Twitter Sentiment Analysis using Spark

Abhinav Mehta
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
mehta34@uwindsor.ca

Divyesh Saraf
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
sarafd@uwindsor.ca

Shivam Dwivedi
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
dwivedi2@uwindsor.ca

Varinder Pal Babool
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
babool@uwindsor.ca

***Abstract*—With increasing demands in Data Analytics and user customization, there is a ton of data that is being collected every day and a plethora of tools to choose for analysis. Every company has their dedicated analytics team to analyze user data, reviews, and feedbacks to make their application one step ahead. Social networking sites like Twitter, Facebook and many more contains millions of data computed every second. Since twitter has an open-source API, the tweets can be analyzed to get an opinion about a specific topic by analyzing the polarity of tweets.**

In this paper, we developed an application that analyzes the tweets of people and determines the polarity of the tweets based on Lexical Approach that is discussed in the Literature review. The application determines the sentiment and trend of a particular topic that is filtered out based on hashtags and provides a report after the analysis.

Keywords—open-source API, Data Analytics, polarity, Sentiment analysis, Lexical Approach

I. INTRODUCTION

Twitter has more 350 million users and through which it gets millions of tweets every day.[1] To process this much amount of data is difficult for anyone and then also coming to any conclusion. But, by offering real-time data, social media has been transformed the way that an individual gets latest and more information. In this project, we will gather the real time data and analyze the tweets, trends or patterns, and hashtags to predict the likely outcomes.

A. Overview

We will be pulling out the real-time data from the twitter and perform sentiment analysis on it. Although, it is a difficult and much more challenging task to process this large data in database we have planned to use different approaches and at the end will be comparing end results of different approaches used in the process.

The final data will be bundled in the form of graphs and the actual data will be presented at the end of the report of project. The graphical data representation will be used as the parameter for predicting the effects. We will also be sharing our actual data on which have performed our sentiment analysis and at last the comparison of each tool used to achieve the goal of study.

B. Motivation

There is an excessive use of hashtags on many topics on twitter but considering the most recent happenings of which every single individual is aware of we have decided to go with the COVID-19 as a topic for this project. We are going to analyze the data related to Covid. With this, we can determine people's opinion and interest in the given topic and analyze the trends and patterns to predict the future happenings. Also, we can notify concerned authorities regarding the trends which can prevent problematic situations before it even creates into something big.

II. RELATED WORK

The goal of the application is to provide free analytics tool to determine the sentiments and trends related to Covid-19. This application is going to help people take inferences about the current situation of Covid-19 and can take certain steps before the actual problem begins. This application can also be helpful to the government authorities in determining the views of people regarding certain policies and mandates.

Below, we have discussed about the approaches for Twitter analysis.

A. Literature Review

The basic method of analyzing the sentiments of people using Twitter is by tracking and monitoring different datasets. This is done by first collecting twitter datasets and apply different filtering techniques to get the desired result and remove unnecessary data. Here, the unstructured data is converted to make a basic structured dataset. After this process, dataset is analyzed using different tools and we can finally come up with a report. According to the paper [2], there can be three approaches for Sentiment Analysis:

1) *Machine Learning Approach:* This uses a model which is trained by the user to detect the polarity of the tweet by training it with huge dataset. This approach takes time as proposed in the work[3], it uses classifiers to detect emotions using SVM LibLinear model. This is a more accurate model since the model is trained with a huge dataset.

2) *Lexicon-based Approach:* This approach uses a list of words annotated by polarity score to determine the opinion score of given text. This uses a dictionary that consists of positive, neutral, and negative words. The tweets can be analysed by

matching to these words to identify the tone of tweets regarding a specific topic.

3) *Hybrid-based Approach:* This approach is using a mix of both approached defined above. Combining both approaches [4] can be more beneficial and produce more accurate results. The work in paper [5] proposed a hybrid method by discussing a real-time sentiment analysis using Apache Spark's machine learning library, Hadoop distributed file system and streaming engine for sentiment prediction. The sentiment classification performance of the proposed system for offline and real-time modes were 86.77% and 80.93%, respectively.

B. *Sentiment Analysis*

It is basically a degree of people's opinion or review regarding a specific topic, a service, a product, or application, or even government. It can be divided into three common sentiment labels [6] such as positive, negative, and neutral by analysing the words in the tweets of the users. The amount of data that can be collected just by tweets is massive which needs to be analysed using big data analytics tool to get the specific information.

For example, if Apple wants to know the response regarding their new color launched for their iPhone, they can do so by analysing the tweets of users from the past 10 years and filtering it with Apple related hashtags like '#iPhone', '#colors' and more. After successfully analysing this data, the design and marketing team of Apple can sit together and decide if the customers take new colors in a positive way and if they really want some new colors for their phone. Also, what are the current marketing trends for the rival companies and if they need to work more on design. What kind of design are people more inclined to now and much more? This can help them to grow more and stay ahead of the competition.

C. *Market Study*

Currently, there are a lot of tools for Sentiment Analysis using Twitter Data sets that are being used by many companies for their products. Some of them are mentioned below [7]:

- Super metrics
- Brand watch
- Native Twitter analytics
- Social Searcher
- Brand24

III. PROPOSED MODEL

The proposed model of the application consists of three stages mainly defined as Data collection and Storage, Data analysis and processing and the last stage of data visualization. The application is free and open-source to keep the application cost down.

A. *Functional Requirement*

The functional requirements are as follows:

- The system should be able to take data from Twitter and save it to HDFS for later analysis.
- After retrieval, the system should be able to process tweets stored in the database.
- The system should be able to analyze data and categorize the polarity of each tweet.
- Calculate tweet's sentiment and store in the database, for later querying.
- The system should be able to determine the recent trend pattern among the given data. Here data can be of any topics such as COVID-19 etc.
- The system should be able to move trends and final data to cloud for further processing and to present in a graphical manner.
- Finally, plot a graph to analyze the performance of the hive and spark approaches in querying databases

B. *Proposed Model*

Below is the proposed architecture of the application.

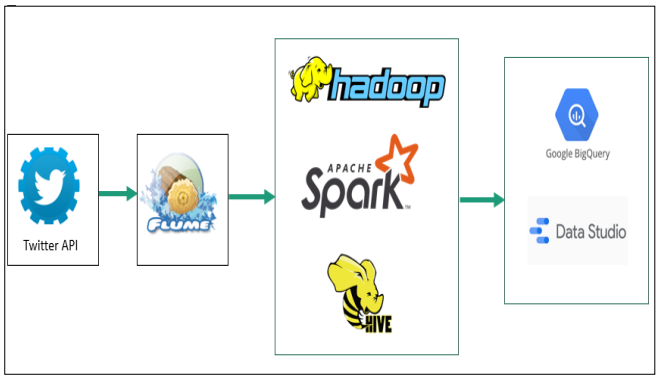


Figure 1: Proposed Architecture

The architecture consists of three phases:

1) Data Collection and Storage:

- Apache flume will stream near real time data through twitter API in JSON format and store it to the Hadoop distributed system (HDFS) in the form of blocks of files.
- Then using Hive, we will move unstructured data obtained in JSON format into structured format Hive tables for further analysis.
- Here we are using Lexicon-based Approach, which requires to have dictionary to determine the polarity and since we have a time constraint, this is the best approach so far. For this requirement, dictionary will be fetched and stored in HDFS.

2) Data Analysis and processing:

- Sentiment Analysis: We will first write multiple views to merge twitter data with dictionary words to find out tweet's polarity. The polarity is positive if tweet contain positive attitude or connotations plus the tweet has more than one sentiment. The polarity is negative if tweet contain negative attitude or connotations plus the tweet has more than one sentiment. The polarity is neutral if tweet does not contain any sentiment or emotions.
- Trends Analysis: A topic is trending when many posts with a specific hashtag are shared or posted in a short period of time. As a result, we base our analysis on recent data. To do this, the unstructured

data exported by Flume into HDFS is directly read and analyzed using PySpark and Apache Spark to represent the current trending COVID-19 subjects.

- Performance Comparison: Performance analysis includes comparing the hive and spark techniques to see which one processes data faster. We store the time taken by hive as well as the time taken by spark to process the same amount of data. Finally, create a graph based on this data to identify which one performance is better.

3) Data visualization

- Google big query: The final data of sentiment analysis and trends analysis will be moved from local HDFS to Google cloud BigQuery. This will allow any end user to access the data and run their model on the top of it.

- Visualizing in Data Studio: Once the data is moved to BigQuery, we will connect BigQuery to Google Data Studio to present data analysis in graphical format.

IV. RESULTS

We have analyzed the tweets in Virtual sandbox using a set of queries in sql and PySpark. These command needs to run in a particular order as defined in the sql file for data analysis.

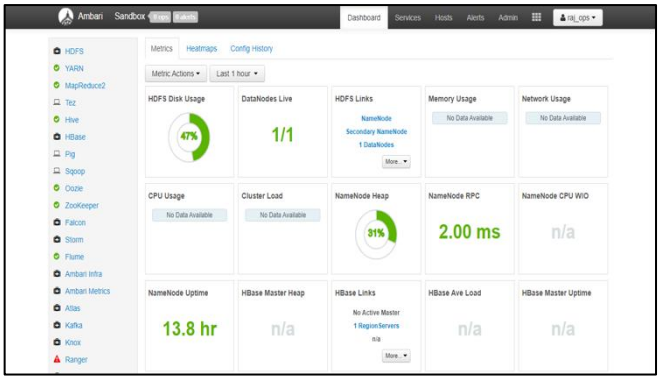


Figure 2: Virtual Sandbox

After this, we moved the data on Google Cloud Platform for further analysis. Here we can run queries on the data that we have moved from Hadoop Filesystem.

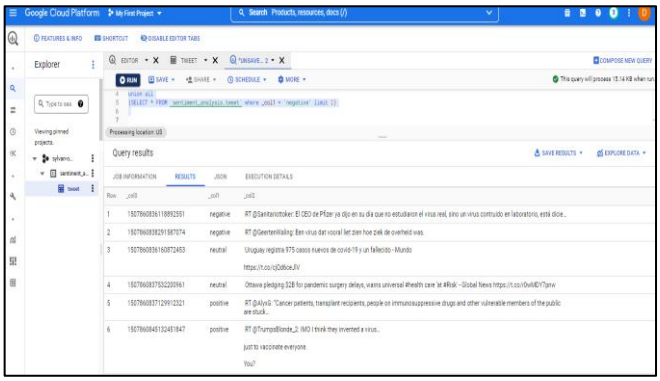
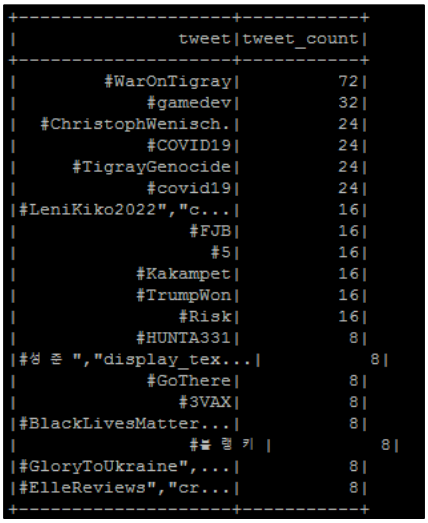


Figure 3: Google Cloud Platform

We also did Trend analysis on the data set in Hive shell as well as in Spark. Below is the result for that.



tweet	tweet_count
#WarOnTigray	72
#gamedev	32
#ChristophWenisch.	24
#COVID19	24
#TigrayGenocide	24
#covid19	24
#LeniKiko2022","c...	16
#FJB	16
#5	16
#Kakampet	16
#TrumpWon	16
#Risk	16
#HUNTA331	8
#성 준 ","display_tex...	8
#GoThere	8
#3VAX	8
#BlackLivesMatter...	8
#날 렐 키	8
#GloryToUkraine",...	8
#ElleReviews","cr...	8

Figure 4: Screenshot of Shell (Trend Analysis)

As the last step, we moved the data to Google Studio for visualization and displayed it as a Pie-Chart. Google Studio gives the option for tables, charts and much more for visualization.

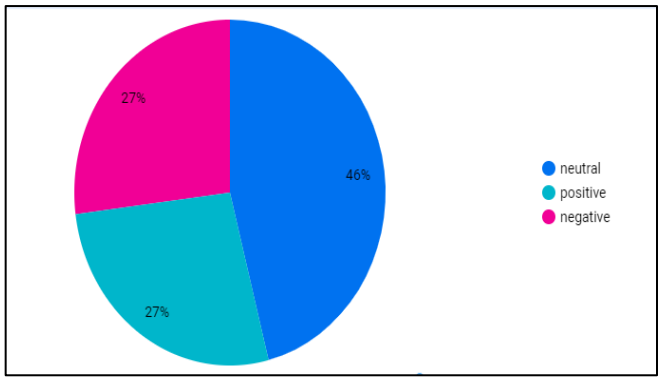


Figure 5: Polarity of Tweets

V. LIMITATION

There are several limitations of the applications which can be enhancements in the future.

A. *Efficiency*

Since there is a time constraint to this application, we can only include Lexicon-based approach which does not give accurate results of the analysis. This approach can give 60-70% correct reults.

B. *Absence of GUI*

The application does not have a Graphical User-Interface which can be intuitive for the user.

C. *High requirements and configuration time*

The application is difficult to set up in a machine and requires at least 8GB of RAM to support Virtual Machine environment. Initially, it also takes more time to configure and setup.

D. *Unable to detect other sentiments*

Some users of twitter express their opinion in a different manner where our approach might have its shortcomings. Since the approach uses the polarity of words and compares

it with a dictionary, detecting emotions like anger, fear, excitement and sarcasm is difficult with the approach.

VI. Conclusion and Future Work

Individuals and many organisations may benefit from analysing the sentiments of tweets and trends, which can lead to an ideal output that they can use to improve and achieve in the future.

Following are future improvements which we may plan to implement for twitter sentiment analysis:

- Since we have used Lexicon based approach during the analysis, due to time restrictions we were not able to compare the performances of all approaches. This can be done to reach the best possible solution.
- Also, a hybrid approach would provide better results since it will have a training set which would train the machine to detect emotions other than the three basic emotions.
- This analysis is done through Character User Interface (CUI). Further this can be done using Graphical User Interface (GUI), so that it can be more easy and feasible to use.
- Lastly, we can move the application on cloud to use the processing power of cloud which can be beneficial in quick analysis of large dataset. This will decrease the processing time for analyzing large datasets.

Acknowledgment

We thank Dr. Shafaq Khan, faculty at University of Windsor, for giving us an opportunity to find a problem statement that piqued our interest and work on its solution. She was a source of guidance and motivation throughout the project. We also would like to sincerely thank the GAs, Ala' Alqaisi, Safiia Mohammed and Sheikh Sadaf Tasin. Additionally, we extend our gratitude to School of Computer Science, University of Windsor, for allowing us to pursue our idea and see it to completion. Last but not the least, we appreciate our classmates for their participation during project presentations and for sharing constructive feedbacks.

References

[1] https://en.wikipedia.org/wiki/Twitter

[2] Twitter Sentiment Analysis Approaches: A Survey, Omar Y. Adwan, Marwan Al-Tawil, Ammar M. Huneiti, Razan H. Al-Dibsi, Rawan A. Shahin, Abeer A. Abu Zayed University of Jordan, Amman, Jordan.

[3] J. Ranganathan and A. Tzacheva, “Emotion mining in social media data,” Procedia Comput. Sci., vol. 159, pp. 58–66, 2019.

[4] A. Giachanou and F. Crestani, “Like It or Not: A Survey of Twitter Sentiment Analysis Methods,” ACM Comput. Surv., vol. 49, no. 2, Jun. 2016.

[5] D. Kılınç, “A spark-based big data analysis framework for real-time sentiment prediction on streaming data,” Softw. - Pract. Exp., vol. 49, no. 9, pp. 1352–1364, 2019.

[6] A Literature Review on Twitter Data Analysis, Hana Anber, Akram Salah, A. A. Abd El-Aziz Internal Journal of Computer and Electrical Engineering.

[7] https://blog.hubspot.com/service/sentiment-analysis-tools