Sentiment Analysis of Trendy Tweets in the US

Ibrahim S. Eldivan

OVERVIEW

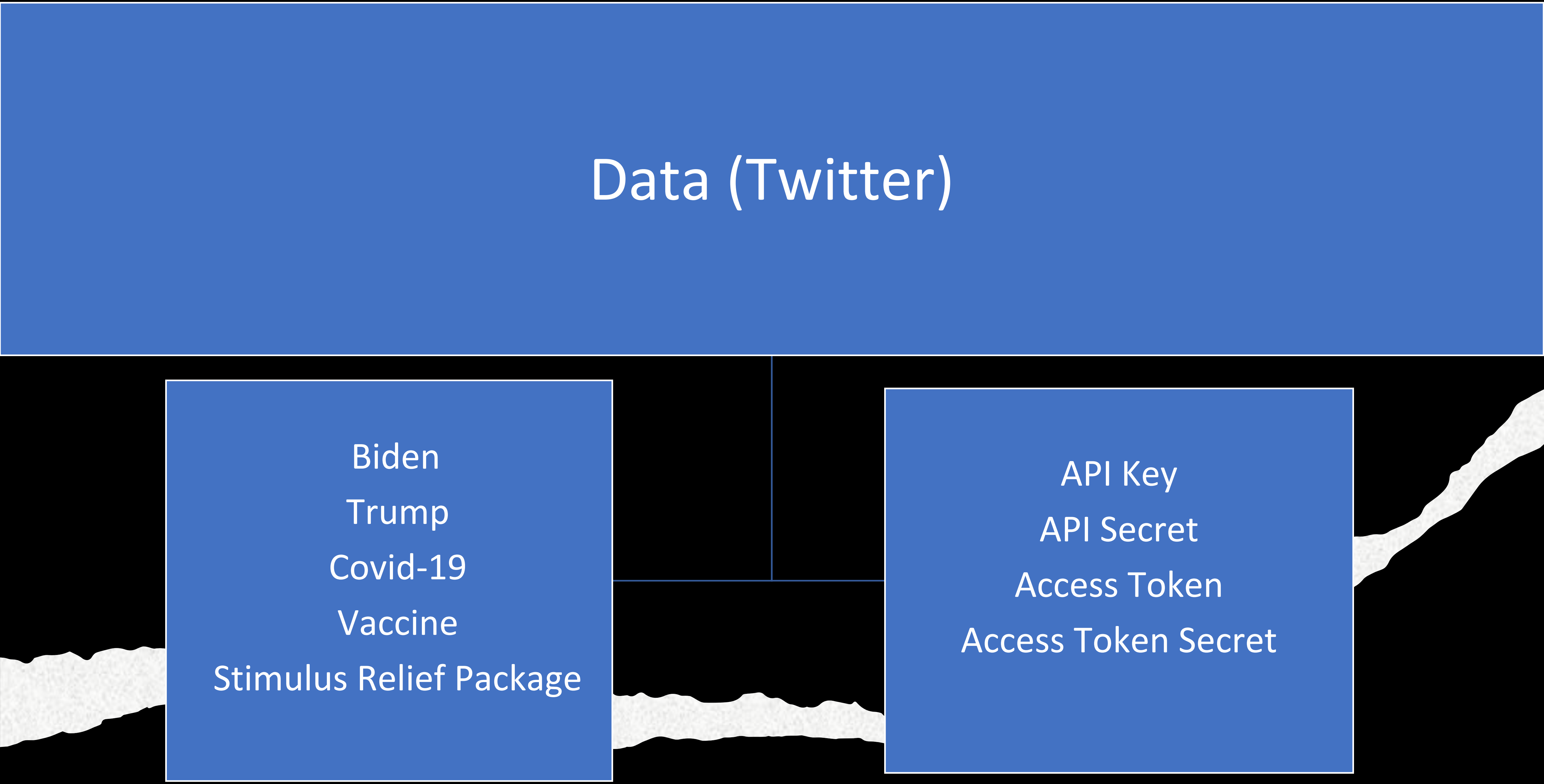
- Twitter is the best indicators of the wider pulse of the world and what’s happening within it among the social media tools.

- Categorizing opinions in the text of tweets-and determine the user’s attitude is positive, negative, or neutral-is highly valuable.

- Sentiment Analysis can help us decipher the mood and emotions of general public and gather insightful information regarding the context.

- The tweets on a specific time period, specific location, and specific subject will be the focus of this project.

- These tweets will be classified by sentiment analysis and output of this analysis will be evaluated.



Data (Twitter)

Biden

Trump

Covid-19

Vaccine

Stimulus Relief Package

API Key

API Secret

Access Token

Access Token Secret

DATASET

- Twitter data retrieved by connecting the Twitter API, 5000 tweets with the extended mode for each subject. All the tweets will be in English as language.

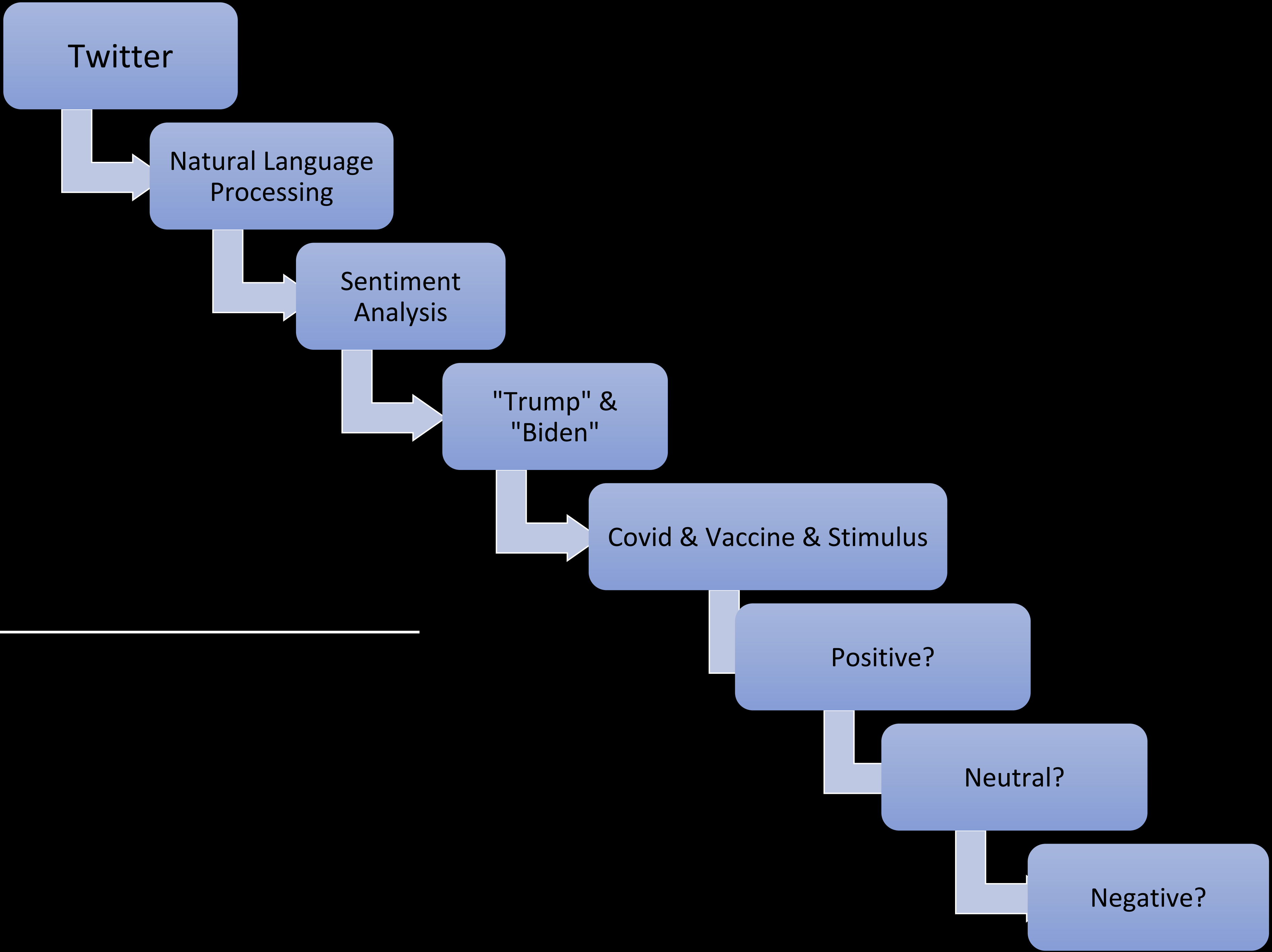
- Used Streaming API that allows to collect tweets on a real-time basis based on search terms, user ids or locations.

- This project will download tweets related to 5 keywords: "Covid", "stimulus", "Trump", "Biden", "vaccine".

- I intentionally selected these subjects because, these topics are the most popular keywords for last months and still they have more attractions than other issues in the thoughts of public.

- In order to access Twitter Streaming API, we need to get 4 pieces of information from Twitter: API key, API secret, Access token and Access token secret.

- I will be using a Python library called "Tweepy" to connect to Twitter API and download the data.



Twitter

Natural Language Processing

Sentiment Analysis

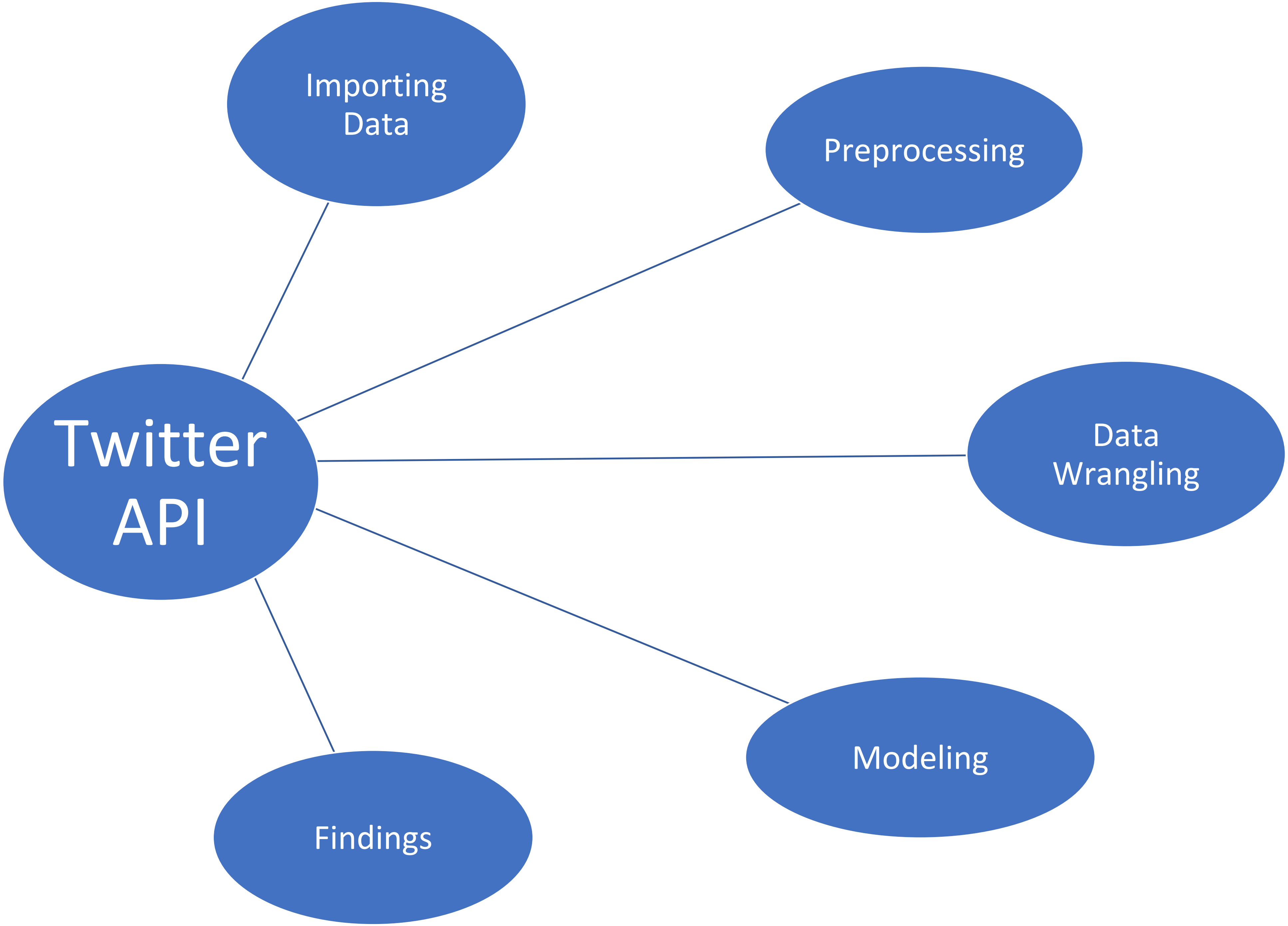
"Trump" & "Biden"

Covid & Vaccine & Stimulus

Positive?

Neutral?

Negative?



Twitter API

Importing Data

Preprocessing

Data Wrangling

Modeling

Findings

Importing & Preprocessing

- Parsed the response from the Twitter API into a structured table.

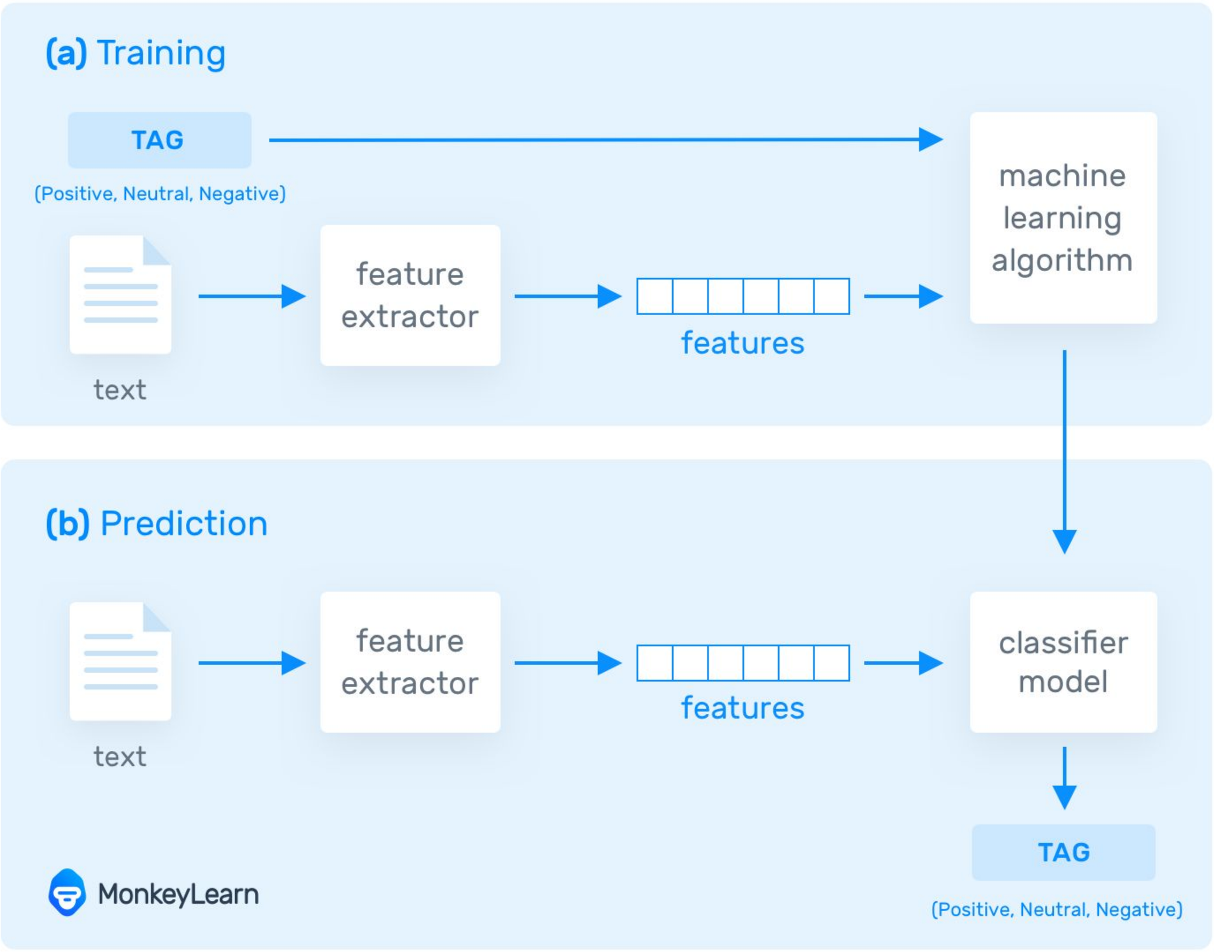
- Run the codes and get the output.

- The data stored various variables, only extract the "full_text", "id" and "the date (created_at)".

- The data has extracted "full_text" because when using 'extended' mode, the “text” is replaced by "full_text" attribute.

- Tweets have the exact same text but have been re-tweeted by different users. In order to extract a variety of hashtags from the tweets and to make the analysis unbiased, duplicate tweets were removed.

How Does Sentiment Analysis Work?



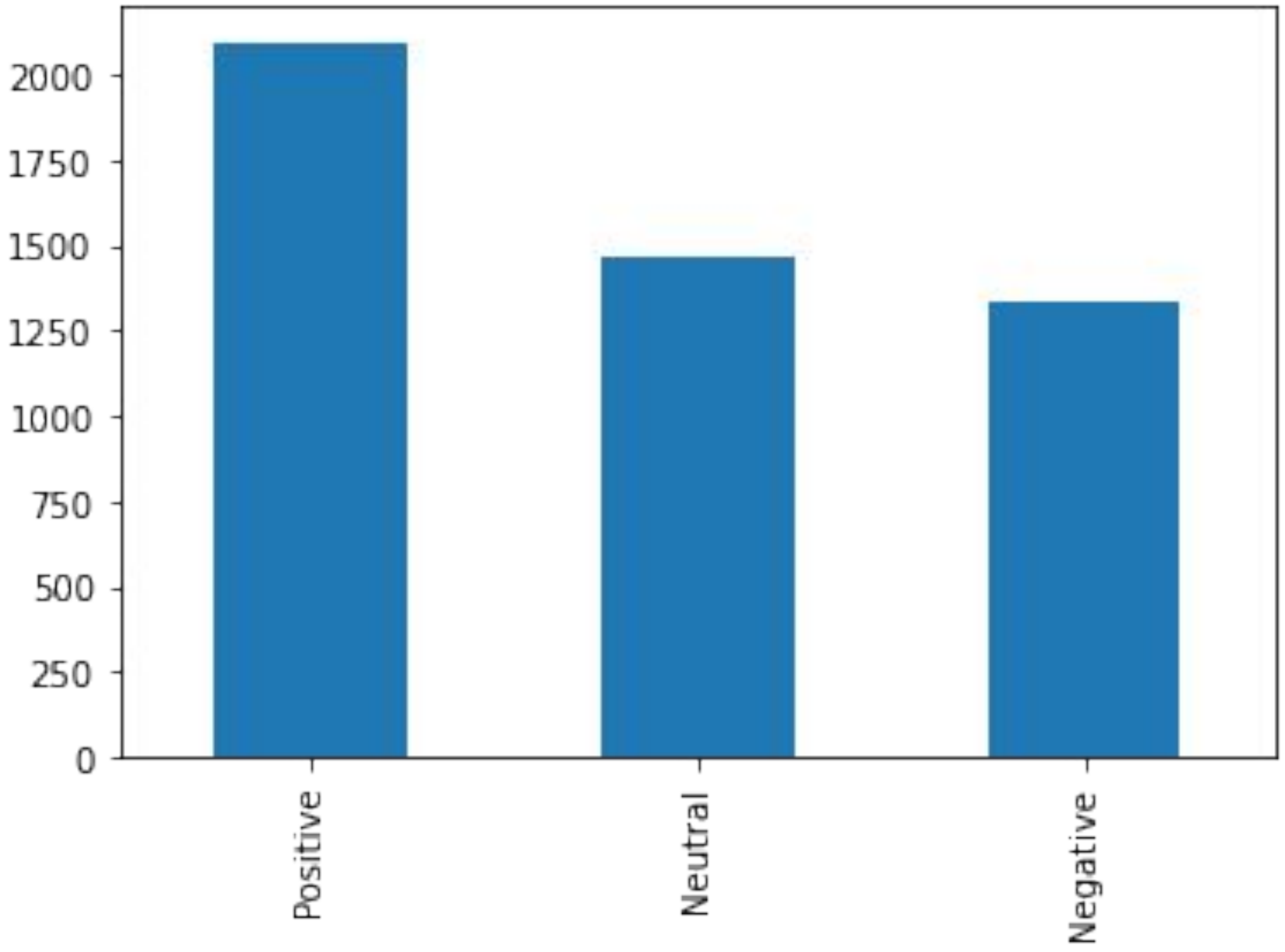
Exploratory Data Analysis

TRUMP

Positive => %43

Neutral => %30

Negative => %27

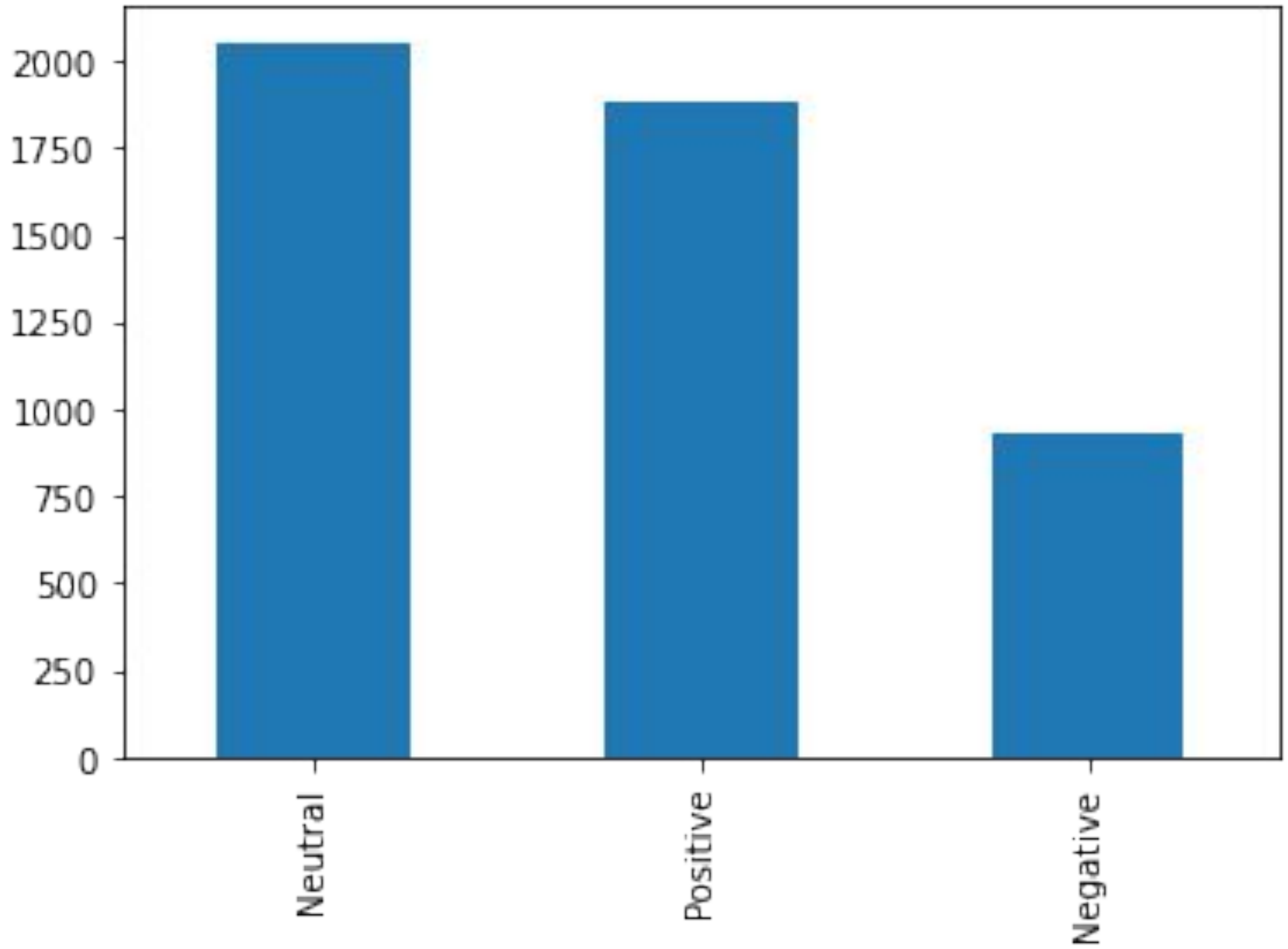


BIDEN

Positive => %39

Neutral => %42

Negative => %19

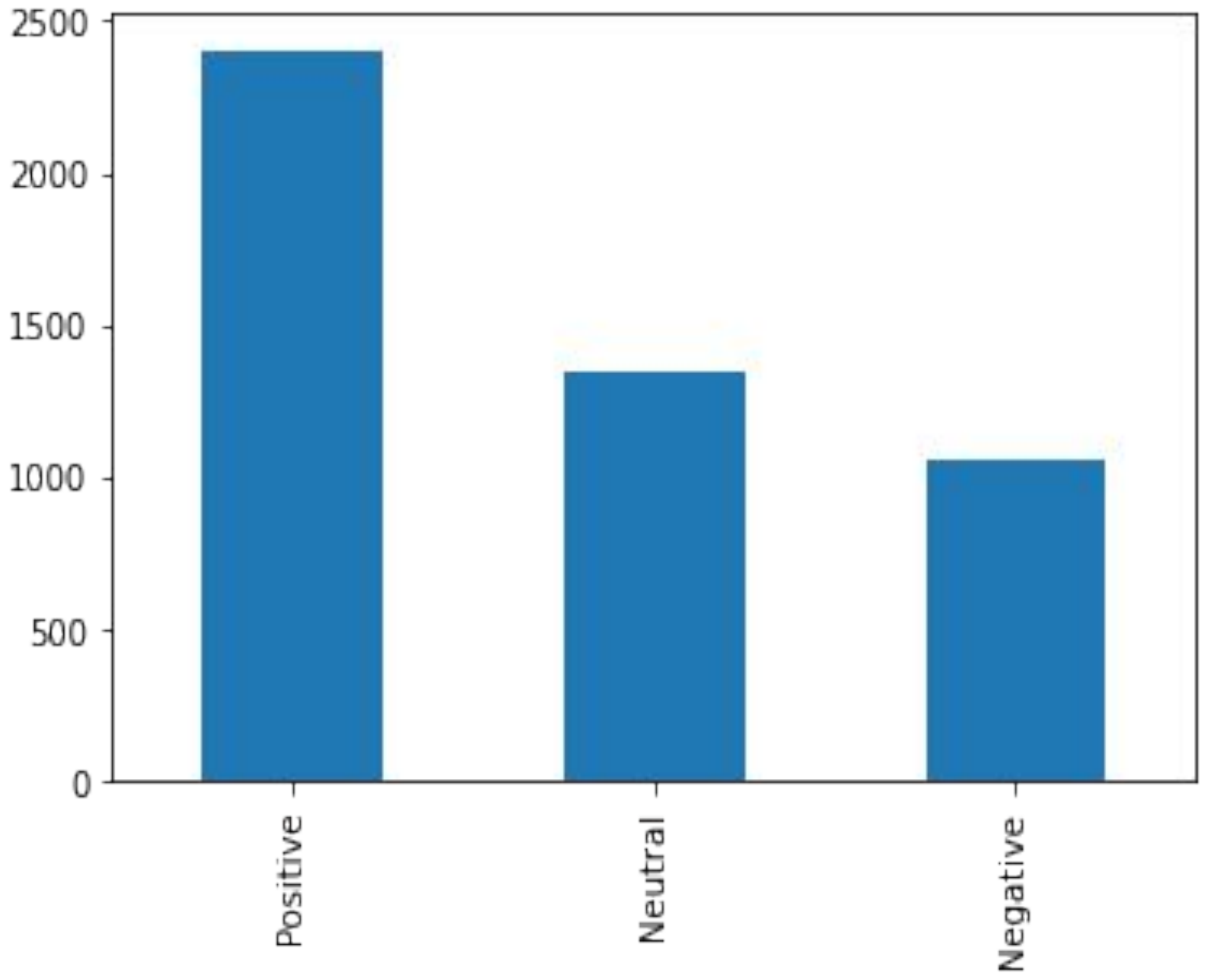


Covid 19

Positive => %50

Neutral => %28

Negative => %22

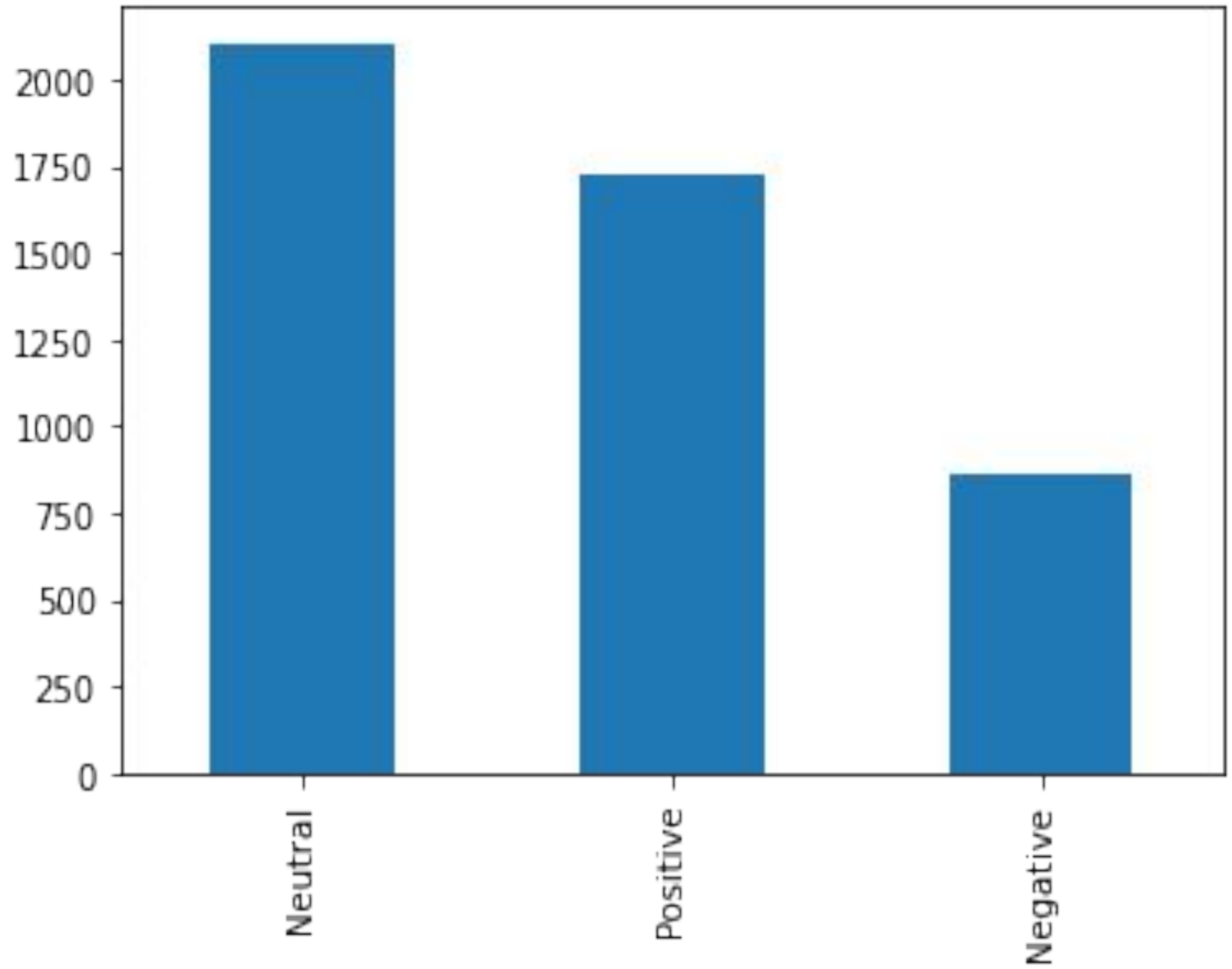


Stimulus

Positive => %36

Neutral => %45

Negative => %19

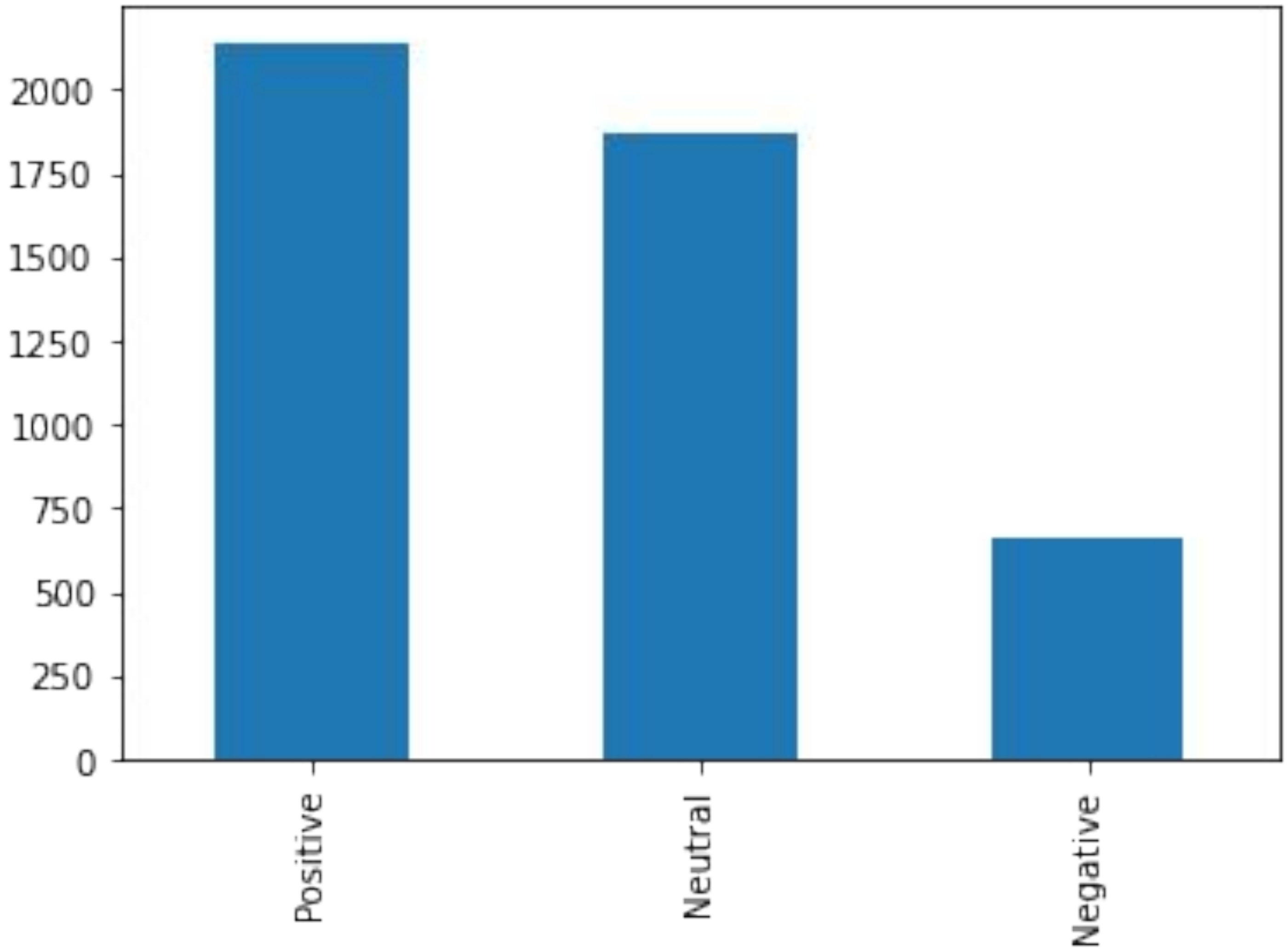


Vaccine

Positive => %46

Neutral => %40

Negative => %14



MODELING

Step 1) Removal of Stop Words (Cleaning)

Step 2) Stemming

Step 3) Lemmatization

Step 4) Sentiment of Lemmatized Data

Step 5) Dropping Irrelevant Columns

Step 1) Divide our dataset into feature and label sets

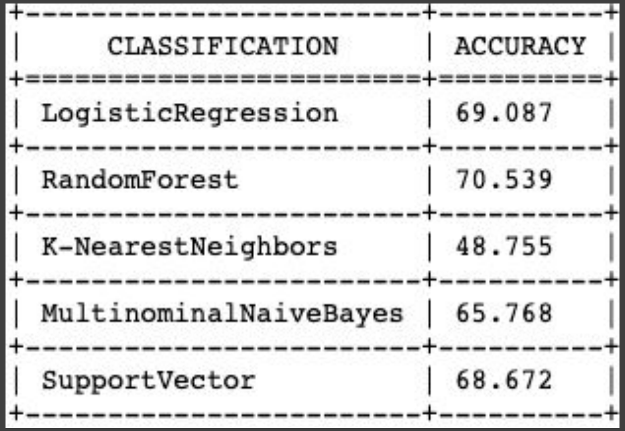
Step 2) Representing Text in Numeric Form (TF-IDF)

Step 3) Dividing Data into Training and Test Sets

Step 4) Training the Model

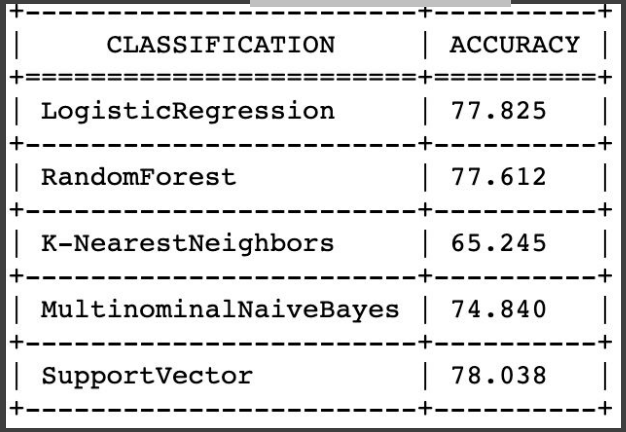
Step 5) Making Predictions and Evaluating the Model

Covid



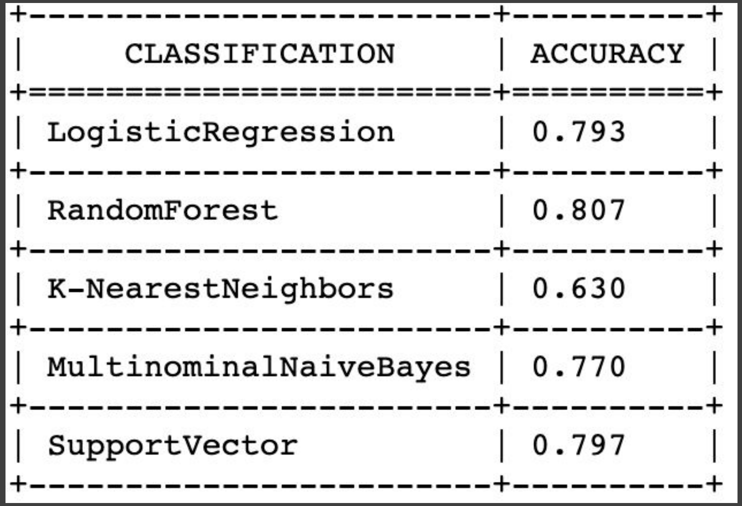
```
+------------------------+----------+
|     CLASSIFICATION     | ACCURACY |
+========================+==========+
| LogisticRegression     | 69.087   |
+------------------------+----------+
| RandomForest           | 70.539   |
+------------------------+----------+
| K-NearestNeighbors     | 48.755   |
+------------------------+----------+
| MultinominalNaiveBayes | 65.768   |
+------------------------+----------+
| SupportVector          | 68.672   |
+------------------------+----------+
```

Stimulus



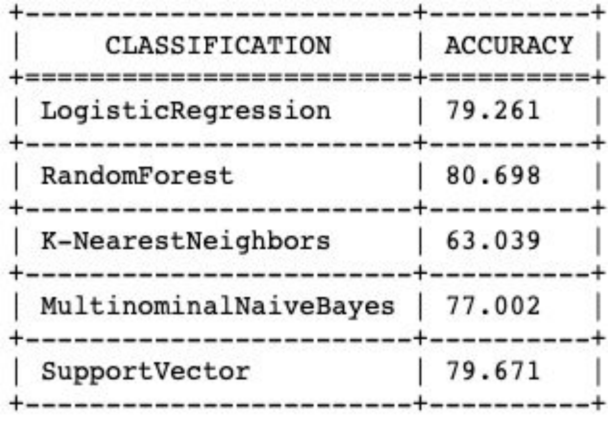
```
+------------------------+----------+
|     CLASSIFICATION     | ACCURACY |
+========================+==========+
| LogisticRegression     | 77.825   |
+------------------------+----------+
| RandomForest           | 77.612   |
+------------------------+----------+
| K-NearestNeighbors     | 65.245   |
+------------------------+----------+
| MultinominalNaiveBayes | 74.840   |
+------------------------+----------+
| SupportVector          | 78.038   |
+------------------------+----------+
```

Trump



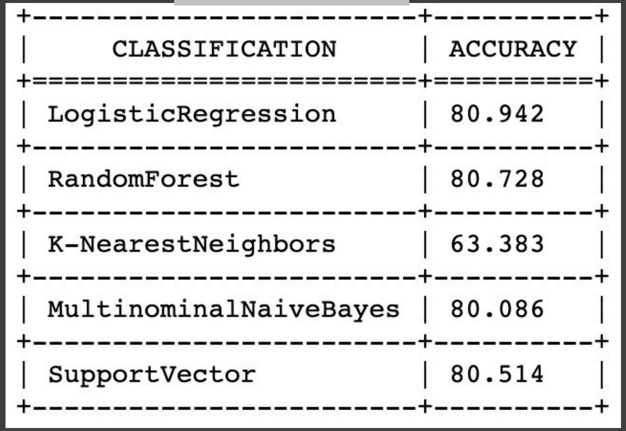
```
+------------------------+----------+
|     CLASSIFICATION     | ACCURACY |
+========================+==========+
| LogisticRegression     | 0.793    |
+------------------------+----------+
| RandomForest           | 0.807    |
+------------------------+----------+
| K-NearestNeighbors     | 0.630    |
+------------------------+----------+
| MultinominalNaiveBayes | 0.770    |
+------------------------+----------+
| SupportVector          | 0.797    |
+------------------------+----------+
```

Biden



```
+------------------------+----------+
|     CLASSIFICATION     | ACCURACY |
+========================+==========+
| LogisticRegression     | 79.261   |
+------------------------+----------+
| RandomForest           | 80.698   |
+------------------------+----------+
| K-NearestNeighbors     | 63.039   |
+------------------------+----------+
| MultinominalNaiveBayes | 77.002   |
+------------------------+----------+
| SupportVector          | 79.671   |
+------------------------+----------+
```

Vaccine



```
+------------------------+----------+
|     CLASSIFICATION     | ACCURACY |
+========================+==========+
| LogisticRegression     | 80.942   |
+------------------------+----------+
| RandomForest           | 80.728   |
+------------------------+----------+
| K-NearestNeighbors     | 63.383   |
+------------------------+----------+
| MultinominalNaiveBayes | 80.086   |
+------------------------+----------+
| SupportVector          | 80.514   |
+------------------------+----------+
```

FINDINGS

- Project retrieved tweets between 11/22/2020 and 11/23/2020.
- The highest number of negative tweets among the 5 subjects are about Trump and highest positive tweets among the 5 subjects are about Covid where twitter users have positive emotions about the "Covid".
- The dataset reveals that almost %43 of the tweets as a subject of "Trump" are positive, %29 are neutral and %27 of the tweets have negative perspective.
- As an opponent of "Trump" during the election, %42 of the tweets used "Biden" as subject are neutral, %39 of these tweets were positive, %19 of the tweets have negative perspective.
- As a 3rd selected subject, %50 of the tweets about "Covid" positive, %28 are neutral and %22 of the tweets have negative emotion.
- The dataset reveals that almost %45 of the tweets used "stimulus" as subject are neutral, %38 of these tweets were positive, %18 of the tweets have negative perspective.
- The dataset reveals that almost %46 of the tweets used "vaccine" in a positive manner, %40 of these tweets were neutral, %15 of the tweets have negative perspective.

- We performed an analysis of public tweets regarding 5 (five) trendy Twitter subjects

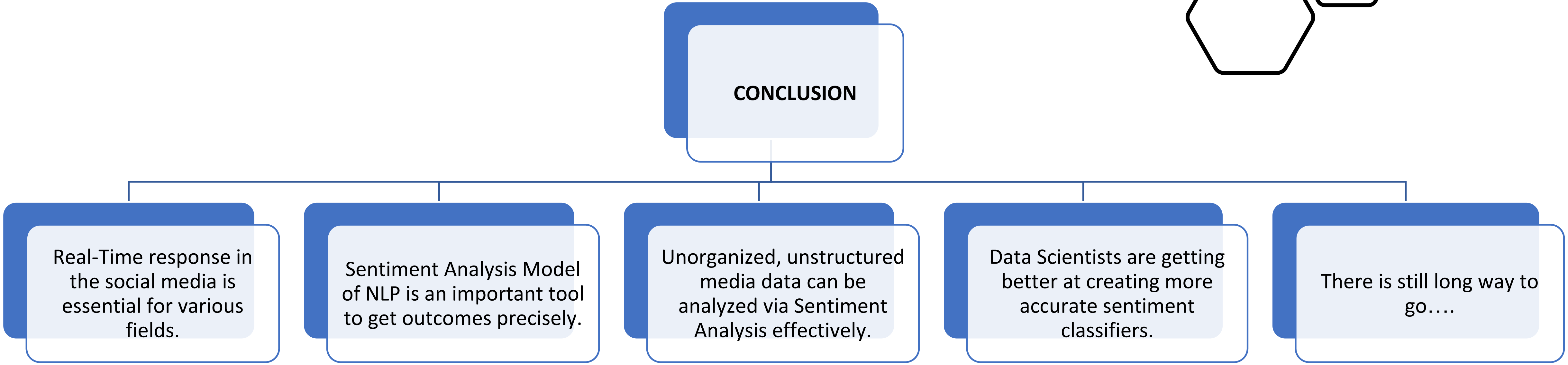
- This is a Sentiment Analysis with Classification

- Logistic Regression, Multinomial Naïve Bayes, KNN, Random Forest and Support Vector Machine Models were used

- Confusion Matrix to evaluate the accuracy of the models was utilized

- The Highest Accuracy scores are from Logistic Regression and Random Forest Models with mostly over 75％ .

- The lowest scores are from K-Nearest Neighbors Model with average 55％.



CONCLUSION

- Real-Time response in the social media is essential for various fields.
- Sentiment Analysis Model of NLP is an important tool to get outcomes precisely.
- Unorganized, unstructured media data can be analyzed via Sentiment Analysis effectively.
- Data Scientists are getting better at creating more accurate sentiment classifiers.
- There is still long way to go….