# *Sentiment Analysis on real-time Twitter data for Epidemics using ELK*

Sathyanarayanan Govindarajan
Department of Computer Science
California State University
Northridge, California,USA

Mayur Rahangdale
Department of Computer Science
California State University
Northridge, California,USA

***Abstract*— Twitter is a free social networking and microblogging service that enables its millions of users to send and read each other's 'tweets'. A recent analysis of the "Twitter stream" revealed that a substantial proportion of tweets contain general chatter, user-to-user conversations only of interest to the parties involved, links to interesting pieces of news content, or spam and self-promotion. Despite the high level of noise, the Twitter stream does contain useful information. Many recent news events have been documented via Twitter directly from users at the site in real time: examples include US Airways flight 1549 landing in the Hudson River, or street riots during Iran's 2009 presidential elections. Because tweets are often sent from handheld platforms on location, they convey more immediately than other social networking systems. These examples suggest that useful information about news and geopolitical events lies embedded in the Twitter stream. Although the Twitter stream contains much useless chatter, by virtue of the sheer number of tweets, it will still contain enough useful information for tracking or even forecasting behavior when extracted in an appropriate manner. For example, Twitter data has been used to measure political opinion, to measure public anxiety related to stock market prices, national sentiment, and to monitor the impact of earthquake effects. In this study, we examine the use of information embedded in the Twitter stream to track rapidly evolving public sentiment with respect to Covid-19 or Coronavirus, and track and measure actual disease activity. Also,it will attempt to design a configurable dashboard to extend the solution to report other pandemics.**

***Keywords—covid19, coronavirus, pandemic, twitter, data-analysis, machine-learning , elk-stack, data-visualisation***

## I. Introduction

There is always a need for powerful assessment and analysis tools for global pandemics. Humankind has faced a lot of challenges tackling such outbreaks in recent years. The goal of our research is to build a similar tool that could benefit humankind from such catastrophes in the coming future. Our proposal involves building a real-time geographical mapping of COVID-19 and their outbreaks, using Twitter data. ELK (Elasticsearch, Logstash, Kibana) stack is used for Data Ingestion, querying and visualization. We have proposed an additional ML pipeline for Sentiment Analysis of target twitter data.

To verify the authenticity of our proposal, we identify correlation between the trends of twitter batch data analysis and the historical data provided by WHO. A significant amount of correlation between the above two implies the correctness of our hypothesis and authenticity of this proposal.

The significance of such a tool is

a) Understanding the crisis event related questions posted by the public on social media.
b) Determination of the outbreak faster and helps to react to the situation.
c) Medical and public welfare preparation after earlier assessment of the outbreak.

## II. Datasets

### *A. Data Acquisition*

The data used in our experiment comes from 3 different sources. For the real time analysis, live tweets are taken from twitter, using the Twitter-API. The Scientific historical data is taken from the WHO (World Health Organisation) dataset called 'Coronavirus (COVID-19) Cases and Deaths' [1]. The batch twitter data used in our exploratory data analysis is taken from a Kaggle dataset published by Gabriel Preda [2].

### *B. Data Description*

The Kaggle Dataset has attributes defining user tweets and descriptions based on location. The WHO dataset involves fields such as new cases and deaths based on location. Table I show fields present in real time data from twitter.

TABLE I. TWEET ATTRIBUTES

| Attributes | Description |
|---|---|
| date | a datetime object in the form of YYYY-MM-DD HH:MM:SS |
| text | the tweet itself |
| hashtags | list of hashtags used in the tweet (without '#' character) |
| source | device used for tweet |
| retweets | number of retweets received at the time the data was collected |
| favorites | number of likes received at the time the data was collected |
| is_retweet | indicates if the tweet is original or a retweet (boolean) |

## III. Technology and Implementation

### A. Technology

ELK stack is mainly used for creating the Realtime COVID sentiment analysis dashboard and Python data processing libraries are used for Exploratory data analysis for the historical tweets.

- Tweepy—Python library for Twitter API used in fetching the tweets directly from Twitter
- Textblob—Python library for text processing and sentiment analysis, which uses NLTK (Natural Language Toolkit) for natural language processing
- Elasticsearch—used for Data Mapping and Requirement-based Querying.
- Kibana—Dashboard, Comparison b/w positive and negative tweets and different kinds of Visualization.
- Python Libraries such as Pandas, NumPy, matplotlib for Exploratory data analysis.

### B. Design and Methodology

The design for our research project is divided into three separate modules. Each module has its own significance and correlates with one another.

The first module is the twitter sentiment analysis from live twitter feed. The approach to extract sentiment from tweets is as follows:

1. Import the Tweepy library to create the connection with the Twitter API and create a StreamListener instance.
2. Fetching tweets real-time using the Twitter developer authentication keys.
3. Using the Textblob to determine the tweet sentiment metrics such as polarity, subjectivity and sentiment.
4. Metrics are attached to the message and pushed to Elasticsearch.
5. Elasticsearch is used to index and store the documents.
6. Kibana is used to create a real time dashboard which describes various visualization and metrics.

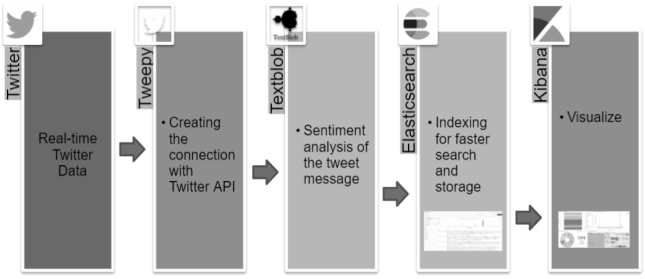


Fig.1. Sentiment Analysis Architecture

The second module involves exploratory data analysis on the batch twitter data [2]. This module is completely built upon python libraries and geospatial analysis is done based on the usage of hashtag 'COVID-19'. The third module involves time series analysis of new cases and deaths based on locations, using the WHO data [1]. ELK Stack is used for this time series implementation.

### C. Sentiment Analysis

Textblob is a python library for NLP tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.Textblob sentiment analysis processing employs two key metrics:

- *Polarity-score* is a float within the range [-1.0, 1.0] where smaller values imply negative statement and larger values implies positive statement.
- *Subjectivity-score* is a float within the range [0.0, 1.0] where smaller values imply objectivity and larger values imply subjectivity.

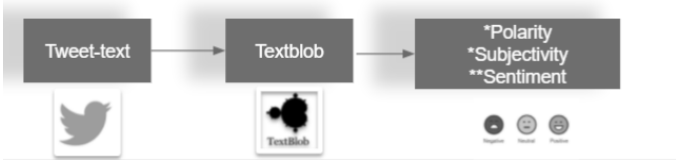


Fig.2. Sentiment Analysis Metrics Pipeline

Below are a few examples of Textblob input and outcome illustrating the classification of text into sentiment classes.

TABLE II. TWEETS WITH DIFFERENT POLARITY

| Text input to TextBlob | TextBlob.sentiment | |
|---|---|---|
| | Polarity-score | Subjectivity-score |
| "So excited to get my vaccine!" | 0.46875 | 0.75 |
| "Is the vaccine painful?" | -0.7 | 0.9 |
| "The Pfizer vaccine is now FDA approved" | 0.0 | 0.0 |

The first input has a positive polarity . Textblob recognizes a relatively positive emotional charge associated with the statement, and based on how an individual feels,the statement is very subjective. For the second input, polarity is negative. TextBlob recognizes a relatively negative emotional charge associated with the statement and based on an individual expressing themselves,the statement is very subjective. For the third input, the values of polarity and subjectivity are null. The statement is identified as neutral in polarity and highly objective.

### D. *Limitations*

The most dominant limitation faced in this experiment is the Geo Validification problem. Almost 40% of the locations used by Twitter users are either inaccurate or does not exist. This leads to noise in the data when performing geospatial analysis.To tackle this problem, a Geo-Validifier on the stream listener side can be added, that returns verified locations using Google-Maps API and coordinates of the user-tweet.

Another limitation is the sentiment analysis tool. A custom built sentiment analysis pipeline can be made, training on the historical batch data. This analyser can be extended to have better sentiment classification than pretrained language models such as textblob.

## IV. Validation and Results

After implementing all 3 modules of our experiment, the comparison was done between the twitter batch data trends in the second module and historical data trends in the third module.

While implementing the Geospatial analysis of twitter batch data, trends like most tweets based on location were studied. The most important trend here is the time stamp analysis, which helps us correlate it with the results of our third module.

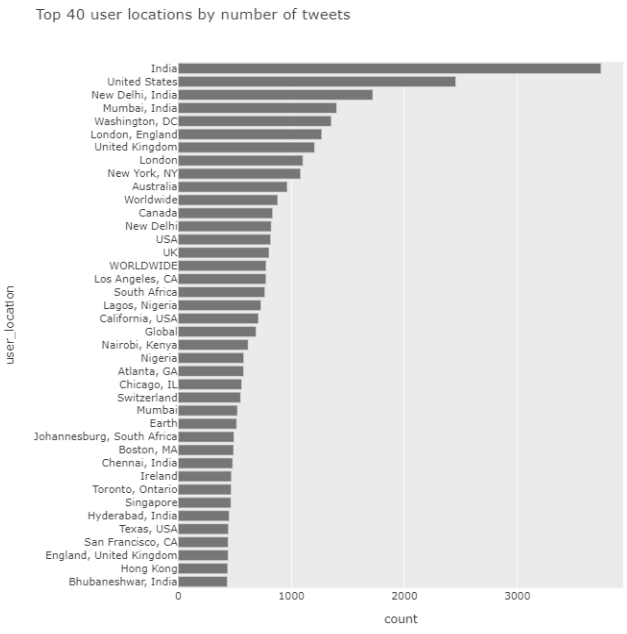


Fig.3. Top 40 User Location by number of Tweets

From figure 3 and 4,we can infer that there is a spike in the Covid-19 cases in locations India and the United States.

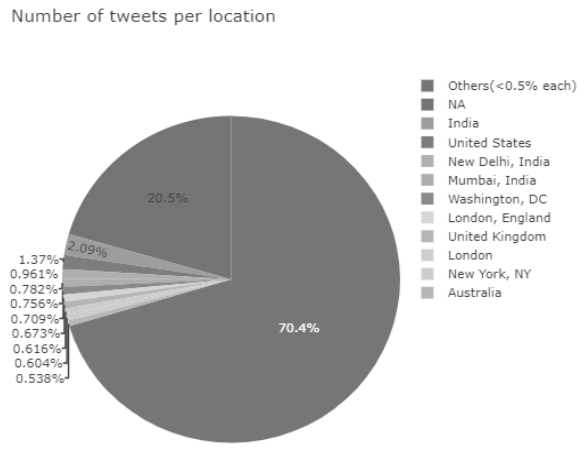


Fig.4. PieChart Showing Number of Tweets per Location

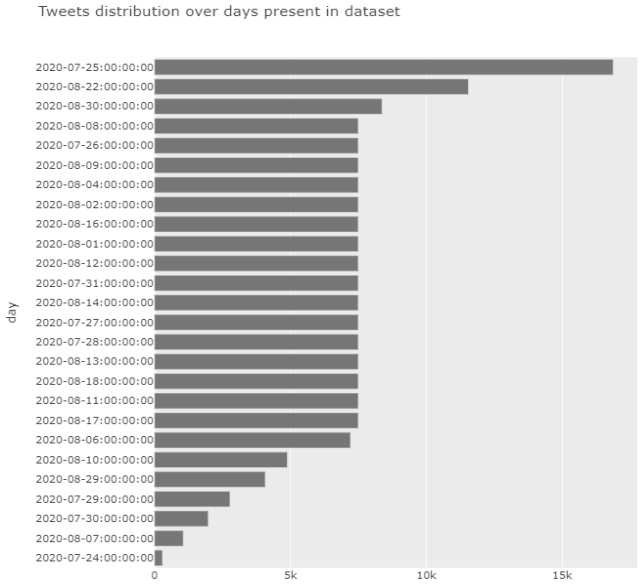


Fig.5.Plot Showing timestamp distribution of Tweets

Figure 5 shows the timestamp distribution of all the tweets in the dataset. From the plot, it can be inferred that all the tweets in the dataset are from 7th and 8th month of year 2020. Overall, from the geospatial analysis, we can infer that there is a spike in new cases in India and the United States in the month of 7th and 8th month of year 2020.

Figure 6 and 7 shows time-series analysis of historical data for locations India and the United States respectively. The curve elevates in both the plots over the span of 7th and 8th month of year 2020. This trend is very similar to what we found in the geospatial analysis.

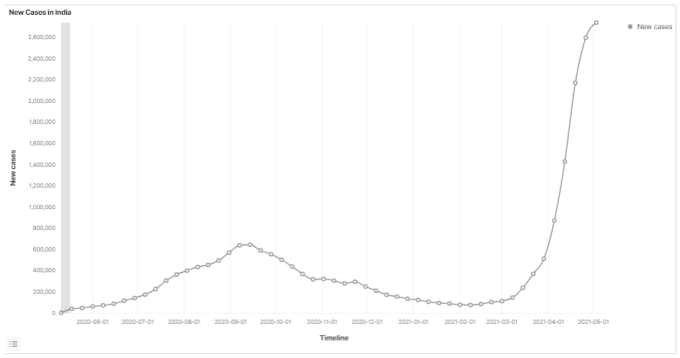


Fig.6.Time Series Plot Showing new cases in India

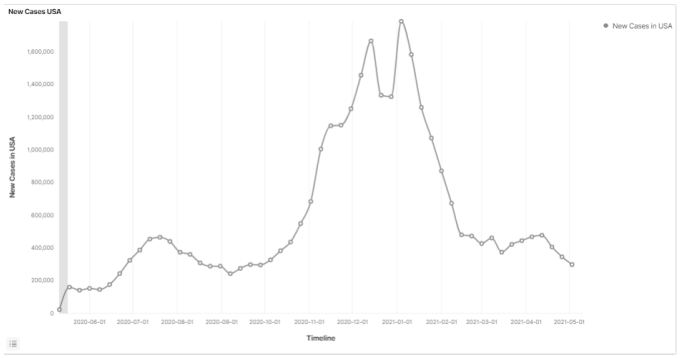


Fig.7.Time Series Plot Showing new cases in United States

## V. Conclusion

There is a major correlation between the geospatial analysis of batch twitter data and the time series analysis of historical Covid data. This significant correlation between the two advocates the authenticity of twitter data analysis. It concludes our experiment that our hypothesis to use real time sentiment analysis for epidemics can be relied upon in real world scenarios.

## Acknowledgment

This paper and the research behind it would not have been possible without the exceptional support of my professor, Dr. Senhua Yu. His enthusiasm, knowledge and exacting attention to detail have been an inspiration and kept our work on track. He motivated us to research on this topic and supported us in the journey from the research proposal to the final draft of this paper.

## References

[1] World Health Organisation 'Coronavirus (COVID-19) Cases and Deaths' https://data.humdata.org/dataset/coronavirus-covid-19-casend-deaths

[2] Kaggle Dataset 'COVID19 Tweets' Gabriel Preda, August 2020 https://www.kaggle.com/gpreda/covid19-tweets