Real time Sentiment Analysis of Tweets using Apache Spark and Scala

V Mareeswari\textsuperscript{1}, Sunita S Patil\textsuperscript{2} & Ramanan G\textsuperscript{3}

\textsuperscript{1,2}Department of Computer Science and Engineering, \textsuperscript{3}Department of Aeronautical Engineering, ACS College of Engineering, Affiliated to VTU, Bengaluru, Karnataka, India. DOI: https://doi.org/10.34293/acsjse.v1i2.9

Abstract - Sentiment Analysis is becoming the field of focus with time considering the user experience weighs much more for the business to grow and for the studies as well. The sentimental expressions refers to the emotions or feeling of a person across certain point of focus or issues. So, in this project, with the assistance of Apache Spark Framework, an open source data streaming and processing platform, sentiment evaluation is done on the tweets from Twitter by the means of real time processing as well as an Ad-hoc Run. Some preprocessing of the textual data has been done upon for better characteristics extraction thus resulting in greater accuracy. The validation of this has been done for achieving better result by comparing the other processes when Naive Bayes algorithm is used.

Keywords: Sentiment Analysis; Apache Spark Framework

I. INTRODUCTION

Sentiment Analysis is finding out the type of emotion from the textual information using several text evaluation techniques. Sentiment evaluation allows businesses to identify user’s sentiment closer to products, brands or services in online conversations and feedback. Sentiment Analysis has grown to become a motivating field for domains which deals with user experience. The expression resultant from sentiment analysis refers to the experience of the user on some issues. Now-a-days social networking sites are playing vital role since majority of the population uses several social media platform to show their views and experience. Because of the majority of the population being on the social media platform, it generates a huge amount of data in the form of comments, articles etc. about certain topic. So by implementing an automated process which investigates and classifies the user experience would be quite essential.

It’s principally a good and thought-about tool for obtaining information directly from the end users. The art of classification of the information gained by their nature is known as Sentiment Analysis. Sentiment Analysis is a part of Natural Language Processing(NLP), and it’s goal is to classify the document’s nature as a positive or negative.

So, for Sentiment Analysis the focus will be towards the data generated by the Twitter, a micro blogging platform where millions of user interact with each other and share their opinion through using some sentences or twitter calls it tweets. Now, the fast-monumental growth of data that is generated needs database for storing and processing, but it’s quite difficult for the databases to handle the depth of information in a short time. Also, keeping in mind that those databases are made for structured data but difficulty arises when the amount of data is huge, so these solutions are not beneficial for the corporations to manipulate the unstructured or big data.
There are several frameworks like Hadoop are present which help process the large amount of data and make it almost effortless in doing so. One of the best ways to do this is by using parallel computing strategies with the multi-core processors while dealing with the that amount of data.

Understanding customers’ feelings is essential for the business corporations since the customers are able to express their minds and feelings now more overtly than ever. By analyzing the consumer feedback, from survey responses to social media conversations, manufacturers are capable of listening to their clients and tailor services and products to fulfill their needs.

II. LITERATURE SURVEY
[1] Shihab Elbagir Saad et al proposed the sentiment analysis of the twitter data using the ordinal regression topic by topic and the result indicated that Decision Tree was giving the best accuracy. [2] Hajime Watanabe et al proposed a system to detect hateful speech on twitter and filter any content that presents hateful language or language inciting to hatred.[3] Ping-Feng Pai et al proposed a model that forecasts the sales of a vehicle using three types of data, namely sentiment scores of tweets, stock market values, and hybrid data.[4] Samah Aloufi et al proposed a domain specific approach for understanding the sentiment of football fans’ during the event. [5] Ankan Mullick et al presented an end-to-end system that identifies the opinion list using the hash tags used on twitter, extracting suitable list answers from relevant tweets. [6] Mark E. Larsen et al describes the “We Feel” for analyzing global and regional variations in emotional expression and report the results of validation against known patterns of variation in mood. [7] Jose Angel Diaz-Garcia et al proposed a new non-query based system which combines mixed proposal of sentiment analysis and association rules to discover patterns in microblogging textual data. [8] Hajime Watanbe et al proposed a approach based on unigrams and patters that are automatically collected from the training set and is used to train the machine learning algorithm. Alec Go [9] proposed a novel approach to classify sentiment of the twitter message automatically and showed that machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) have accuracies above 80% when trained with emoticon data.

III. PROPOSED SYSTEM AND IT’S ADVANTAGES

Figure 1: Number of users Per Million on Several Social Media Platforms
In this project, we analyze the sentiment of the twitter data by using Apache Spark engine and Scala programming, which we know is popular an open-source distributed data processing and analysing the streaming data on any platform it helps to sanitize on several memory-saving techniques. The main reason for using Apache Spark’s Machine learning library (MLIB) is to process a very large amount of data with producing very efficient results. We recommend some Pre-processing of twitter data and Machine learning text classification of feature extraction process for getting greater results in Sentiment Analysis of the twitter data for text classification. We deployed all of our sentiment analysis methods and text classification algorithm on to the Amazon Web Service (AWS) cloud which will be easy for the user to access it from anywhere in the world to get the classification results. AWS provides various cloud service like computing instances, storage bucket and other cloud services but in our project, we used AWS computing instances to launch our project online and evaluate all analysis of the twitter data on the AWS ubuntu machine.

Advantages
- Overcomes the drawbacks of the existing systems
- Proven High accuracy
- Memory and Time-efficient
- A solution made available to the public

A. Microblogs

Microblogging evolved as social networking made its mark on the internet. When the dawn of the internet there was nothing in early 2005 the expansion of social networking came into existence microblogging became the most tremendous way of sharing someone feeling or any sentimental thoughts over the internet in a matter of seconds due to the expansion of the internet. Anyone with access to the internet or simply we can say with a smartphone now a days smartphone plays a very important role in the social networking phenomenon. Now come to the definition of the microblogging it’s nothing but the sharing or updating simple text messages on any of the microblogging platforms. In the beginning, microblogging services offered the text limit of sharing the message from 120 to 150 character which was supported by the tumble logs. Tumble logs was the first microblogging platform which supported the sending messages over the internet or the user who they have subscribed to their channel or follow them. And it was just the beginning latter on several platforms came into the existence of the microblogging. In 2006 twitter came into existence with that several other platforms like tumbler, Jaiku, Paunce in2007 and Plurk inn 2008. It was not only the microblogging platforms there were several other social networking platforms which supported microblogging like Facebook which is the major holder of the user of the social media networks. And after that MySpace, Linked in like that all came into existence and in all of that twitter and Facebook hold the most number of the active user according to the source 2.5 billion monthly users in the starting of 2020 and/or the twitter 325 million monthly active users and 140 daily active users on the twitter which is quite a bit of expansion over the year on the active user over those microblogging platforms. With this amount of active
users on the social media sites companies started the opinion mining or we can say the sentimental analysis of the daily user's twits or status which helps them to improve their quality of the service and the quality of the product by gaining the knowledge of the user's thoughts by sentiment analysis.

B. Motivation

The future of social media and its interest of people deals with microblogs which can be easily used to its services which provides an open platform to share their ideas and opinion where more people connect to each other to share their thoughts. From such opinion analysis and monitoring provides a great opportunity for both private and public sectors. As it is concerned with a private organization which affects its company status with different rumours and negative opinion through the social network. so the company realizes the opinion of people to enhance the product and to build better relations to the customer by microblogging, and for the public sector, recently it shows activities on the social network provides the outcome to political parties. For example, Egypt’s president resigns which changes the political environment to total tweets made on twitter. Hence, Twitter is a good platform to analyse and predict the popularities of political parties. Therefore, tracking and analysing provides a better environment to public opinion during elections or other events.

C. System design

System design is an important phase of the software development process. The purpose is to plan the solution of a problem specified by requirement documentation. System design is the process to define the architecture, components, modules, interface, and data for a system. The system design consists of two phases such as high-level design and low-level design.

D. System Implementation

The project is implemented on two aspects such as usability aspect and technical aspect. In the usability aspect, the project is implemented on a java application and user-friendly interface using java's view architecture. In the technical aspect, servers consist of apache tomcat to develop the product and JBoss application server to host the product as well as databases.

Twitter Live Streaming using Spark and Scala

Spark Streaming is the further extension of the core Spark API that is highly scalable with high-throughput less fault-tolerant stream processing of live data streams. In this streaming, the data can be inputted from different sources like Kafka, Twitter, or TCP sockets, which is processed using complex algorithms and is expressed with high-level functions like map, reduce, join and window. In the end, the processed data can be moved out to the file systems, databases, and live dashboards.
The application consists of operation to end-users as:

**Figure 2: System Architecture**

**Dashboard:** in this component user sees a summary of sentiment analysis of all tweets which are streamed from scala. It shows sum of positive counts and sum of the negative count of live tweets. **Tweets Analysis:** In this case, users will see a detailed analysis of tweets with current tweets.

**Figure 3: Data Flow from Twitter to Dashboard**

**Figure 4: User Application Architecture**
Adhoc Run: this component is used to analyse the sentiment of text imputed to text box before posting to digital media.

Requirements
The requirements are of two types:

**Functional Requirements**
- The tool builds emotion word ontology efficiently and effectively.
- The tool allows user to analyse the text for emotion in file or directory form.
- The user also is able to pull tweets from a specified twitter handle.

**Non-Functional Requirements**
- It should be easier to access from various browsers available.
- Response time should reflect real-time observations.
- Algorithms should not fail in any of the test cases.
- There should n't be any security concern on merged data.

E. Results
For testing of the solution proposed by this, several experiments have been done on the twitter data which is collected and is trained using Naive Bayes Algorithm.

Before the analysis the data was preprocessed to remove any kind of noise present in the dataset and was made into the format required by the algorithm. The datasets are split into two different datasets, test and train in the ratio of 3:7 i.e 30% of the datasets were used for testing and 30% for testing. Upon comparing different algorithms we found out that Naive Bayes was giving better results in the form of accuracy.

In the end, we check the impact of using multiple computing nodes in the Spark Cluster in comparison to the total running time. By using the Apache Spark streaming services which is a part of Apache Spark’s library, we were able to process the tweets in the real time prediction and sentiment focused on particular subject.

F. Conclusion and Future work
Data collected from the twitter cannot be processed directly using the algorithm and it needs to be preprocessed in order for it to be fit for sentiment evaluation.

Tweets in the form of a review, comment etc. is considered as big data and it demands proper tools for the processing. Upon utilizing Apache Spark Framework’s machine learning library for executing different machine learning classification algorithms the results focused towards the enhancement in accuracy using the Naive Bayes Algorithm is the dataset were increasing exponentially.

For achieving better experience, usability and making this project scalable, we’ve utilised the Amazon Web Services which in being more effective as it is online and can be accessed from anywhere using the internet.
In the near future, we will focus on adding several other features for better understanding the human sentiment and also for analysing the larger datasets.

IV. REFERENCES


