Twitter Analysis Report

Abdullah McDonald *Department of Computer Sciences University of the Western Cape* [3708949@myuwc.ac.za](mailto:3708949@myuwc.ac.za)

1. INTRODUCTION

Twitter is a social media platform in which people and communities express their views on worldly matters. Because of this, Twitter has become a source of big data for companies. Tweets have been used in the various analysis in today’s world. For example, [1] was able to use machine learning to predict spikes in social tension by analyzing tweets. [2] did sentiment analysis on tweets to determine whether they had a negative, positive or neutral effect towards the use of mobile technologies, laptops, etc. [3] used Twitter data to predict people’s mood on the stock market and then used these predictions to analyze how the market fluctuates. Fascinatingly, [4] did a real-time analysis of tweets to determine whether a tweet is associated with a traffic event or not and then tried to reduce that problem.

With that said, this report details how tweets about the ongoing pandemic, coronavirus or COVID-19, can be used to determine whether the words used in these tweets were offensive or not and whether any of them were misspelt.

The structure of this report is as follows; section II explains how various technologies were set up to carry out this project, section III displays the results and gives an analysis of the results, and section IV concludes the report.

1. METHODOLOGY

Since we need access to tweets, we first need to access Twitter. This is done by using the Twitter API, Tweepy. Next, we need to set up a TCP socket with an end that sends tweets and one that receives tweets. The sending side will be accessed by the Tweepy application.

Before we can send the tweets, we first need to establish a streaming context. This was done by setting up a Spark Streaming session via a Python library, Pyspark. Once this is set up, we allow Pyspark to establish a connection to the TCP socket that has a loopback address of 127.0.0.1 and uses port 5555. Next, we allow the streaming context to create Spark RDDs every 60 seconds. Spark RDDs will contain the tweets and will allow Spark to do parallel processing of these tweets.

Now that both sides of the TCP socket have been set up, we send the tweets through the socket. Since we are only interested in tweets that contain #coronavirus and #COVID- 19, we only send these tweets. As tweets are streamed, we also saved them to a csv file as an unstructured data format. When the tweets arrived at the streaming context, all punctuation in the tweets is removed to allow for earlier processing. The tweets are then split into words and saved to an SQL table so that they can be easily queried to produce real-time analysis. To provide the results needed, we checked whether each word was offensive or not and whether each word was misspelt or not. This was done using the Python profanity-check and autocorrect libraries.

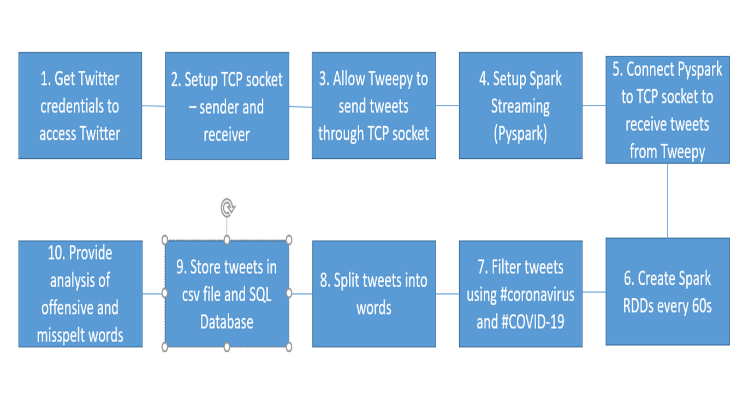


Fig. 1: Project Setup

The process of setting up this project will be explained using Figure 1.

1. RESULTS AND ANALYSIS

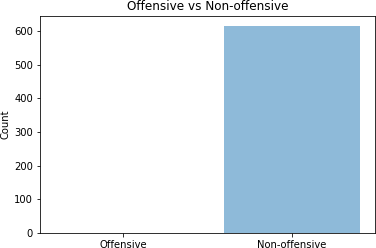


Fig. 2: Offensive vs Non-offensive Words

From Figure 2 and Table 3, we see that there were 0 offensive words and 615 non-offensive words. This clearly



Fig. 3: Offensive vs Non-offensive Word Counts

shows that not many people used offensive words when tweeting about the ongoing pandemic.

1. E. D’Andrea, P. Ducange, B. Lazzerini and F. Marcelloni, ”Real- Time Detection of Traffic From Twitter Stream Analysis,” in IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 4, pp. 2269-2283, Aug. 2015, doi: 10.1109/TITS.2015.2404431.
2. R. Li, K. H. Lei, R. Khadiwala and K. C. Chang, ”TEDAS: A Twitter-based Event Detection and Analysis System,” 2012 IEEE 28th International Conference on Data Engineering, Washington, DC, 2012, pp. 1273-1276, doi: 10.1109/ICDE.2012.125.

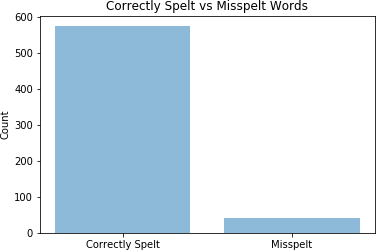


Fig. 4: Correctly Spelt vs Misspelt Words



Fig. 5: Correctly Spelt vs Misspelt Word Counts

From Figure 4 and Table 5, we see that there were only 41 misspelt words and 574 correctly spelt words. This clearly shows that people often spelt words correctly when tweeting about the coronavirus.

1. CONCLUSION

In this report, we streamed tweets from Twitter using Tweepy and Pyspark to determine whether Twitter users were using offensive words in their tweets and whether they misspelt any words in their tweets. Results showed that they never used any offensive words and quite a few words were misspelt. Improvements on this project would be to clean the data more effectively so that results can be produced faster and more efficiently.

REFERENCES

1. P. Burnap, O. F. Rana, N. Avis, M. Williams, W. Housley, A. Edwards,

J. Morgan, and L. Sloan, “Detecting tension in online communities with computational Twitter analysis,” Technological Forecasting and Social Change, vol. 95, pp. 96–108, 2015.

1. M. S. Neethu and R. Rajasree, ”Sentiment analysis in twitter using machine learning techniques,” 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, 2013, pp. 1-5, doi: 10.1109/ICCCNT.2013.6726818.
2. Mittal, Anshul, and Arpit Goel. ”Stock prediction using twitter sentiment analysis.” Standford University, CS229 (2011 <http://cs229.stanford.edu/proj2011/GoelMittalStockMarket> PredictionUsingTwitterSentimentAnalysis. pdf) (2012).