An Unsupervised Clustering Approach for Twitter Sentimental Analysis: A Case Study for George Floyd Incident

Balaji Karumanchi

[#]*Sr Software Engineer & Natsoft Corporation*
6804 Prompton Bnd, Irving, Texas, USA

Abstract *- Performing sentiment analysis is vital which can be used to find out the public review about a product or ongoing events in the world. Public can easily and efficiently express their perspectives and ideas on a wide variety of topics like events, services and brands via social networking websites. Social networks especially Twitter is continuously updated with public views, expressions and opinions. In this we have performed twitter sentimental analysis to review public opinion about George Floyd incident using Twitter data. Text mining and sentimental analysis are used Text mining and sentiment analysis to analyse unstructured tweet text to extract positive and negative polarity about this incident. Moreover, tweet frequency analysis has been done to view trend in public opinion across 9 days' tweet text data. We found out that majority of the people have attitude towards this incident by using 3 hashtags and overall data.*

Keywords — *George Floyd, twitter sentimental analysis, K-Means clustering, data mining*

I. INTRODUCTION

Catastrophic national events have a ground effect on the lives of national citizens or even on people around the globe. The consequences of such a scenario are experienced in diverse and multiple ways. The outcome and reaction of general public can be felt through media, news reports and social media. One such footprint has been left since the murder of George Floyd on 25 Jun, 2020. Since his death, protests have flared across the country demanding the arrest of the officers involved in the killing and for systemic change to put an end to police brutality. There has also been overwhelming response on social media.

One of the best mechanisms to capture human views and emotions is to analyse the content they post on social media. People's opinions are always an important piece of information for businesses and other people to make them aware of the current trends. Since the introduction of World Wide Web, the world has become a global village. According to [1], approximately 4.57 billion people have access to internet which makes up more than 57% of world's population. And out of these 4.57 billion people 3.81 billion use internet for social media. Hence public opinions on social media are best source to know about a person's mood and opinion about a product or an event [2].

This paper is an effort to analyses the current trends and reaction of people towards this incident using statistical analysis through Twitter data. Here we have proposed two methods to achieve our goal; one is through tweet frequency analysis and second sentiment analysis using unsupervised learning to learn about polarity of tweets.

In the first proposed method we are going to see the trends in top 3 hashtags selected from downloaded tweets. Based on these hashtags we will give a strong idea about the reaction of people. This method is also vital in pre-processing of data before passing it for sentimental analysis. Sentiment analysis is a technique that can be used to gauge the polarity of a writing. It can help analyse the attitudes in a text related to a particular event or subject. A mathematical model is created to know about people's opinions and expressions. In [3], researchers have affectively used for political predictions, marketing strategy, e-commerce, and brand reputation management. Since we have used unlabelled data we will be using unsupervised sentimental analysis techniques to derive results and opinions.

The remaining paper is structured as follows: In section 2 we discuss the related work on sentimental analysis. Then in section 3 we discuss about methodology which includes discussions on dataset curation, pre-processing and brief introduction of model. In section 4 we discuss about the results. And final section concludes the paper.

II. RECENT WORK

With the introduction of World Wide Web and Web2.0 technology, we can see a sudden surge in consumer voices and public opinion over social media. One of the resulting emerging field is sentiment analysis [4]. Natural Language Processing (NLP) is being widely used in opinion mining. A review of existing methods on opinion mining and sentimental analysis is done by Pang and Lee [5].

The power of social media marketing is influencing the consumer and companies as well by

spreading useful information and exchange of positive or negative values. Companies are learning the customer views and discussion to support their own mission and performance goals. In [6], Rathod et. al. have employed weightage classification model based on self-learning model to study public opinion about smart phone products. Based on words from tweet they categorize the tweet to be positive, negative or neutral.

In [7], the authors used hashtags, URLs and emoticons to create new tweet specific features. Becker et al. [8] proposed an online clustering framework to identify different types of real-world events and their associated social media documents. This technique can categorize alike events and non-events. Bhuvan et al. [9] proposed sentimental model using naïve-based algorithm to classify the polarity of trained dataset and to validate the model to get the percentage for three categories like positive, negative, or neutral for the automotive industry.

Several methods have been deployed for sentimental analysis including work based on Support Vector Machines [10], Naïve Bayes [11] and K-Means clustering [12]. In [13], the authors have reviewed several works for Twitter data analysis using machine learning techniques and Naïve Bayes classifier for public opinion extraction. In a more recent work [14], the authors have used graphs using the Clauset-Newman-Moore algorithm to create clusters and groupings.

Latest advent of deep learning has also been leveraged in sentiment analysis by [15]. They provide a comparison between traditional machine learning methods, polarity based methods and deep learning methods. They employed various datasets to train deep learning models hence covering a wide variety of tweets. Their results show that deep learning methods can provide up to 97% accuracy as compared to machine learning model which achieved a maximum of 84 % accuracy. The problem with deep learning methods (LSTM, CNN) is that they require a large dataset, very high computation power and take time to train. Hence for specific type of sentimental analysis traditional deep learning methods are recommended to be used.

III. METHODOLOGY

The proposed methodology for sentimental analysis is shown in Fig. 1. Each step is explained below in details.

A. Dataset

The dataset is curated using scrapper built with Python. We retrieved 8,86,579 tweets by using 5 hashtags (blacklivesmatter, georgefloyd, icantbreathe, riots) over a span of 11 days from 25 May, 2020 to 4 Jun, 2020. The columns in dataset include username, tweet text, date of tweet and link to the tweet. The dataset is further divided into four sub-datasets which are created using top 2 hashtag (blacklivesmatter and goergefloyd). Whereas to give insight about public reaction to protests, tweets related to protest are separated for tweet frequency analysis.

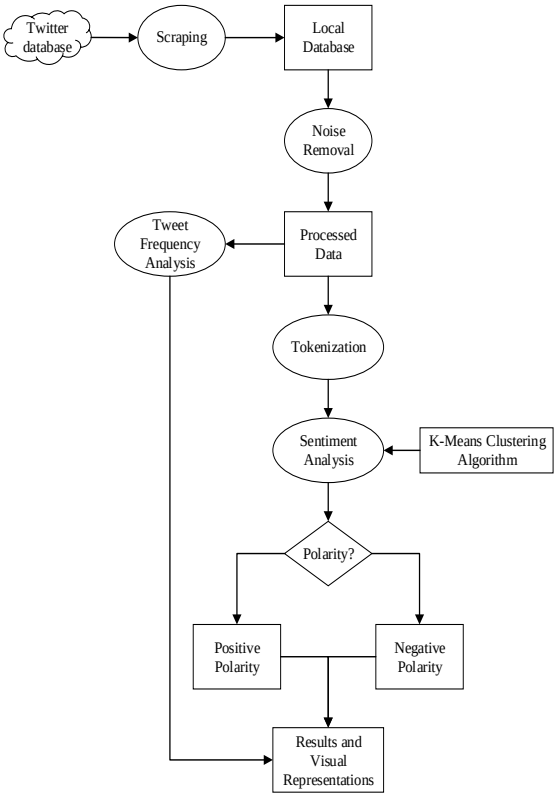

Fig. 1 Proposed methodology

B. Pre-Processing

The text data from tweets is raw in its original form. The text is filled with noise data like punctuation marks, emoji text, removal of duplicate data, hashtags, mentions and links. So it is necessary to clean the data to make it feasible for next stage. Python's regular expression library (re) and Natural Language Toolkit (nltk) have been used for this purpose. Moreover, subsets are also created for tweet frequency analysis based on 3 hashtags. During this stage the tweet feature vector using tokenization to create unigrams is also created which is passed to the clustering algorithm to perform sentimental analysis.

C. Tweet Frequency Analysis

In this stage we analyse the trend of public reaction using all data and three hashtags. The tweeter feature vector and dates are used to create groups by date. The number of dates in each group are created and bar graphs are created. This gives insight into public trend towards this incident.

D. Sentimental Analysis

We employed un-supervised machine learning algorithm called K-Means clustering [16] for the clustering of polarity of each tweet for the sentiment analysis. This seemed most suitable for the given problem, since the data is unlabelled and it is not

possible to manually annotate 800k+ tweets. It works by taking an input number N of necessary clusters, and outputs coordinates of calculated central points of discovered clusters. Basically the algorithm aims at minimizing the objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left| x_i - v_j \right| \right)^2$$

Where, V corresponds to the current cluster, xi is the center of current cluster and vj is the position of data points. It is an iterative algorithm, in which in first step N random data points are chosen as coordinates for center of clusters. In each step all using Euclidean distance points are assigned to their closest centroid. Then new coordinates of centroids are calculated by taking mean of coordinates of all the data points in that cluster. These steps are iterated till until a minimum value of mean squared error between points assigned to centroids is achieved.

Since the output of K-Means algorithm is distance of data points from centroids, so we have to convert it to polarity score. For this purpose, the distance is multiplied by inverse of closeness score. After this step we will have a dictionary containing a word and weighted sentimental score. Later then we calculated the term frequency–inverse document frequency score (tfidf). This is a numerical statistic which points out how much important a word to a sentence/document and is used as a weighing vector for information retrieval or text mining. To achieve this, we used sklearn library. After this step we have 2 vectors for each sentence; one vector containing weighted sentimental score and other one tfidf score. Finally, these 2 vectors are multiplied to achieve the final polarity of sentence being positive or negative.

IV. RESULTS AND DISICUSSIONS

A. *Tweet Frequency Analysis*

The analysis for whole data is given in Fig. 2. It shows number of tweets per day spanning over a period of 9 days. It can be seen from the figure that the first tweet arises on early 26 May, 2020. And it starts rising till 28 May, 2020. Later it remains almost constant and then start decreasing from 02 Jun, 2020.

The tweet frequency analysis for hashtag George Floyd is shown in Fig. 3. A similar trend as seen in whole data plot can be seen here. The plot first increases to 25 May, 2020, then remains almost constant till 04 Jun, 2020 and then starts decreasing. The tweet frequency analysis for hashtag

The tweet frequency analysis for hashtag protest are shown in Fig. 5. We can see that during the days when protests were on peak we can see a similar trend in tweet data also. The protests were on its peak during these days. Then we can see a drop in data as there were less protests during that time.

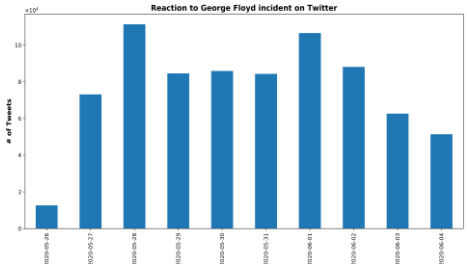

Fig. 2 Tweet Frequency Analysis (all data)

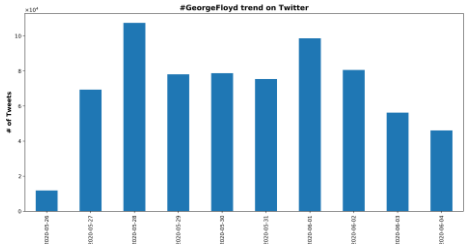

Fig. 3 Tweet Frequency Analysis (#GeorgeFloyd)

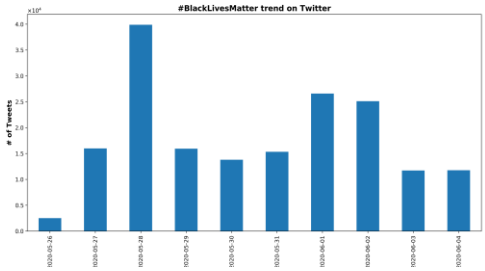

Fig. 2 Tweet Frequency Analysis (#BlackLivesMatter)

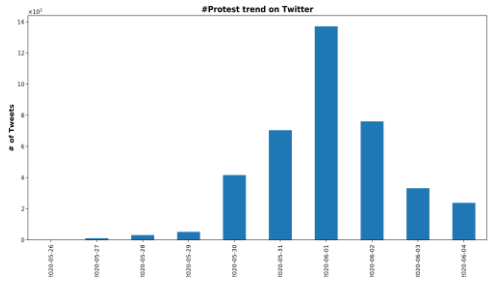

Fig. 3 Tweet Frequency Analysis (#protest)

B. *Sentimental Analysis*

After the implementation of sentiment classification, we got the values of sentiment distribution of negative and positive polarities. Each tweet is analyzed to be positive or negative tweet based on a query term and polarity clustering. Some tweet samples of positive and negative polarities are given in Table I.

Looking at the table, we can conclude that tweets having harsh words or too much anger are marked as negative whereas tweets with good words or very less harsh words are marked as positive. Moreover, it is obvious from Fig. 6, Fig. 7 and Fig. 8 that majority of the people have a positive attitude towards this matter which means that majority of the tweets do not show hatred but still want justice for George Floyd and have a positive attitude towards this incident.

Table I. Tweet Samples and their Polarity

Tweet	Polarity
can you imagine feeling so empowered, that you allow yourself to be recorded while you commit murder in broad day light infront of several witnesses??? #blacklivesmatter	Positive
rest in peace # icantbreathe #georgefloyd	Positive
it is heartbreaking & terrifying living in a country where i wouldn't call the police if i needed help, in fear that someone in my family could be wrongfully killed. #minneapolis #philandocastille #centralparkkaren #minneapolispolice #blacklivesmatter #icantbreathe	Negative
im so at a loss for words. this shit is just so hard to watch. #georgefloyd #rip! i hope all them cops a painful slow death fr	Negative

It can be seen that for all data we have 72.2 % positive and 27.3 % negative tweets. For #GeorgeFloyd we can see that 67.9% percent people have positive and 32.1 % people have negative polarity. Finally, for #BlackLivesMatter it can be seen that 70% people have positive and 30 % people have negative attitude.

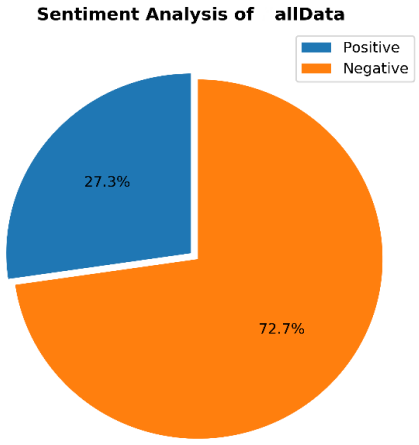

Fig. 4 Sentiment distribution of all tweets

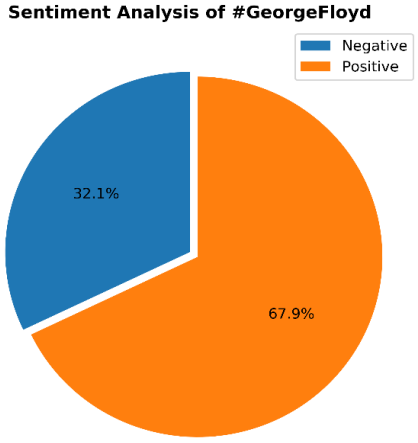

Fig. 5 Sentiment analysis of tweets (#GeorgeFloyd)

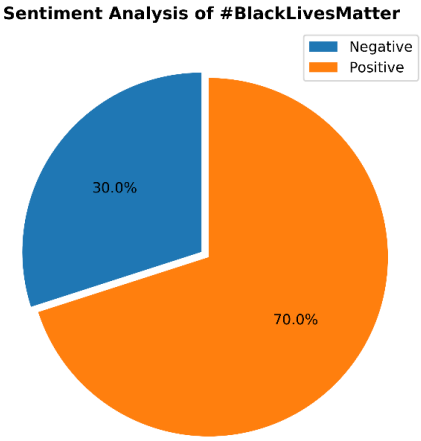

Fig. 6 Sentiment distribution of tweets (#BlackLivesMatter)

V. CONCLUSIONS

Sentiment analysis is one of the most attractive and shinning field text and data mining with vast applications in numerous sectors. In this paper, we have shown how a dataset can be curated, its pre-processing stages. Then we have done Tweet Frequency Analysis to monitor the trend towards this incident. Then we did sentimental analysis using K-Means clustering algorithm to know about the polarity of tweets i.e. positive or negative. The results showed that the twitter response after 2 days of incident was at its peak then it declined slowly. Moreover, we have shown that on average 70.2 % people have a positive attitude towards this incident. The polarity results were consistent in all three samples of dataset.

In future work, these results can be further improved by labelling the dataset. Then using this labelled dataset, a comparison between different machine learning algorithms can be done. Moreover, deep learning algorithms like CNN. LSTM or RNN can also be employed for improvement in polarity clustering.

REFERENCES

[1] J. Clement, "*Global digital population as of April 2020*," Statistica.com, 2020, [Online]. Available: https://www.statista.com/statistics/617136/digital-population-worldwide/, Accessed on: Jun 10, 2020.

[2] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "*New avenues in opinion mining and sentiment analysis*," *IEEE Intelligent systems*, vol. 28, no. 2, pp. 15-21, 2013.

[3] B. Liu, "*Sentiment analysis and opinion mining,*" Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1-167, 2012.

[4] N. Majumder *et al.*, "*Improving Aspect-Level Sentiment Analysis with Aspect Extraction*," *arXiv preprint arXiv:2005.06607*, 2020.

[5] B. Pang and L. Lee, "*Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval*," 2008.

[6] J. A. Rathod, S. Vignesh, and A. J. Shetty, "*Sentiment Analysis of Smartphone Product Reviews Using Weightage Calculation*," in Advances in Computing and Intelligent Systems: Springer, 2020, pp. 427-437.

[7] A. Ritter, S. Clark, and O. Etzioni, "*Named entity recognition in tweets: an experimental study*," in *Proceedings of the conference on empirical methods in natural language processing*, 2011: Association for Computational Linguistics, pp. 1524-1534.

[8] H. Becker, M. Naaman, and L. Gravano, "*Beyond trending topics: Real-world event identification on twitter,*" in Fifth international AAAI conference on weblogs and social media, 2011.

[9] Malladihalli S Bhuvan , Vinay D Rao , Siddharth Jain , T S Ashwin , and R. M. R. Guddeti, "*Semantic sentiment analysis using context specific grammar*," presented at the International Conference On Computing, Communication and Automation., 2015.

[10] C. C. Chen and Y.-D. Tseng, "*Quality evaluation of product reviews using an information quality framework*," Decision Support Systems, vol. 50, no. 4, pp. 755-768, 2011.

[11] M. Gayathri, S. S. Nisha, and M. M. Sathik, "*Twitter Sentiment Analysis using Naive Bayes Classification*," Studies in Indian Place Names, vol. 40, no. 71, pp. 1473-1478, 2020.

[12] H. Suresh, "*An unsupervised fuzzy clustering method for twitter sentiment analysis*," in 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2016: IEEE, pp. 80-85.

[13] T. Vaseeharan and A. Aponso, "*Review On Sentiment Analysis of Twitter Posts About News Headlines Using Machine Learning Approaches and Naïve Bayes Classifier,*" in Proceedings of the 2020 12th International Conference on Computer and Automation Engineering, 2020, pp. 33-37.

[14] W. Ahmed, J. Vidal-Alaball, J. Downing, and F. L. Seguí, *"COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data*," *Journal of Medical Internet Research*, vol. 22, no. 5, p. e19458, 2020.

[15] Y. Chandra and A. Jana, "*Sentiment Analysis using Machine Learning and Deep Learning,*" in *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2020: IEEE, pp. 1-4.

[16] J. MacQueen, "*Some methods for classification and sanalysis of multivariate observations,*" in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. 1, no. 14: Oakland, CA, USA, pp. 281-297.

[17] B.Srinivasa Rao, S.Vellusamy Raddy, "*A Hard K-Means Clustering Techniques for Information Retrieval from Search Engine*" SSRG International Journal of Computer Science and Engineering 4.2 (2017)