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*—This electronic document is a "live" template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. **CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.* (*Abstract*)**

Keywords—component, formatting, style, styling, insert (*key words****)***

I. INTRODUCTION

Twitter has more 350 million users and through which it gets Millions of tweets every day. 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 do the sentiment analysis on it. Although, it's 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 we have used.

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 plethora of hashtags being used and many more topics on twitter but considering the most recent happening and of which every single individual is aware of we have decided to go with the COVID-19 as a topic for this project. We will be working upon the data which we will be gathered involving around the COVID. We will be finding the people's opinion and interest in the given topic and analyze the trends and patterns to predict the coming happenings. With this, we can notify concerned authorities regarding the trends which can prevent problematic situations before it even creates into something big.

II. RELATED WORK

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

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/filtered to make a basic structured dataset. After this, analysis is done using different tools and we can finally come up with a report. According to the paper [4], 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 [7], 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 [5] can be more beneficial and produce more accurate results. The work in [6] 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 [1] 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 [3]:

a) Super metrics
b) Brand watch
c) Native Twitter analytics
d) Social Searcher
e) Brand24

III. PROPOSED MODEL

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. *Functional Requirement*

The functional requirements are as follows:

1) The system should be able to take data from Twitter and save it to HDFS for later analysis.
2) After retrieval, the system should be able to process tweets stored in the database.
3) The system should be able to analyze data and categorize the polarity of each tweet.
4) Calculate tweet's sentiment and store in the database, for later querying.
5) 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.
6) The system should be able to move trends and final data to cloud for further processing and to present in a graphical manner.
7) Finally, plot a graph to analyze the performance of the hive and spark approaches in querying databases

B. *Proposed Model*

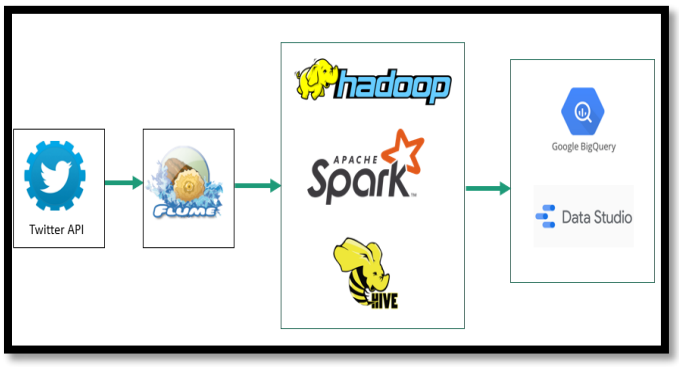

Figure 1: Proposed Architecture

The architecture consists of three phases:

1) *Data Collection and Storage*

a) 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.

b) Then using Hive, we will move unstructured data obtained in JSON format into structured format Hive tables for further analysis.

c) 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:*

a) 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.

b) 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.

c) 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

a) 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.

b) 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 (Screenshots and Output)

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

V. Limitation

VI. Conclusion and Future Work

VII. References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.