Assignment 3

Spark Streaming

Reetish Chand Guntakal Patil (RXG190006)

Rohan Vannala (RXV190003)

Introduction

The biggest problem of the year 2020 is clearly Covid-19 (or) the Corona Virus, which is affecting the world in a lot of ways. We have taken the opportunity in this assignment to check how netizens are tweeting about “Covid-19” in either a Positive, Negative or Neutral way.

Method

There a lot of NLP libraries available, among which Stanford NLP is most widely used. Thus, Stanford NLP has been used in this assignment.

Technology

- Spark
- Scala
- Kafka
- Elasticsearch
- Kibana
- Logstash

Analysis

Visualization of data was done using Kibana. Figure 1 below, for the query string “Covid-19”shows the sentiment analysis of the tweets classified into Positive, Negative and Neutral categories. This graph is the output of a two-hour timeline from 13:30 to 15:30 depicting the positive neutral and negative responses for the chosen query string.

At the time instance shown in the figure, 14:23:38 there were –

- 557 (Negative tweets)
- 55 (Positive tweets)
- 56 (Neutral tweets)

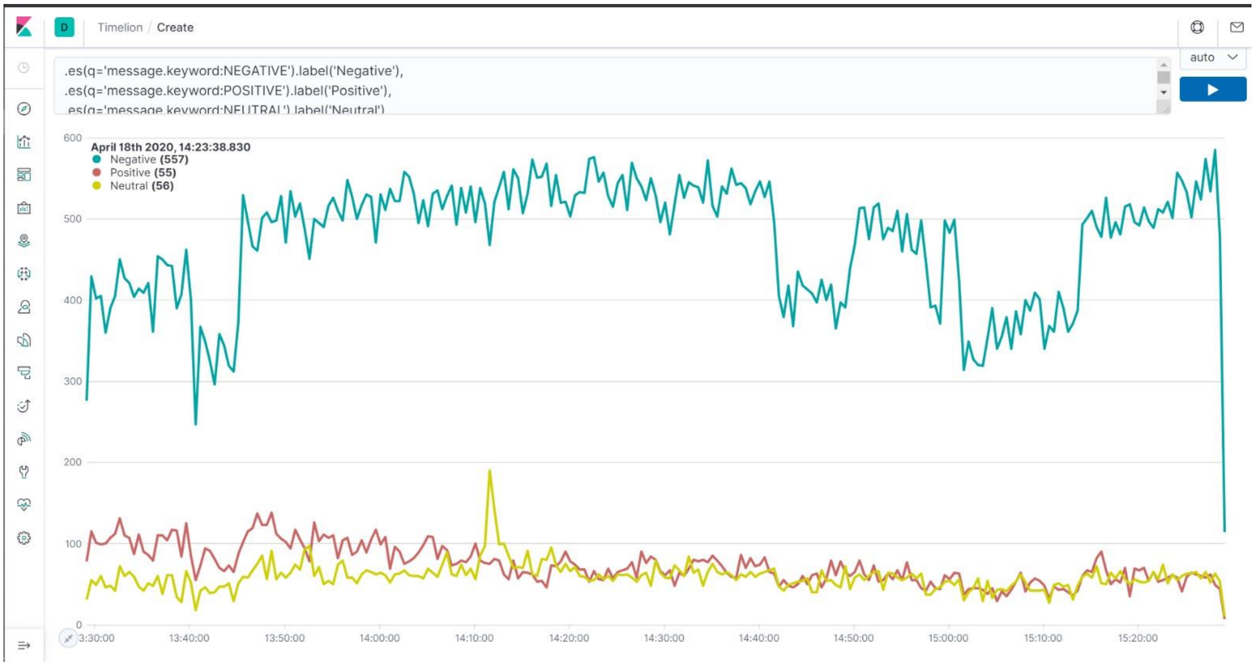
Compared to the negative tweets there were significantly smaller number of positive and neutral tweets. This clearly shows that most people have not been happy with the way things are turning out due to the virus outbreak. It might take a while to see a decrease in the negative number of tweets in the future. There are a small percentage of positive and neutral tweets may indicate the following:

- Fewer death rate in certain places
- Decline or stagnation in the number of positive tested cases per day
- Very less percentage requiring intensive care
- Progress in the vaccine testing

The vast majority of tweets at any given instance are considered negative which may indicate:

- Rapid spreading of the virus
- Decline in economy
- Business shutdown or stagnation
- Lockdowns
- Public not realizing the seriousness of the situation
- No vaccine or drug to treat the virus yet

Visualization



```
.es(q='message.keyword:NEGATIVE').label('Negative'),
.es(q='message.keyword:POSITIVE').label('Positive'),
.es(q='message.keyword:NEUTRAL').label('Neutral')
```

April 18th 2020, 14:23:38.830
● Negative (557)
● Positive (55)
● Neutral (56)

Fig 1: Time-series visualization for query string “Covid-19“

Conclusion

Overall, a large percentage of the tweets obtained for query string “Covid-19” have been negative in analysis.