BIG DATA

ASSIGNMENT 3 -PART 2

22L-7468

Sentiment Analysis:

- **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 **0.7423**

- **PySpark Bayes**

Using StringIndexer, I convert sentiment values (positive, negative, or neutral) into numerical labels (label). The CountVectorizer transforms the no_emojis column into a feature vector (features) for numerical representation. I divide the data into 80% training and 20% testing sets. A Naive Bayes classifier (nb) is trained with labelCol set to label. I make predictions on test data and evaluate accuracy, precision, recall, and F1-score using MulticlassMetrics. I managed to obtain an accuracy of **0.6730**

Analysis:

- **Sentiment distribution:**

I found the sentiment distribution by using group by on sentiments in dataframe. I found that **668516** tweets were positive, **264996** were negative and **2502936** were neutral

- **Top keywords:**

I used group by on keywords after exploding the ‘no_emojis’ text to find the top keywords.

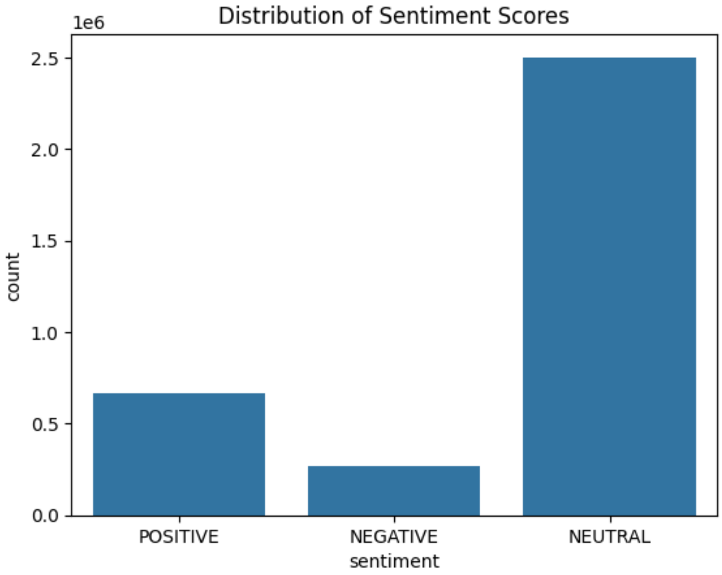
- **Trends:**

I found the trending words by using group by along with the count to find the count of the most used words among the tweets of 6 million seeing the result it was found that they were positive in most of the cases.

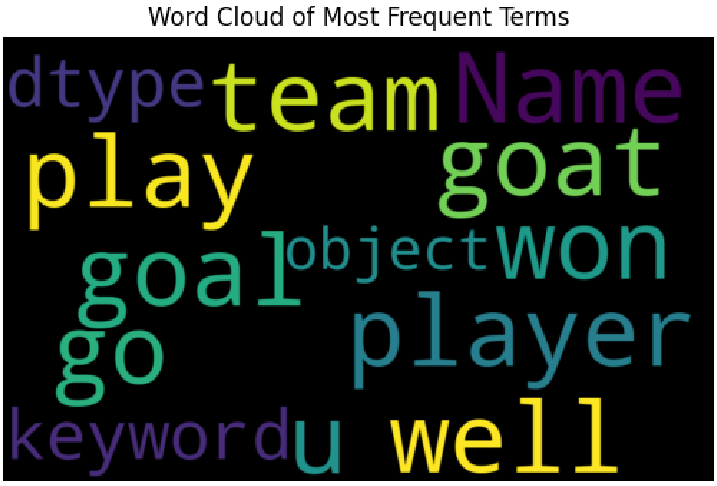
Graphs:

- **Distribution of sentiment scores**

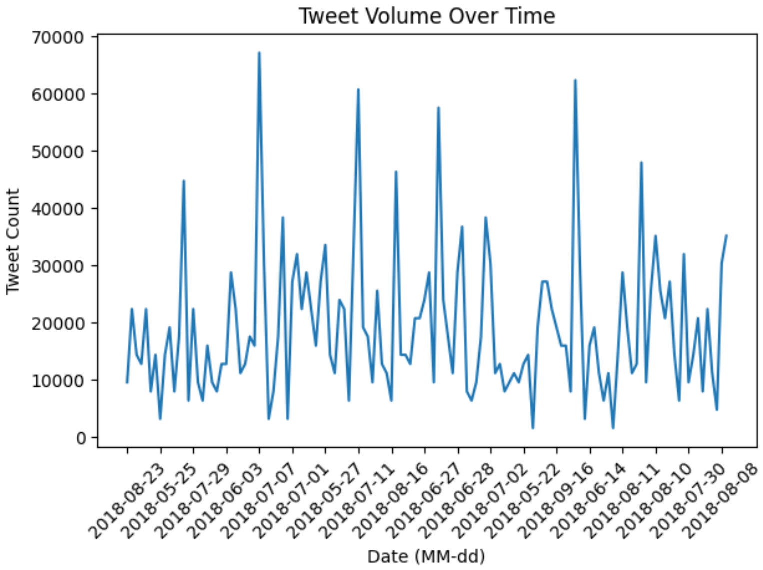
Using the seaborn and matplotlib I ploted the barplot to view the distribution of sentiment scores.



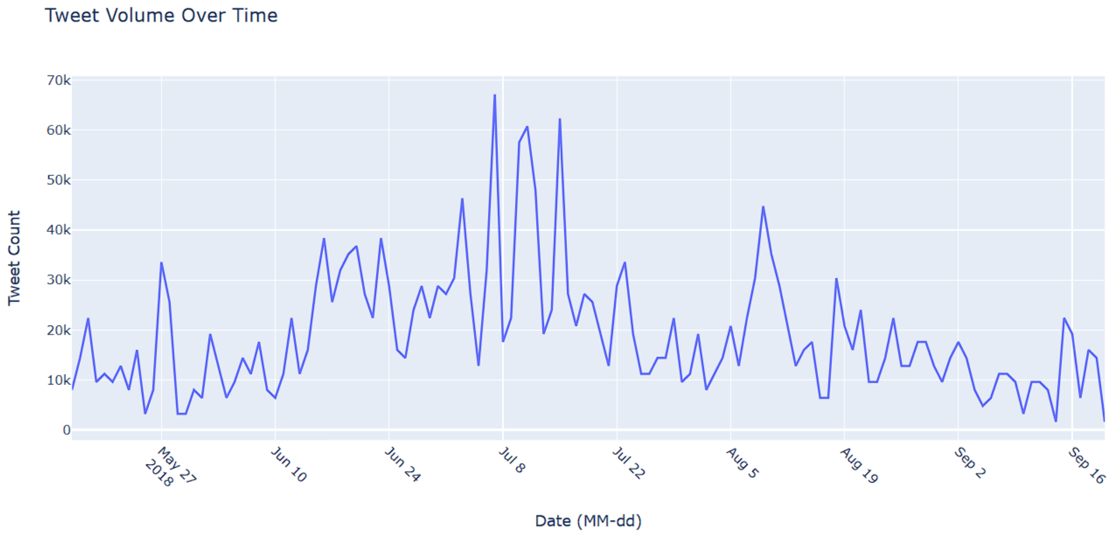
- **Word count of most frequent used terms**



- **Tweet volume over time**



- **Using plotly to print tweet volume over time**


Tweet Volume Over Time
Tweet Count
70k
60k
50k
40k
30k
20k
10k
May 27
Jun 10
Jun 24
Jul 8
Jul 22
Aug 5
Aug 19
Sep 2
Sep 16
Date (MM-dd)