**BIG DATA**

**ASSIGNMENT 3**

**22L-7468**

**Introduction:**

This report is about a project where I used tweets related to the FIFA World Cup 2018 to understand what people were saying about the players and teams. All work was done in local environment using jupyter notebook and locally installed HADOOP

**Setting Up the Environment**

I set up a system on my laptop called Hadoop Distributed File System (HDFS). I used Hadoop 3.3.6 and Java 8 versions to set it up. I also downloaded Apache Spark, which is a tool that helps handle big data well.

**Storing the Data:**

I stored all the tweets data in the HDFS with file name TweetsDataset.csv inside directory named data. Then, I loaded this data into a PySpark DataFrame.

**Cleaning the Data:**

The tweets had some extra information that I didn’t need, like **duplicate tweets** and some extra columns. So, I removed these. I also made sure that the tweets didn’t contain any links or retweets, as these could affect the results. I also removed the **mentions (@)** and **trends (#)**. Then I **tokenized** the tweets and removed the **stop words**. I also changed all the words to their base form. For example, ‘running’ was changed to ‘run’. This process is called **stemming** and **lemmatization**. Then I handled the emojis by removing them.

**Sentiment Analysis:**

* **Pandas Data frame (Logistic Regression)**

I converted the PySpark data frame to pandas data frame and used Logistic Regression to understand the sentiment of each tweet. This method looks at the words in the tweets and figures (X) out if the sentiment (Y) is positive, negative, or neutral. I was able to obtain accuracy of **0.7785**

* **PySpark Logistic Regression**

I converted the TweetSentiment (original) column into a numeric ‘label’ column. This is necessary because the machine learning algorithm I used (Logistic Regression) requires the labels to be numbers in pyspark. Then, I converted the TweetText column into a ‘features’ column. This column is a vector that represents the frequency of each word in the column. This is my input feature for the machine learning model.

I split my Data Frame into a training set (80% of the data) and a test set (20% of the data). I then created a Logistic Regression model and trained it on the training set. After the model was trained, I used it to make predictions on the test set. I was able to obtain accuracy of **0.7038**

* **Sentiment Analysis with NLTK**

I performed sentiment analysis on the TweetText column using the NLTK library’s **SentimentIntensityAnalyzer**. This tool calculates the sentiment of each tweet based on the words it contains. It returns a ‘compound’ score which is a weighted sum of the positive, negative, and neutral sentiments of the words in the tweet. If the ‘compound’ score is positive, the sentiment is ‘positive’. If it’s negative, the sentiment is ‘negative’. If it’s zero, the sentiment is ‘neutral’. Then I created a confusion matrix and found that most of the the predictions were correct.