Data Incubator Proposed Capstone project Presentation

Project Title: Tweet Sentiment Analysis

Presented by

Aklilu

- ***Objective***

✓ *To build a predictive Machine Learning algorithm(ML) to predict the sentiment of each tweet*

✓ *To compare the different classification ML algorithms performance on predicting the sentiment of each tweet*

- ***Motivation***

✓ *LyX combines the power and flexibility of TeX/LaTeX with the ease of use of a graphical interface*

✓ *it is good for preparing a technical reports such as theses, academic paper, books*

✓ *It is an open source and easy to use*

✓ *It has some draw backs in exporting files from lyx to Microsofts.*

➢ *Introduction*

- ✓ *Lyx is an open source graphical user interface document processor based on the LaTeX typesetting system.*
- ✓ *It was developed by Matthias Ettrich in 1995 with the name of **Lyrix**.*
- ✓ *Unlike most word processors, which follow the WYSIWYG ("what you see is what you get") paradigm, LyX has a WYSIWYM ("what you see is what you mean") approach*
- ✓ *what shows up on the screen roughly depicts the semantic structure of the page and is only an approximation of the document produced by TeX*
- ✓ *It is a very important editor for preparing technical reports and reports.*

➢ *Data Structure*

- ✓ *In this study 1,049,074 number of tweets with corresponding six variables were included.*
- ✓ ***Dependent variable: Polarity_Tweet****:the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)*
- ✓ ***Other variables****:*
  - ***ID****:the id of the tweet,****Date****:the date of the tweet*
  - ***Query****:If there is no query, then this value is NO_QUERY,*
  - ***User****:the user that tweeted (robotickilldozr)*
  - ***Text****:the text of the tweet are included*

➢ *Exploratory Data Analysis*

❖ ***Table 1.1 below summarizes:***

- *800,177 tweets have a negative sentiment ➔ About 76% of the total tweets*
- *248,758 tweets have a positive sentiment ➔About 23.7% of the total tweets*
- *Only 139 tweets have a neutral sentiment ➔About 0.0132% of the total tweets.*

Table 1.1: Summary of the polarity tweets

Polarity Tweet	Number of Tweet	Percentage of Tweet
Negative Tweet	800177	76.27460%
Positive Tweet	248758	23.71215%
Neutral Tweet	139	0.01325%

Figure 1.1: Bar chart of the Polarity Tweet

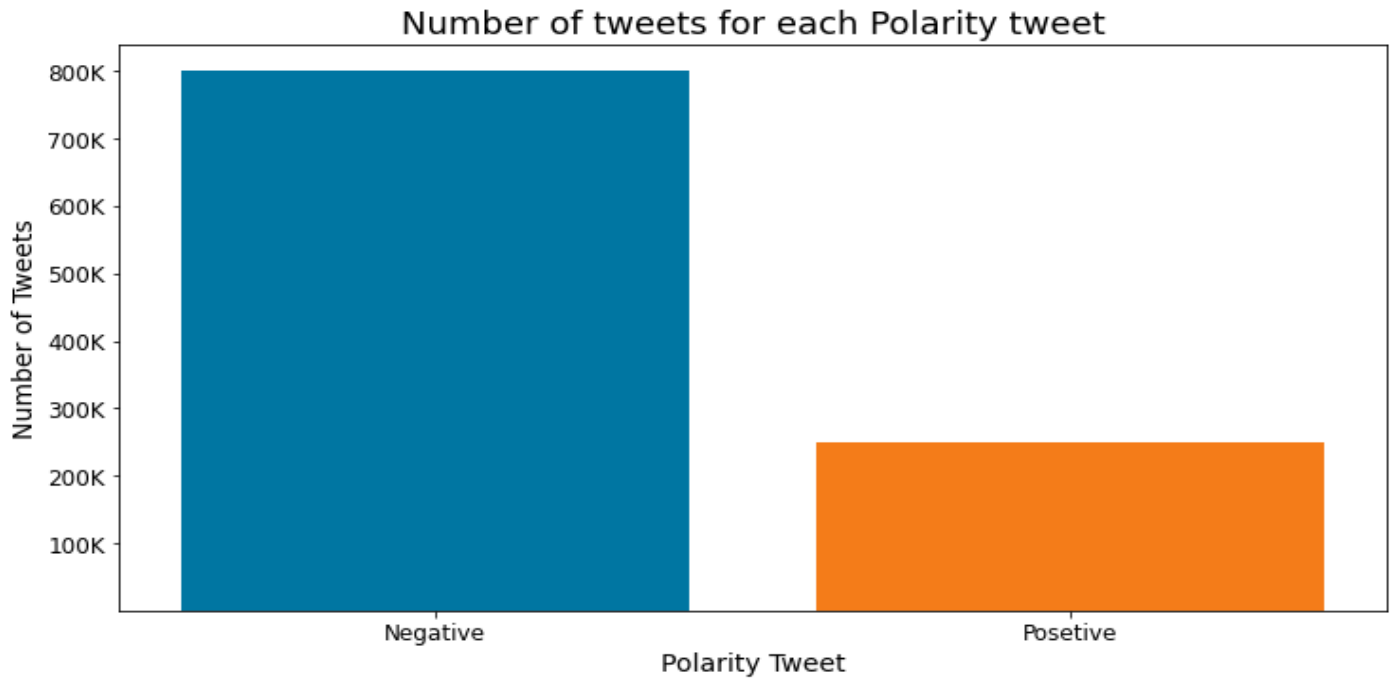

Figure 1.1 above shows

- ✓ *About 800,000 tweets of tweets has a negative sentiment tweets*
- ✓ *About 200,000 tweets of tweets has a positive sentiment tweets*
- ✓ *Positive sentiment tweets is almost one-fourth of the negative sentiment tweets*

Figure 1.2:Number of Tweets for each month

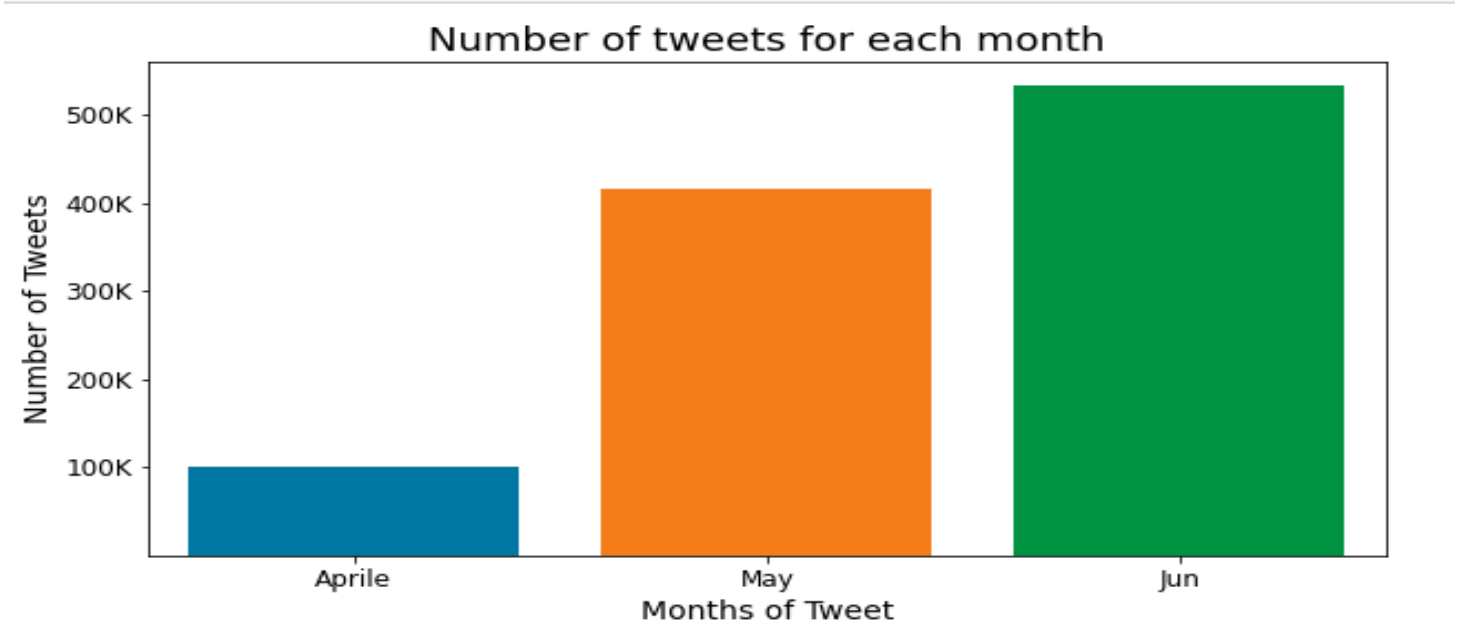

Figure 1.2 above shows

- ✓ *Most of the tweets were tweeted on Jun*
- ✓*Least number of tweets were tweeted on April*
- ✓*Almost five fold of the tweets tweeted on April were tweeted on Jun*
- ✓*On May almost four fold of tweets tweeted on April were tweeted on May*

Figure 1.3: Bar chart of tweets in each month

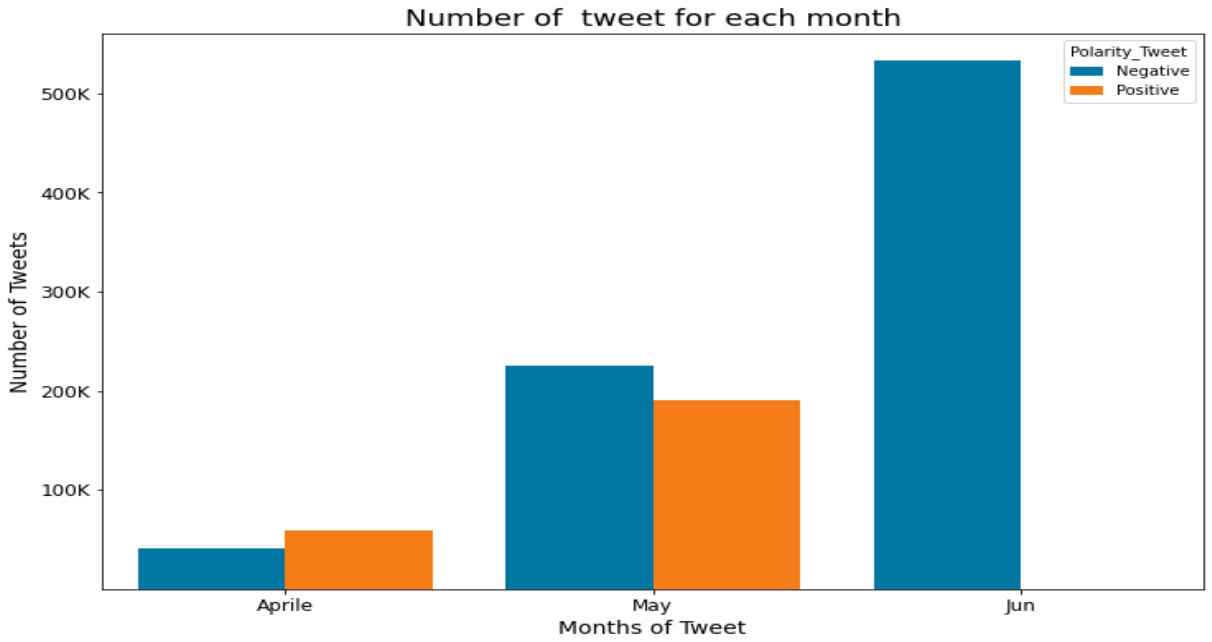

Figure 1.3 above shows

✓*Most number of the tweeted tweets had a negative sentiment. .*

✓*Least number of negative sentiment tweets were tweeted on April*

✓*In may most number of positive sentiment tweets were tweeted*

✓*Not positive sentiment tweets were tweeted on Jun*

Figure 1.4: Bar chart of tweets on April

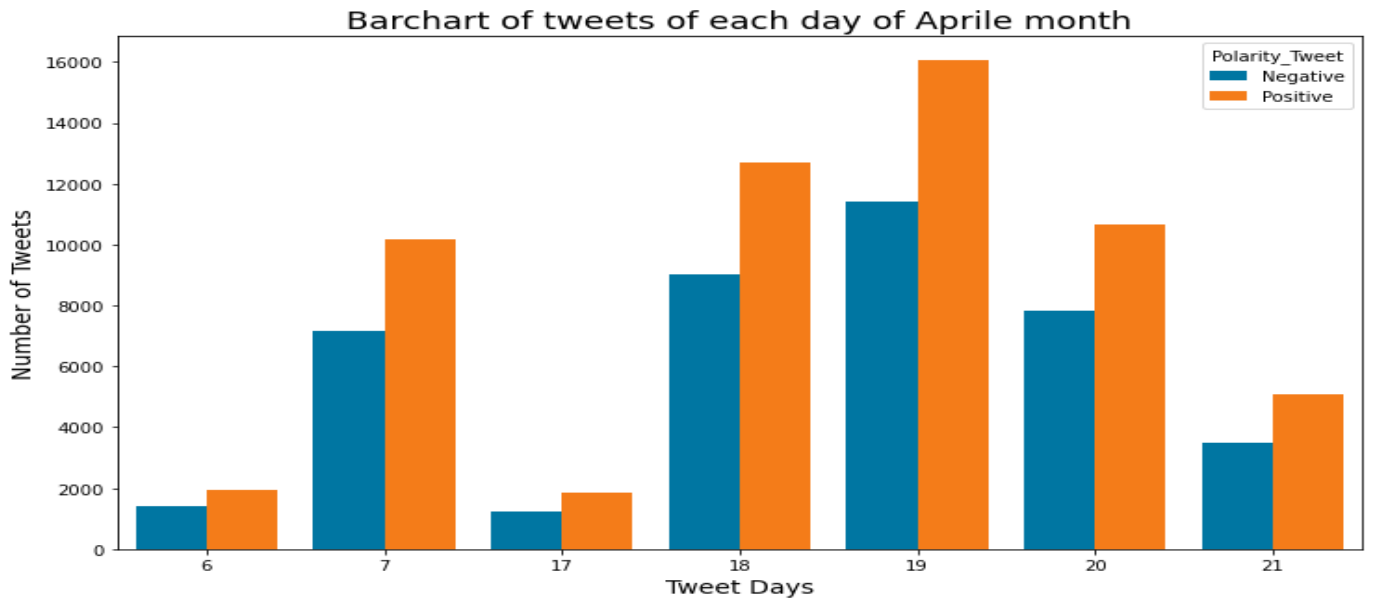

Figure 1.4 above shows

- ✓ In all tweet days in April, there were lower number of negative sentiment tweets
- ✓ in April 19, there were a highest positive and negative sentiment tweet tweeted
- ✓ In April 6 and 17, almost similar number of negative and positive tweets tweeted.

❖*Feature Engineering*

✓ *Create Term Document-Matrix*

✓ *Preparing the Feature and Target variable for modeling*

✓ *Standardizing/normalizing the data set*

✓ *We can use PCA,FA to reduce the dimension the feature variable*

❖*Data Modeling*

✓ *Data will be splinted into train, validation and test data set in order to get a generalizable model*

✓ *The validation data set will be used for hyper tunning hyper parameters*

✓ *The distribution of the negative and positive sentiment tweets are 76% and 24% respectively*

❖*In this study different Classification Machine learning algorithms will be assessed so as to select the better ML in predicting the sentiment of the tweet such as:*

- *Logistic Regression*
- *KNN*
- *SVM*
- *Linear Discriminant Analysis*
- *Quadratic Discriminant Analysis*
- *Naive Bayes*

- *Decision Trees*
- *Random Forest*
- *Gradient Boosting*
- *Adaptive Boosting*
- *CatBoosting Classifier*
- *Light Gradient Boosting*
- *LSTM*
- *Extreme Gradient Boosting*