8157 – Project Milestone 1

**Ali Hasan, Md Tafsirul Islam, Muzakkir Qadri Mohammed, Abhilash Harsh Saksena**

# **Problem statement**

Scan a set of natural language texts for sentiment analysis, from an e-commerce websites comment section for predictive classification purposes by populating a database of search index. User reactions can be utilized as a reference to improve service quality, future product development, reduce suer churn and improve how the product is presented. This project can possibly be used to provide specific marketing strategies. We also plan to show gathered information in graphs.

# **Literature Review**

## **Real-Time Analysis of Twitter Streaming data for Stock Prediction**

*Authors: Sushree Das, Ranjan Kumar Behera, Mukesh kumar , Santanu Kumar Rath*

In this study, an attempt has been made for making financial decisions such as stock market prediction, to predict the potential prices of a company’s stock and to serve the need of this, Twitter data 1 2 has been considered for scoring the impression that is carried for a particular firm. Streaming data proves to be a perennial source of data analysis collected in real-time. Streaming data basically deals with the continuous flow of data which carries information from sources like websites, mobile phone applications, server logs, social websites, trading floors, etc. The major characteristics of such data being its accessibility and availability, help in proper analysis and prediction of user behavior in a ceaseless manner. The classifying model made out of historical data can be relentlessly honed to give even more accurate results since its outcome is always compared to the next tick of the clock. Spark streaming has been considered for the processing of humongous data and data ingestion tools like Twitter API and Apache Flume have been further implemented in this analysis.

## **Real Time Sentiment Analysis of Twitter Data using Hadoop**

*Authors: Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde*

This paper provides a way of sentiment analysis using Hadoop which will process the huge amount of data on a Hadoop cluster faster in real time .Twitter, one of the largest social media site receives tweets in millions every day. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. Their project uses emoticons but the use of hashtags to determine the context of the tweet is not done. Hence with the current limitations the accuracy is found to be 72.27 %.

## **A Big Data Analysis Framework Using Apache Spark and Deep Learning**

*Authors: Ritvik Shrivastava, Pulkit Kumar, Sreyashi Nag, Anand Gupta, Hardeo Kumar Thakur*

Frameworks such as Apache Hadoop and Apache Spark have gained a lot of traction over the past decades and have become massively popular, especially in industries. It is becoming increasingly evident that effective big data analysis is key to solving artificial intelligence problems. Thus, a multi-algorithm library was implemented in the Spark framework, called MLlib. While this library supports multiple machine learning algorithms, there is still scope to use the Spark setup efficiently for highly time-intensive and computationally expensive procedures like deep learning. In this paper, we propose a novel framework that combines the distributive computational abilities of Apache Spark and the advanced machine learning architecture of a deep multilayer perceptron (MLP), using the popular concept of Cascade Learning. We conduct empirical analysis of our framework on two real world datasets. The results are encouraging and corroborate our proposed framework, in turn proving that it is an improvement over traditional big data analysis methods that use either Spark or Deep learning as individual elements.

## **A Review of Sentiment Analysis in Twitter Data Using Hadoop**

*Authors: L. Jaba Sheela*

Twitter is an online social networking site which contains rich amount of data that can be a structured, semi-structured and un-structured data. In this work, a method which performs classification of tweet sentiment in Twitter is discussed. To improve its scalability and efficiency, it is proposed to implement the work on Hadoop Ecosystem, a widely adopted distributed processing platform using the Map Reduce parallel processing paradigm. Finally, extensive experiments will be conducted on real-world data sets, with an expectation to achieve comparable or greater accuracy than the proposed techniques in literature. It is proposed to stream real time live tweets from twitter using Twitter API, and the large volume of data makes the application suitable for Big Data Analytics

# **Functional Requirements**

* The system will have a main page.
* System will fetch the reviews from social media and ecommerce websites.
* System will extract the useful information from the data fetched.
* The extracted useful information will be stored in the database
* System will perform the sentiment analysis on the stored data.
* It will separate both the positive and negative reviews based on the performed sentiment analysis.
* The results of analysis will be displayed on the main web page in a graphical form.

# **Usability/Non-Functional Requirements**

## **Performance**

The response time of the system will be minimum against any user input.

## **Usability**

Users of age above 12 will be able to use the system easily because of the friendly user interface.

## **Reliability**

The system will be available to users all the time.

## **Maintainability**

The system will experience minimum downtime in case of any update or failure.

# **UML Diagram**

In the UML diagram below we have two types of actors here- the users, and the administrators. Users will view products and possibly leave a review. Our task as administrators is to-

1. Get the reviews by extracting them using scripts
2. Store data in HDFS
3. Execute several querries and saving those results
4. Show results to manager in a graphical format

Diagram

Description automatically generated