, pp. 1–4.

[3] D. P. Haider-Markel, M. D. Allen, and M. Johansen, “Understanding variations in media coverage of us supreme court decisions: Comparing media outlets in their coverage of Lawrence v. texas,” Harvard International Journal of Press/Politics, vol. 11, no. 2, pp. 64–85, 2006.

[4] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-

Volume 10. Association for Computational Linguistics, 2002, pp. 79–86.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm? id=944919.944937

[6] K. Tian, M. Revelle, and D. Poshyvanyk, "Using latent dirichlet allocation for automatic categorization of software," in Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on. IEEE, 2009, pp. 163–166.

[7] S. Tasci and T. Gungor, "Lda-based keyword selection in text categorization," in Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on. IEEE, 2009, pp. 230–235.

[8] I. Putri and R. Kusumaningrum, "Latent dirichlet allocation (lda) for sentiment analysis toward tourism review in indonesia," Journal of Physics: Conference Series, vol. 801, no. 1, p. 012073, 2017. [Online]. Available: http://stacks.iop.org/1742-6596/801/ i=1/a=012073

[9] P. Raina, "Sentiment analysis in news articles using sentic computing," in Data Mining Workshops (ICDMW), 2013 IEEE
13th International Conference on. IEEE, 2013, pp. 959–962.

[10] X. Grangier, "Goose-Extractor," https://pypi.python.org/pypi/ goose-extractor/, 2015.

[11] K. Paramesha and K. Ravishankar, "Analysis of opinionated text for opinion mining," arXiv preprint arXiv:1607.02576, 2016.

[12] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009, pp. 375–384.

Gaurav M Pai: Under-graduated student in Vidyavardhaka College of Engineering major in Computer Science

Paramesha K.: Associate Professor, Department of Computer Science, Vidyavardhaka College of Engineering.

K C Ravishankar: Professor in Govt. Engineering College, Hassan India