Metoo Movement Analysis through the Lens of Social Media

P.Asha, K. Sri Neeharika, T. Sindhura

Abstract*—Sentiment analysis is an errand which is used to analyse people's opinions which has been derived out of textual data seems productive for palpating various NLP applications. The grievances associated with this task is that, there prevails variety of sentiments within these documents, accompanied with diverse expressions. Therefore, it seems hard to whip out all sentiments employing a dictionary which is commonly used. This work attempts at constructing the domain sentiment dictionary, by employing the external textual data. Besides, various classification models could be utilised to classify the documents congruent to their opinion. We have also implemented topic modelling, emoticon analysis and optimized gender classification in our proposed system. Many sectors have been identified where women are being abused. Clusters are formed for these sectors and the most affected sector is also identified.

***Keywords*—Sentiment analysis, cluster, Classifier, Modelling.**

I. Introduction

Based on vast augmentation in networks, internet has turned out to be a basic need for human survival. This development in internet increased the connectivity among people around the globe. People are getting more exposed to social media platforms in every possible way. So public opinion analysis has become a trend in the society before any further step in any industry. Hence the insistence for sentiment analysis along with opinion mining is burgeoning. In this era of machine learning sentimental analysis [1-3] plays a pivotal role in creating awareness through analyzing a big sample of social media users who share their thoughts, emotions and opinions. In this work, text mining helps is used to obtain results. As there are many social media platforms on the internet, one among them is twitter. The main use of this social network is that it contains hashtags which makes our task easier for data collection.

II. Literature Review

Probabilistic Latent Semantic Analysis (PLSA), an unsupervised learning technique was proposed, which was formulated on statistical Latent class model. The authors affirmed that their approach seems to be more of principal oriented than the Conventional Latent Semantic Analysis (LSA), as it possess a strong statistical foundation which adopts Annealed Likelihood function as its optimization criterion [4-7]. The privileges of this PLSA is considered as a promising and productive unsupervised learning method, which covers a wide spectrum of applications with respect to text learning [8].The authors stated the employment of semantic features in the twitter sentiment classification. They explored three other approaches for assimilating the collected tweets for effective analysis. These approaches include replacement, augmentation and interpolation [9]. Replacement includes replacing the words with meaningful words, deleting the unnecessary words. Augmentation simply means adding. Approaches in augmentation include adding noise and applying transformations on existing data. In sparse areas imputation and dimensional reduction are also used for augmentation in the data sets. Interpolation is a process of drawing new data points from the existing range of known data points. Mainly interpolation helped the model to achieve best results by interpolating the generative words into unigram language model of Naïve Bayes (NB) classifier [10].

A new approach to sentimental analysis was introduced, which uses support vector machines (SVM). Mainly, this SVM is used to bring together potentially pertinent information from different sources [11-13]. This also includes various favourability measures for phases and adjectives related to topic of the text in the tweet. Merits of this approach includes the incorporation of various words (with the help of SVM) where, previously it was limited to the specific words that are present in the tweets. Due to this incorporation of words from various sources, efficiency of the model was declined.

III. Proposed System

Initially data has been collected in the first step. Later on it is pre-processed. The pre-processed data is used for getting valuable insights through different visualization techniques. Finally clusters are formed and the most affective cluster is identified (Fig. 1).

Revised Manuscript Received on September 15, 2019

Dr. P.Asha, Asst. Prof.,Dept. of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai. ashapandian225@gmail.com

K. Sri Neeharika, Asst. Prof.,Dept. of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai. neeharikasri6@gmail.com

T. Sindhura, Asst. Prof.,Dept. of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai. sindhura2397@gmail.com

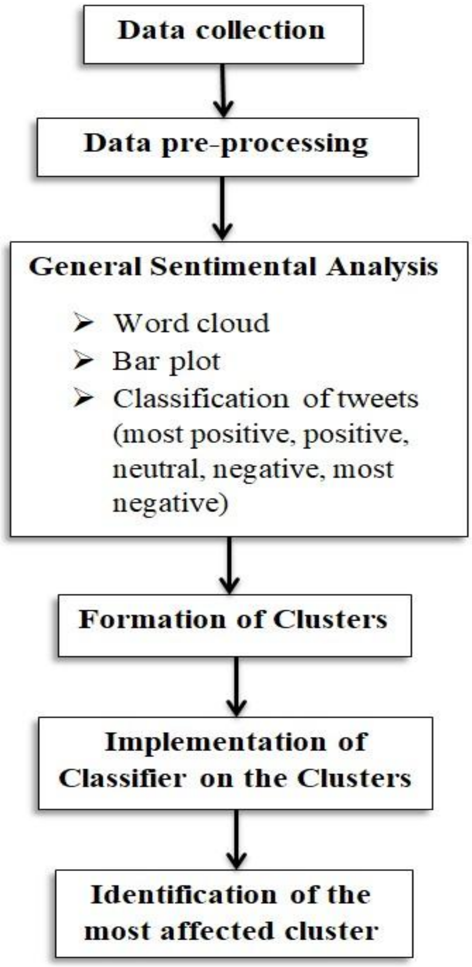
Data collection
Data pre-processing
General Sentimental Analysis
➢ Word cloud
➢ Bar plot
➢ Classification of tweets (most positive, positive, neutral, negative, most negative)
Formation of Clusters
Implementation of Classifier on the Clusters
Identification of the most affected cluster

Fig 1. System Architecture

The input data has been collected from twitter using various hashtags (#Metoo, #politics, #Education, #work) using twitter consumer key, API, secret key. All the extracted tweets are stored in a .csv format. With the help of hashtags like politics, education and work abuses in those fields are identified and data will be stored under different sectors which helps in the formation of clusters. Storage of tweets are done because extraction of tweets depends on the number of people tweeting using a particular hashtag. So if we take it dynamically sometimes the number tweets can be low. To avoid such constraints the required input data is stored in csv format. Two dictionaries are formed with a catalogue of positive and negative words in it.

Data pre-processing is to be done to the collected input as it contains so much of noise. Here noise includes like punctuations marks, numbers, stop words, tags, URL'S, un-parliamentary language, missing end marks, splitting the sentence into words. Words like RT, CRT, amp, thi, CrT are also removed. These words are present at the beginning of each tweet. So all the above stated things are removed during pre-processing.

A. Topic Modelling

Topic modelling groups the similar words into one cluster which helps to identify the hidden patterns in it.

B. Emoticon Analysis

Emoticon analysis is used to calculate the reaction of the tweeting person. In this analysis we replaced emoticons to a suitable word, so that we can take the emoticon into consideration while categorizing the tweet.

C. Gender Classification

Twitter doesn't disclose the gender of the tweeting purpose. But we can find out the gender of the user through their usernames. This can be achieved using traditional dictionary libraries.

D. Visualization

Generally visualizations are used to understand the reactions of the people in an easy manner. Various types of visualizations helps us to recognize any hidden semantic patterns in a precise manner.

Some of the visualizations used in the project are

1. Bar plot
2. Histogram
3. Word cloud

IV. Results and Discussion

Calculation of the score of the tweet is an important step for this analysis because this distinguishes the type of tweet and further classifies it. This classification is done with the help of the dictionaries which includes a wide range of positive and negative words in it. Hence all the extracted tweets including the cluster data undergoes this process. Calculation of the score of a tweet involves the number of positive and negative words in it. The ultimate score will be the difference of positive and negative words. Depending upon the final score, that particular tweet is further categorized into any one of the 5 categories. These categories contain headers like most positive, positive, neutral, negative, most negative which is shown in figure 1. This is achieved by using laply function. Hence the above stated process is done to the cleansed data obtained after pre-processing.

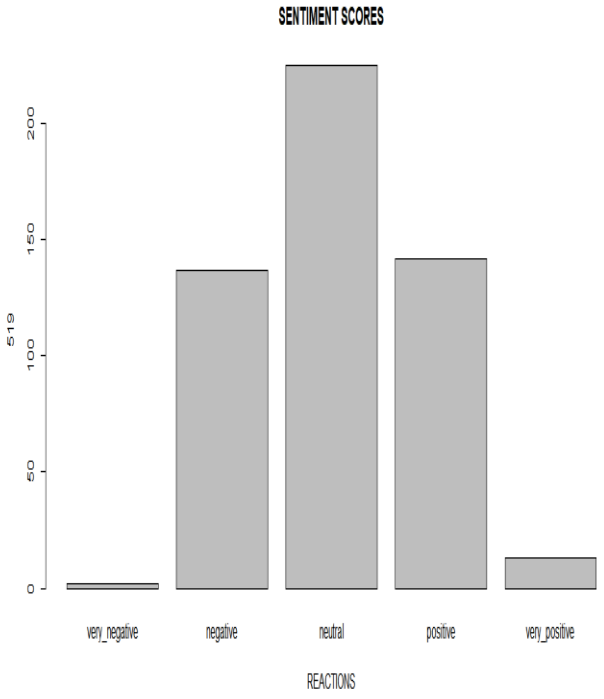

Fig 2. Classification of tweets of #MeToo

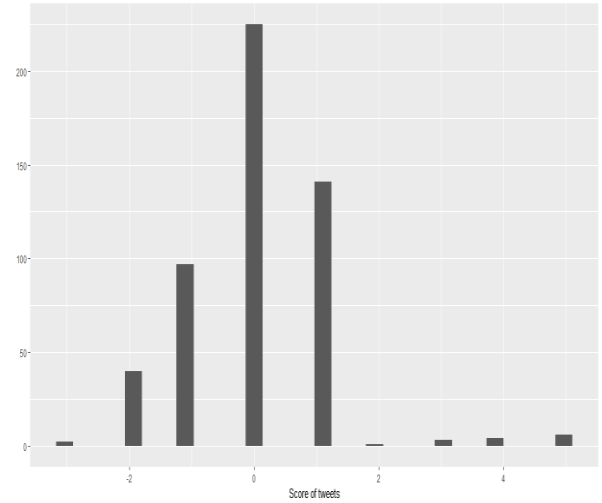

Fig 3. Scores of the tweets

A. Identification of the Most- Affected Cluster

Sentiment scores calculated are grouped into negative and positive. Scores with a negative number falls under negative category and score with a positive number falls under positive category. So sentiment scores are calculated for the clusters and all the negative categories are compared and a bar plot is drawn to identify the most-affected cluster among the three of the clusters.

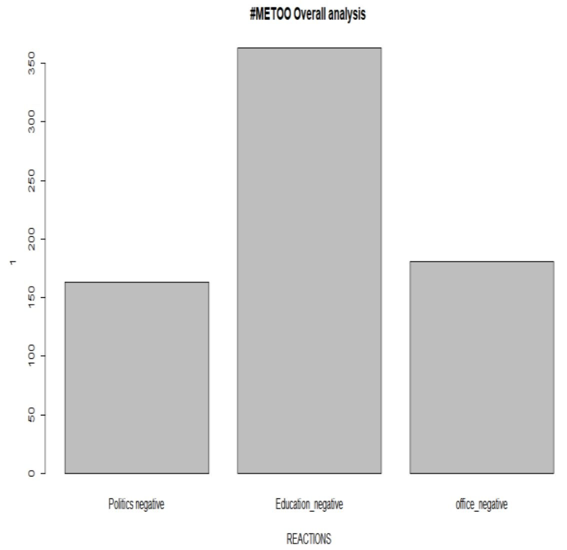

Fig 4. Identification of the most-affected cluster

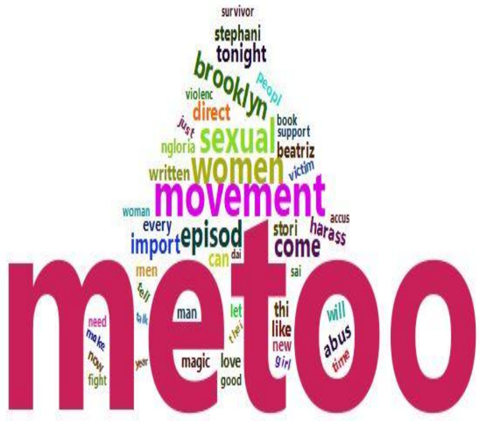

Fig 5. Word cloud representing the most frequent words in tweets

This is a word cloud that displays the most frequent words used in the tweets. The bigger the word is the bigger its size (occurrences) in the cloud. The specialty of this cloud is that if we hover on any of the word in the cloud it displays the frequency of that.

V. CONCLUSION

Data is collected from twitter and is not limited to a single platform. It can be collected from any social media platform, but the collected data should be accurate. Hence the collected data is pre-processed. Many pre-processing techniques are done to the data such that, the factual data is supplied as an input to the process. The pre-processed data is visualized to get valuable insights from it. Visualizations include bar plot, word clouds etc. Hence, the proposed strategy classifies the tweets based on the sentiment scores into 5 different categories. For further analysis they are classified as Positive and Negative tweets, which boosts up sentiment analysis which assists in identifying the most affected sector.

REFERENCES

1. J. Yi, T. Nasukawa, R.B., Niblack, W.: Sentiment analyser: Extracting sentiments about a given topic using natural language processing techniques. In: 3rd IEEE Conf. on Data Mining (ICDM'03). (2003)
2. Lloyd, L., Kechagias, D., Skiena, S.: Lydia: A system for large-scale news analysis. In: String Processing and Information Retrieval (SPIRE 2005). Volume Lecture Notes in Computer Science, 3772. (2005) 161–166
3. Andreevskaia, A., Bergler, S.: Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In: EACL. (2006)
4. Mehler, A., Bao, Y., Li, X., Wang, Y., Skiena, S.: Spatial analysis of news sources. IEEE Trans. Visualization and Computer Graphics 12 (2006).
5. Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, Ming Zhang," Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach", ACM, CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK, 2011.
6. Asha P., Albert Mayan J., Canessane A. (2018) Efficient Mining of Positive and Negative Itemsets Using K-Means Clustering to Access the Risk of Cancer Patients. Soft Computing Systems. ICSCS 2018. Communications in Computer and Information Science, vol 837. Springer, Singapore
7. Andrius Mudinas, Dell Zhang, Mark Levene," Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis", WISDOM' 12, August 12, 2012, Beijing, China Copyright 2012, ACM.
8. Asha, P., Jebarajan, T.: SOTARM: size of transaction based association rule mining agorithm. Turk. J. Electr. Eng. Comput. Sci. 25(1), 278–291 (2017)
9. Asha, P., Srinivasan, S.: Analyzing the associations between infected genes using data mining techniques. Int. J. Data Min. Bioinform. **15**(3), 250–271 (2016)
10. Chenghua Lin, Yulan He, Richard Everson, Member, IEEE, and Stefan Ru¨ger," Weakly Supervised Joint Sentiment-Topic Detection from Text", IEEE trans. on knowledge and data engineering, vol. 24, no. 6, June 2012.
11. Xiaohui Yu,Yang Liu, Jimmy Xiangji Huang, and Aijun An,," Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain", IEEE Trans.On knowledge and data engineering,vol. 24, no. 4, April 2012.
12. Danushka Bollegala, David Weir, and John Carroll," Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", IEEE trans. on knowledge and data engineering, vol. 25, no. 8, August 2013.
13. Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan,"Mining Social Media Data for Understanding Students' Learning Experiences" IEEE trans. on learning technologies, vol. 7, no. 3, July- September 2014.