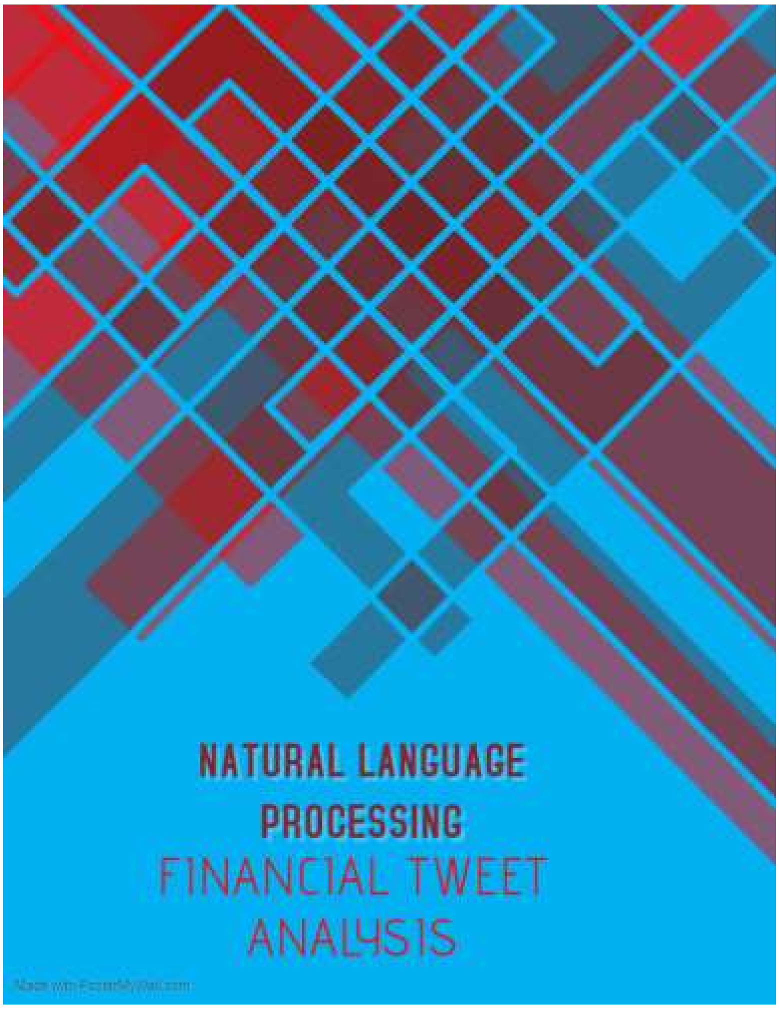
FA 692 NATURAL LANGUAGE PROCESSING IN FINANCE



Written By-

Jayesh Kartik
CWID-20012210
MS in financial Analytics
Stevens Institute of Technology, Hoboken, New Jersey
Under the guidance-Dr. Zach Feinstein
Date-24thApril'2023

CONTENTS

ABSTRACT

This research work centres on the application of Natural Language Processing (NLP) techniques for the analysis of financial tweets. The provided dataset encompasses a voluminous assortment of

more than 28,000 tweets that pertain to publicly traded companies as well as cryptocurrencies. These tweets are marked by the corresponding company symbol and other pertinent fields. The initiation of the project involves executing Extract, Transform, Load (ETL) operations on Python, which entail the conversion of the CSV file into a data frame. Subsequently, an Exploratory Data Analysis (EDA) is carried out to gain a comprehensive understanding of the data through visual means. The effective analysis of data frames necessitates the application of techniques such as text cleaning and tokenization. A polarity score is derived from the analysis of sentiment exhibited in all tweets. A word cloud is developed to show the frequency of the words that has appeared on the financial tweets. The study delves into the realm of Natural Language Processing (NLP) by employing machine learning methodologies such as logistic regression, Random Forest, and Multinomial Naive Bayes. The data frame was subjected to K-means cluster analysis. The assessment of algorithmic proficiency typically involves the calculation of accuracy and confusion matrix measures to ascertain the efficacy of said methods. In general, this study offers valuable perspectives on the utilization of Natural Language Processing (NLP) and machine learning techniques within the financial sector.

MOTIVATION

The motivation for the present undertaking is to investigate the utilization of Natural Language Processing (NLP) in the realm of finance. The expanding adoption of social media within the finance sector has led to a plethora of data accessible in the form of tweets. Such data holds the potential to furnish advantageous perspectives pertaining to market trends and sentiments. Conducting sentiment analysis on financial tweets has the potential to evaluate the prevailing public mindset toward a corporation or cryptocurrency, and this information can hold significant implications for investors and traders. The objective of this endeavour is to utilize natural language processing (NLP) methods and machine learning algorithms to get meaningful insights from an extensive financial tweet dataset. The knowledge gained from the analysis may be effectively utilized to facilitate well-informed investment decision-making practices, as well as enhance the efficiency of financial models. The project evinces the capability of Natural Language Processing (NLP) in the sphere of finance and its efficiency in thus deriving significant inferences from extensive volumes of unformatted data.

I have the dataset because:

- The dataset contains many financial tweets that are publicly available, making it a valuable source of information for sentiment analysis and market trends.
- The dataset is tagged with the company symbol, which enables the analysis of tweets about specific companies and cryptocurrencies.
- The dataset provides an opportunity to apply various NLP techniques and machine learning algorithms to gain insights into the financial domain.
- The dataset is relevant to the field of finance and can be used to make informed investment decisions.
- The dataset is readily available and easy to access, which reduces the effort required for data collection.

INTRODUCTION

The field of Natural Language Processing (NLP) has experienced substantial growth and has garnered noteworthy attention in contemporary times owing to its diverse applications in various areas of study. The realm of finance constitutes an area in which Natural Language Processing (NLP) techniques have the potential to be applied towards examining the sentiments conveyed within financial tweets, thereby facilitating judicious investment decisions. The aim of this project endeavour is to employ Natural Language Processing (NLP) techniques in scrutinizing an extensive compilation of financial tweets and conduct sentiment analysis as a means of acquiring a comprehensive understanding of market trends and emotional states.

The dataset utilized in the current project encompasses more than 28,000 financial tweets that are publicly accessible and classified according to the corresponding company symbol. The Twitter posts pertain to corporations that are listed on the stock market, as well as a limited number of digital currencies. Such tweets present a substantial resource for analytical purposes. The Python programming language was utilized to convert the dataset into a data frame. To enhance comprehension of the data, an Exploratory Data Analysis (EDA) was conducted.

The project is motivated by multiple factors, the exact nature of which might be delineated as follows:

- The utilization of social media across diverse industrial sectors has undergone a substantial surge in the recent years, where tweets have emerged as a pivotal source of information for the purpose of decision-making. The scrutiny of tweets has the potential to yield significant contributions towards the comprehension of customer perspectives, market patterns, and attitudes.
- The current job market places a high degree of importance on the acquisition of practical knowledge, particularly for individuals studying or working in analytical roles, pertaining to the utilization of natural language processing techniques and machine learning algorithms for the purpose of data analysis.
- The present project endeavour presents a chance to employ several techniques of Natural Language Processing (NLP), comprising text cleansing, tokenization, sentiment analysis, and machine learning algorithms, to a practical dataset. This undertaking offers significant prospects for developing skills and learning.
- The dataset is conveniently accessible and presents a prospect for the examination of the utilization of Natural Language Processing (NLP) methodologies and machine learning (ML) algorithms to conduct financial analysis, a subject that holds a great significance to financial experts and investors.
- The disquisitions obtained from this scrutiny may be utilized to make judicious investment verdicts, enhance the functionality of fiscal paradigms, and attain an advantageous position in the market.

ABOUT THE DATASET –

Link for the dataset-

https://www.kaggle.com/datasets/davidwallach/financial-tweets?resource=download

I have put it in a google drive folder. The google drive link is attached here-

https://drive.google.com/file/d/1A4rtKICXyEmhSELDDyFECOtZlP06X2Z4/view?usp=sharing

There are 28268 tweets and 8 columns in the dataset-

The **id** column is representative of the distinct identification code that is linked to every individual tweet. The numerical value, which is produced through automatic means by the system utilized by Twitter, represents a significant aspect of the platform's functionality.

The **text** column denotes the factual content conveyed in each individual tweet. The primary focus of the analysis pertains to the message that has been posted by the user on Twitter.

The **timestamp** column in question denotes the chronological data pertaining to both the time and date of the tweet's initial publication. This text comprises data pertaining to the precise temporal occurrence of the tweet, which may facilitate the scrutiny of patterns that emerge in users' posting tendencies.

The **source** column denotes the origin of the tweet, encompassing the platform or device whereupon the tweet was disseminated. The acquisition of knowledge regarding how users are accessing Twitter, as well as their interactions with the platform, can prove to be beneficial.

The **symbol** column denotes the alphanumeric codes that serve as identifiers for the stocks referenced within the tweet. This analytical tool has the capacity to scrutinize the emotional tone revolving around corporations or shares, thus proving advantageous in the domain of finance.

The **company** column in question pertains to the appellation of the business that is linked to the stock symbol indicated in the Twitter post. The utility of sentiment analysis lies in its ability to scrutinize and gauge public sentiment surrounding individual corporations and/or stocks, thus rendering it an advantageous tool for financial scrutiny and evaluation.

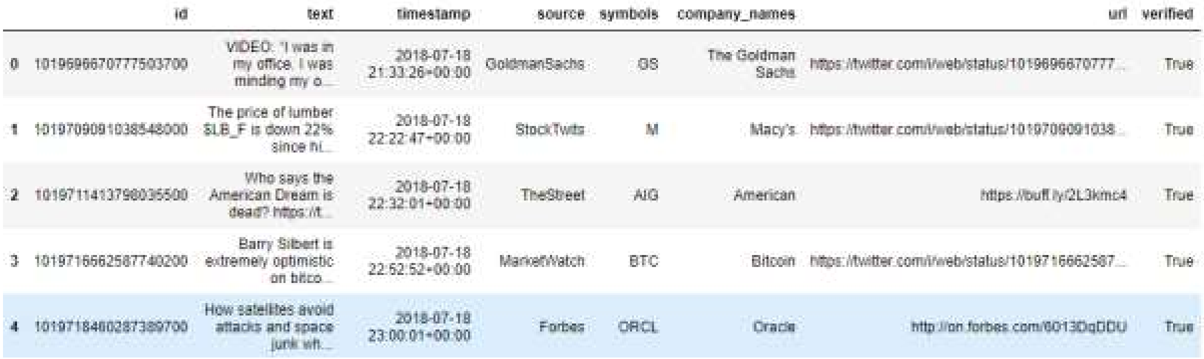
The **URL** column denotes the Uniform Resource Locator (URL) linked with the respective tweet. The functionality of retrieving a tweet directly from Twitter has the potential to enhance the analytical process by enabling the examination of the sharing and retweeting mechanisms employed by users.

This **verified** column denotes the verification status of the account holder who authored the tweet, indicating whether they have undergone the confirmation process administered by Twitter to authenticate the account's identity. A verified account serves as an indication of the user's status as a public figure or celebrity and can prove to be instrumental in scrutinizing the prevailing influence or impact of a tweet.

EXPLORATORY DATA ANALYSIS-

Analysing and comprehending a dataset with the intention of finding intriguing patterns, trends, and insights is the process of data exploration. Data exploration in the context of stock market data often entails looking at historical price and volume data for a certain stock.

I have downloaded the data from the above link. The data is saved in the CSV format. I uploaded the CSV file in the data frame. The data has been explored in python using the NumPy and Panda Library. The following screenshot shows the ‘head’ of the dataset-



	id	text	timestamp	source	symbols	company_names	url	verified
0	1019696670777503700	VIDEO: "I was in my office. I was minding my o...	2018-07-18 21:33:26+00:00	GoldmanSachs	GS	The Goldman Sachs	https://twitter.com/i/web/status/1019696670777...	True
1	1019709091038548000	The price of lumber $LB_F is down 22% since hi...	2018-07-18 22:22:47+00:00	StockTwits	M	Macy's	https://twitter.com/i/web/status/1019709091038...	True
2	1019711413798035500	Who says the American Dream is dead? https://t...	2018-07-18 22:32:01+00:00	TheStreet	AIG	American	https://buff.ly/2L3kmc4	True
3	1019716662587740200	Barry Silbert is extremely optimistic on bitco...	2018-07-18 22:52:52+00:00	MarketWatch	BTC	Bitcoin	https://twitter.com/i/web/status/1019716662587...	True
4	1019718460287389700	How satellites avoid attacks and space junk wh...	2018-07-18 23:00:01+00:00	Forbes	ORCL	Oracle	http://on.forbes.com/6013DqDDU	True

DATA VISUALISATION

The practice of visualizing data involves the representation of information through graphical or pictorial means with the aim of enhancing comprehension and interpretation amongst its audience. In the given financial tweet's dataset, the application of data visualization can facilitate a more lucid comprehension of the prevalent trends and patterns within the dataset.

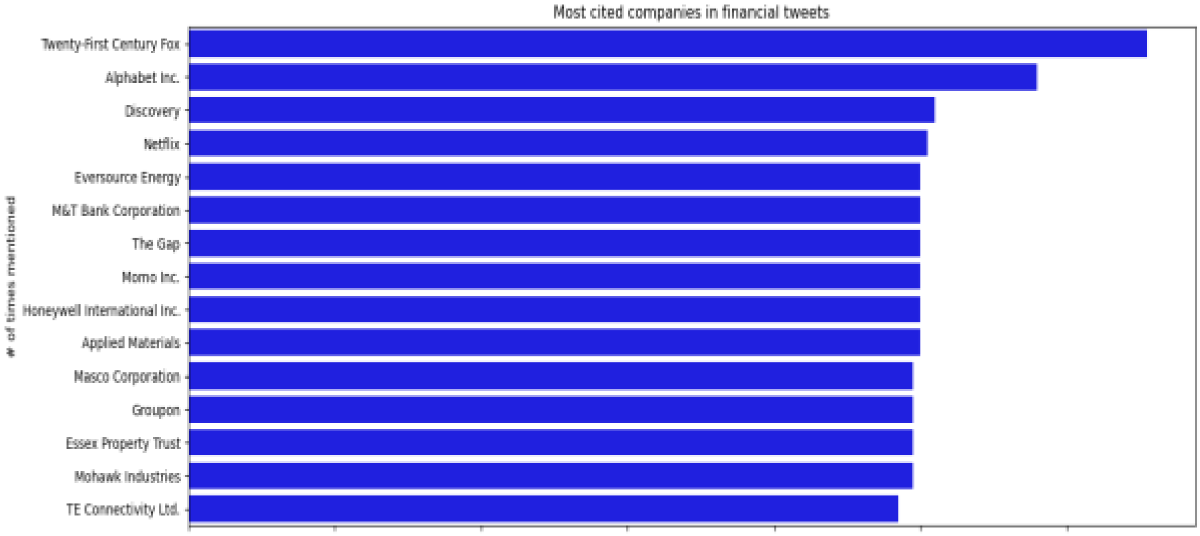


Figure 1 shows the most cited companies in financial tweets.

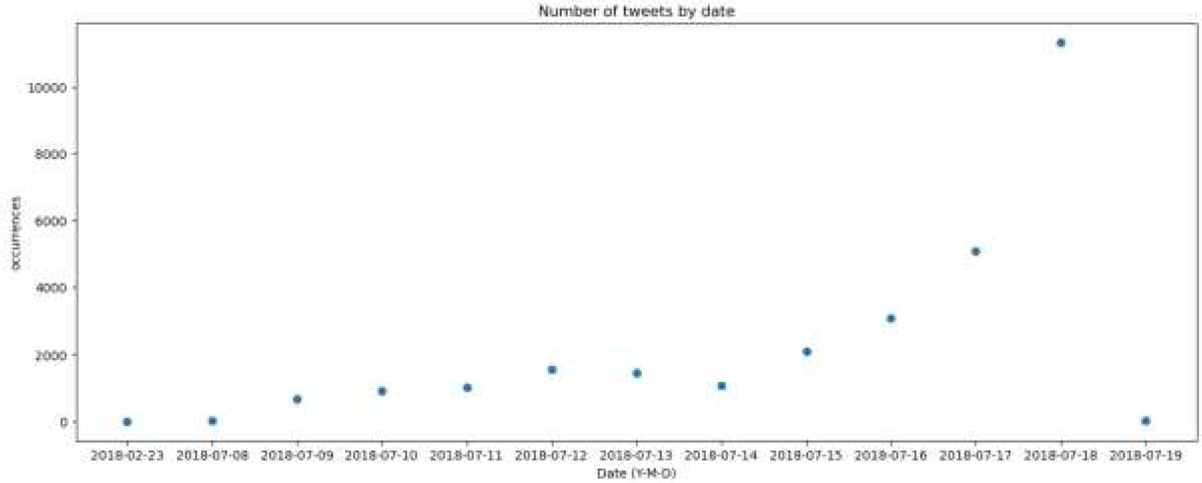


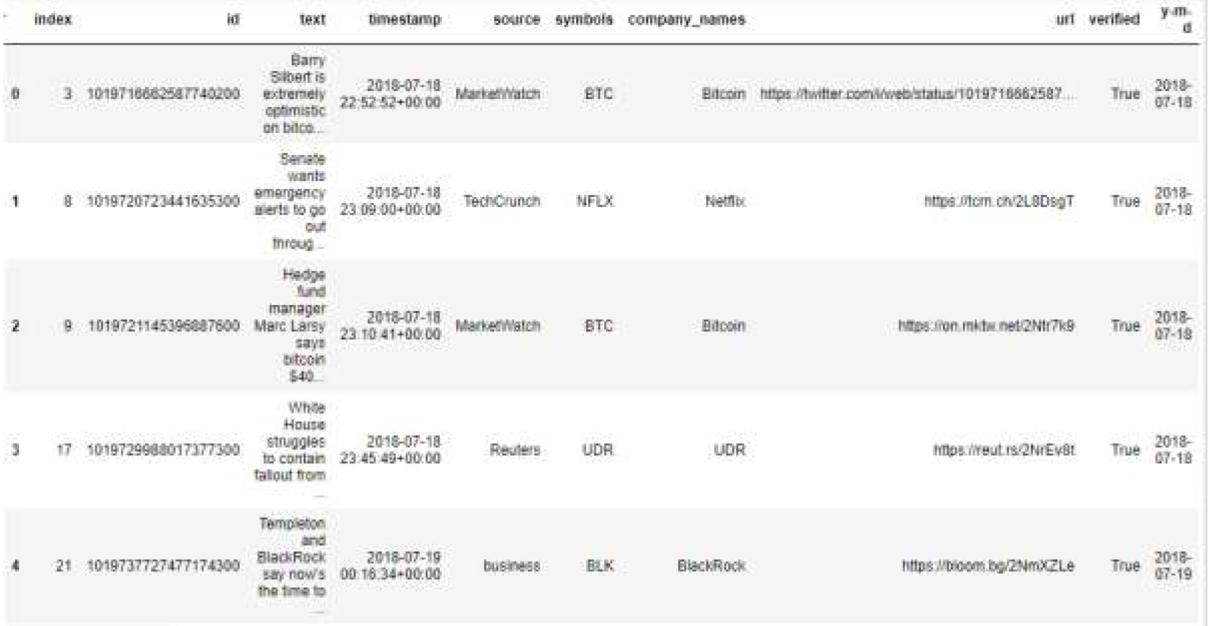
Figure 2 shows the date with the frequency of tweets.

TEXT CLEANING

In the next step, I created a new DataFrame called data_filtered that contains the filtered data.I have considered the date in the 'timestamp' column is after or equal to July 9, 2018.

I considered only those companies that have been mentioned in at least 50 tweets and then the company names are extracted and stored in a Python list called highly_mentionned_companies.I filtered the data_filtered DataFrame to include only the rows where the company name in the 'company_names' column is in the highly_mentionned_companies list.

Lastly using the I called reset_index() method on data_filtered to reset the row index of the DataFrame, since some rows may have been removed during the filtering process.



	index	id	text	timestamp	source	symbols	company_names	url	verified	y-m-d
0	3	1019716662587740200	Barry Silbert is extremely optimistic on bitco...	2018-07-18 22:52:52+00:00	MarketWatch	BTC	Bitcoin	https://twitter.com/i/web/status/1019716662587...	True	2018-07-18
1	8	1019720723441635300	Senate wants emergency alerts to go out throug...	2018-07-18 23:09:00+00:00	TechCrunch	NFLX	Netflix	https://tcrn.ch/2L8DsgT	True	2018-07-18
2	9	1019721145396887600	Hedge fund manager Marc Larsy says bitcoin $40...	2018-07-18 23:10:41+00:00	MarketWatch	BTC	Bitcoin	https://on.mktw.net/2Ntr7k9	True	2018-07-18
3	17	1019729988017377300	White House struggles to contain fallout from ...	2018-07-18 23:45:49+00:00	Reuters	UDR	UDR	https://reut.rs/2NrEv8t	True	2018-07-18
4	21	1019737727477174300	Templeton and BlackRock say now's the time to ...	2018-07-19 00:16:34+00:00	business	BLK	BlackRock	https://bloom.bg/2NmXZLe	True	2018-07-19

In the next step, I implemented some basic text preprocessing steps that are commonly used in Natural Language Processing (NLP) tasks-

The remove_punctuation() function takes a text string as input and returns a new string with all punctuation marks removed. The preprocess_txt() function takes a text string as input, removes any URLs contained in the text, converts all characters to lowercase, and then removes all punctuation marks using the remove_punctuation() function. These preprocessing steps are used to clean and standardize the text data, making it easier to analyze using NLP techniques.

NLP is a field of study that involves developing algorithms and models that can analyze and understand human language. Some common tasks in NLP include sentiment analysis, named entity recognition, and language translation. In order to perform these tasks, the text data must be preprocessed to remove noise and standardize the format of the text. Once the text has been preprocessed, it can be tokenized (split into individual words or phrases), analyzed using statistical models, and used to train machine learning algorithms.

I performed frequency analysis on the preprocessed text data by first concatenating all the preprocessed text and then tokenizing it using the word_tokenize function of the nltk library. The frequency distribution of the tokens is then calculated using the FreqDist function of the same library. The resulting distribution is converted into a Pandas dataframe and this analysis gives insights into the most commonly used words in the financial tweets, which can be helpful in understanding the topics and sentiments discussed in the tweets.

TEXT ANALYSIS

Before implementing machine learning models on the data, I used the TfidfVectorizer from the sklearn library to convert the pre-processed text data into a numerical vector representation, which can be used as input for machine learning models.

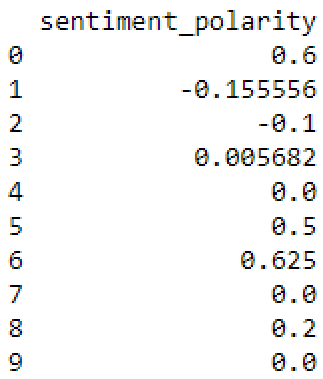
TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure that reflects the importance of a term in a document or corpus, considering how frequently it appears in the given text and how rare it is in the corpus. The TfidfVectorizer function computes the TF-IDF values of each token in the text data and generates a sparse matrix representation of the data.

The use of TF-IDF is essential in natural language processing (NLP) and text mining because it helps to identify the most informative and relevant tokens in the dataset, while filtering out the noise and common words that do not provide much useful information. By encoding the text data into a numerical format, it enables the use of various machine learning algorithms, such as classification or clustering, to analyse and derive insights from the data.

In the given context, this step is essential for analysing the Twitter data and identifying patterns or trends in the tweets related to the highly mentioned companies. The encoded sparse matrix can be used to train and evaluate machine learning models that can classify or cluster the tweets based on their content and other features.

1. SENTIMENT ANALYSIS

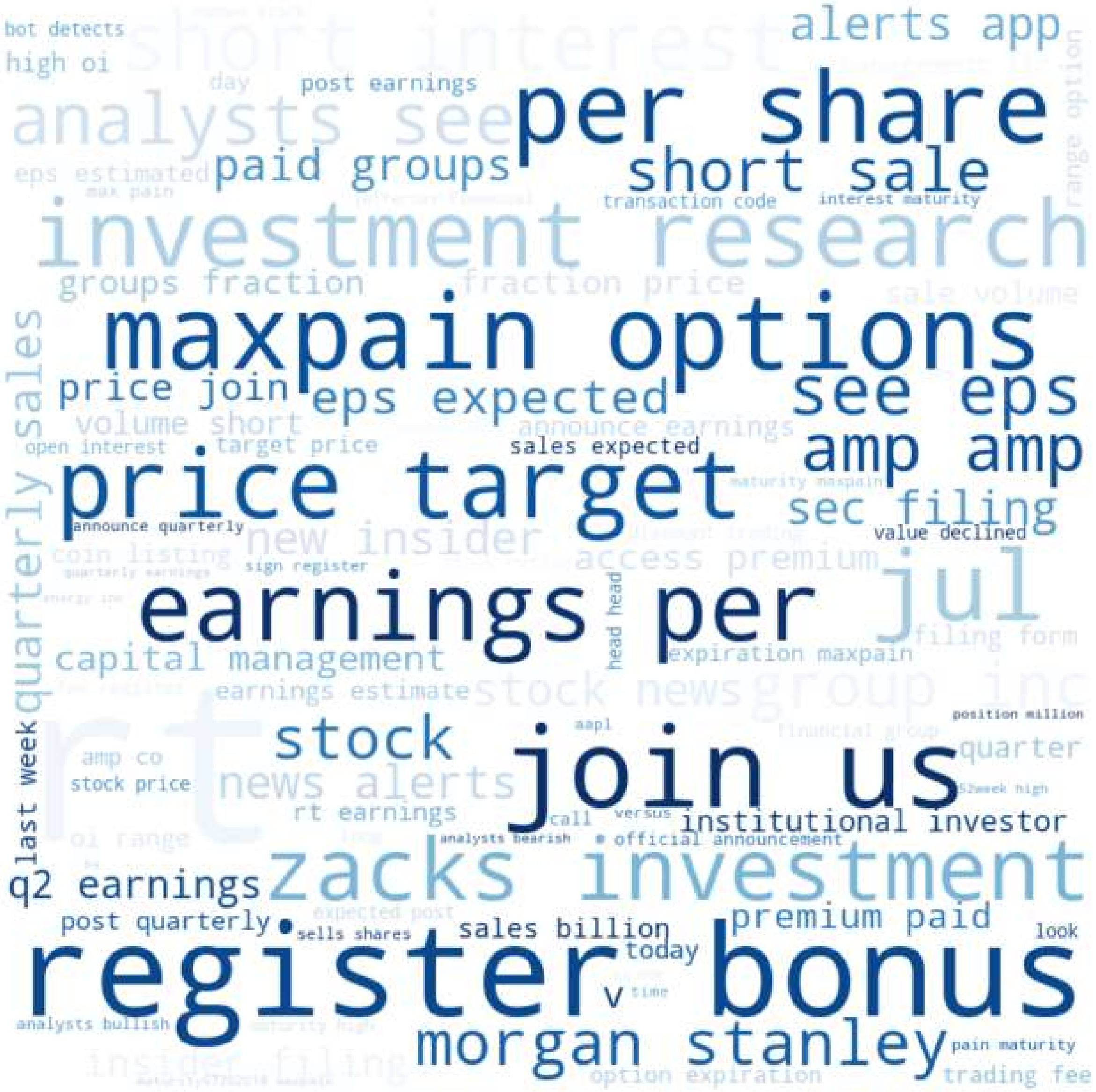
I performed sentiment analysis on the 'text' column of the 'data' data frame using the TextBlob library. Firstly, I created a new column called 'sentiment_polarity' in the 'data' data frame to store the sentiment polarity score for each tweet. Then I iterated over each row in the 'data' data frame, procured the text of the tweet, performed sentiment analysis using TextBlob, and stored the sentiment polarity score in the 'sentiment_polarity' column for that particular tweet. Finally, I printed the first 10 rows of the 'data' data frame to display the sentiment polarity scores for those tweets.



```
   sentiment_polarity
0                 0.6
1           -0.155556
2                -0.1
3            0.005682
4                 0.0
5                 0.5
6               0.625
7                 0.0
8                 0.2
9                 0.0
```

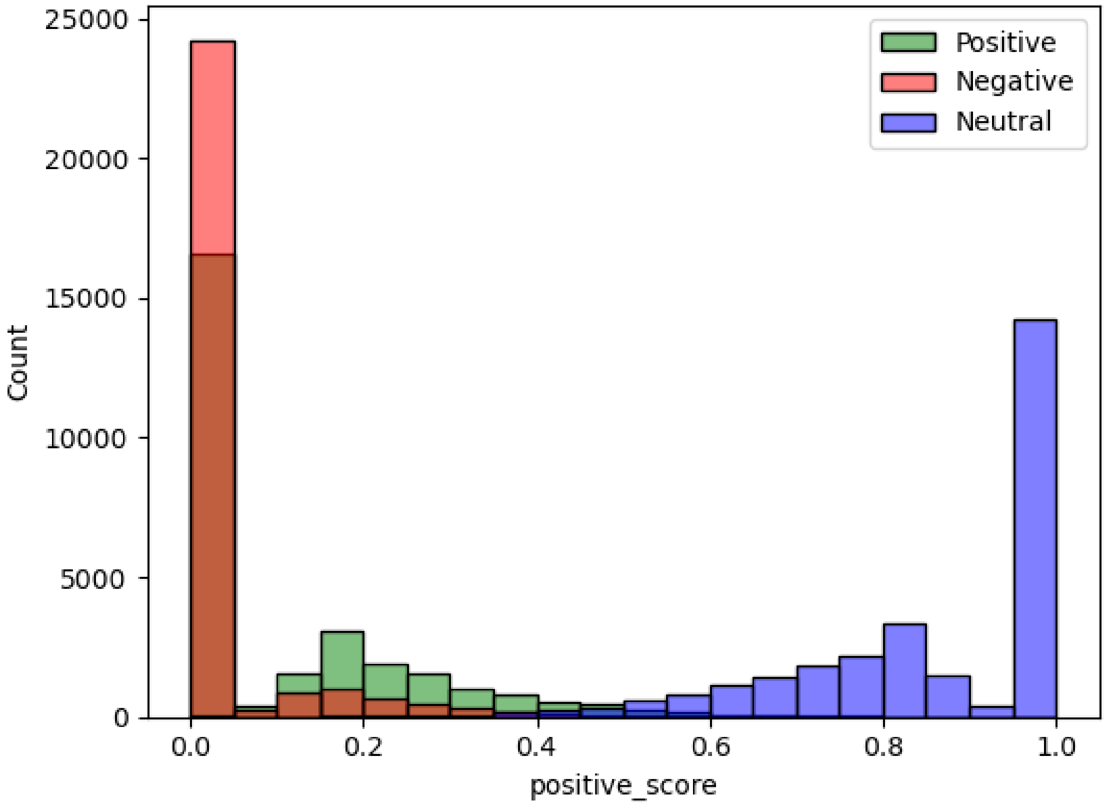
2.WORD CLOUD ANALYSIS

Until now, I have removed URLs and punctuation marks from the text data. I tokenized the text by splitting it into individual words, converts each word to lowercase, and removes stop words using the Natural Language Toolkit (NLTK) library. I stored in the 'text' column of the Data Frame. I finally wanted to implement word cloud analysis on the data as it gives us deeper understanding of the data. The Word Cloud object is initialized with the specified parameters, including the width and height of the word cloud image, the maximum number of words to display, and the color map to use.



3.SENTIMENT ANALYSIS(CONTINUED)-

The above code is performing sentiment analysis on a dataset of tweets. It first initializes the sentiment analyzer and applies it to each tweet in the dataset, storing the positive, negative, and neutral scores in separate columns of the data frame. The code then classifies each tweet as positive, negative, or neutral based on the sentiment scores, and plots the distribution of sentiment scores using a histogram. Finally, the code prints the number of positive, negative, and neutral tweets in the dataset. The sentiment analysis and visualization provide insights into the overall sentiment of the tweets, which can be useful for various applications such as market research or public opinion analysis.



DATA SPLITTING

I have split the dataset into training and testing sets. I am using the train_test_split() function from scikit-learn to randomly split the data Data Frame into two sets, X_train and X_test containing the tweet text data, and y_train and y_test containing the corresponding sentiment labels. The test_size parameter specifies the proportion of the dataset that should be reserved for testing (in this case, 20%), and the random_state parameter sets the seed for the random number generator to ensure that the results are reproducible.

DATA MODELLING

Data modelling refers to the process of creating a conceptual or mathematical representation of a set of data. The goal of data modelling is to create a structure that can be used to organize, understand, and analyse the data.

Before starting to model the data, I have implemented a code which is vectorizing the text data using the CountVectorizer method from the scikit-learn library. First, I have created a CountVectorizer object which will be used to transform the text data into a matrix of word counts. I then called the fit_transform method of the vectorizer on the training data (X_train), which fits the vectorizer to the training data and transforms it into a matrix of word counts. I have stored the resultant data in the X_train_vec variable. Then, I called CountVectorizer on the test data (X_test), which transforms it into a matrix of word counts using the vocabulary learned from the training data. The resulting matrix is stored in the X_test_vec variable.

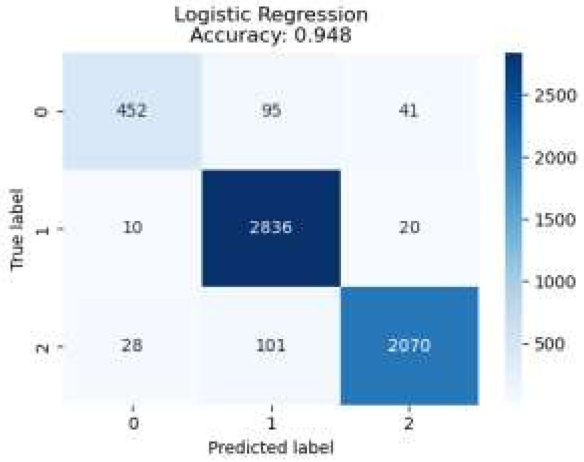
Logistic Regression Model-

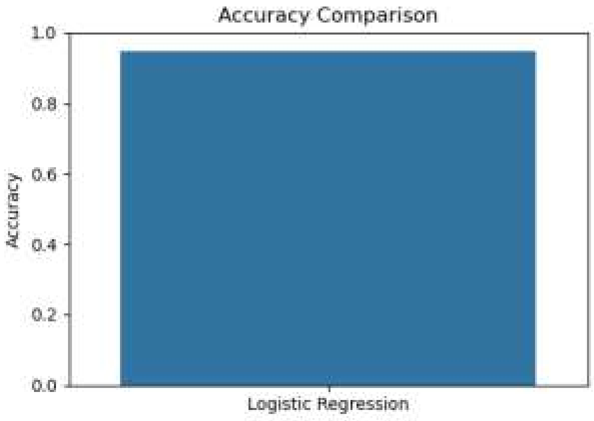
The logistic regression machine learning model is used to classify the sentiment of text data as positive, negative, or neutral. It is a supervised learning algorithm that learns from the labelled training data and then makes predictions on the test data.

The logistic regression machine learning model is used to classify the sentiment of text data as positive, negative, or neutral. It is a supervised learning algorithm that learns from the labelled training data and then makes predictions on the test data.

I have implemented a code which is training and testing a logistic regression model. First, I have created an instance of the logistic regression model using the default parameters. Then, the fit() method of the model is called to train the model on the training data. The trained model is then used to predict the sentiment of the test data using the predict() method.

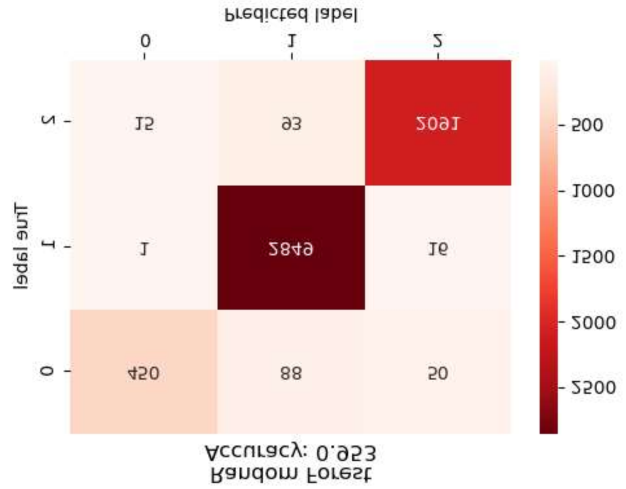
The accuracy_score() function is used to compute the accuracy of the model's predictions on the test data. The f1_score () function is used to compute the F1 score, which is a measure of the model's performance that takes into account both precision and recall. Finally, the confusion_matrix() function is used to compute the confusion matrix, which shows the number of true positives, false positives, true negatives, and false negatives of the model's predictions on the test data.

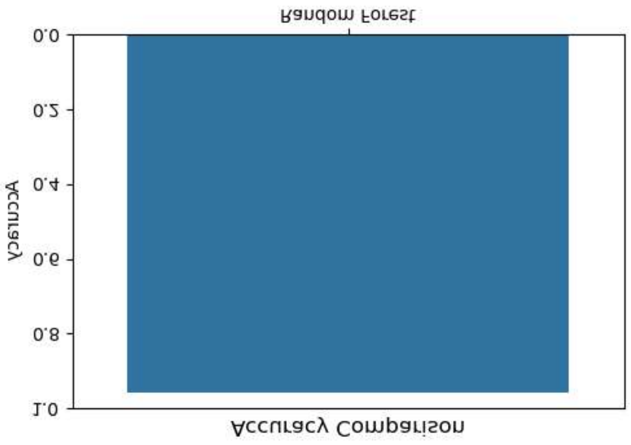




RANDOM FOREST ALGORITHM

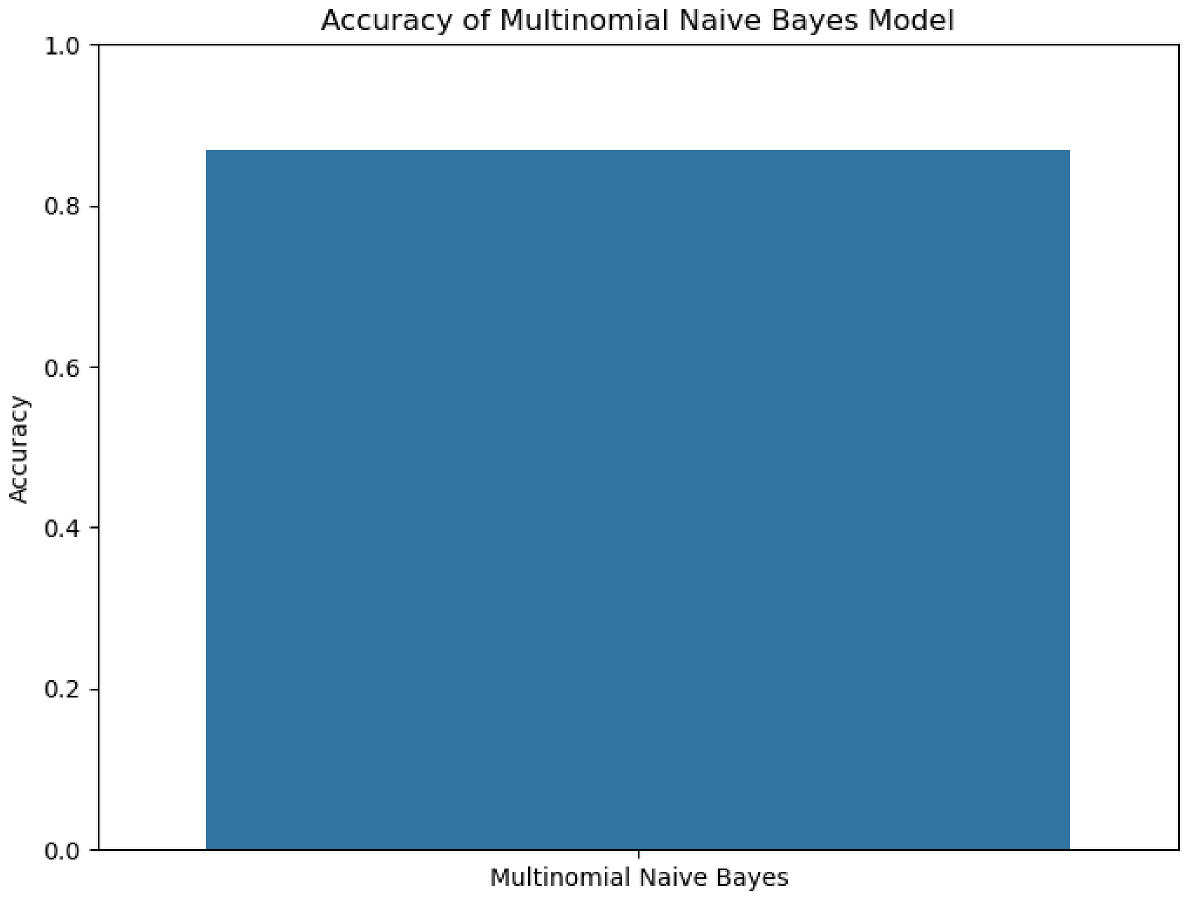
Random Forest is a powerful and versatile machine learning algorithm that is capable of handling both classification and regression tasks. It works by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests are often used in natural language processing tasks such as sentiment analysis because they can handle large feature spaces and noisy data, and they are less prone to overfitting than other machine learning models like decision trees.

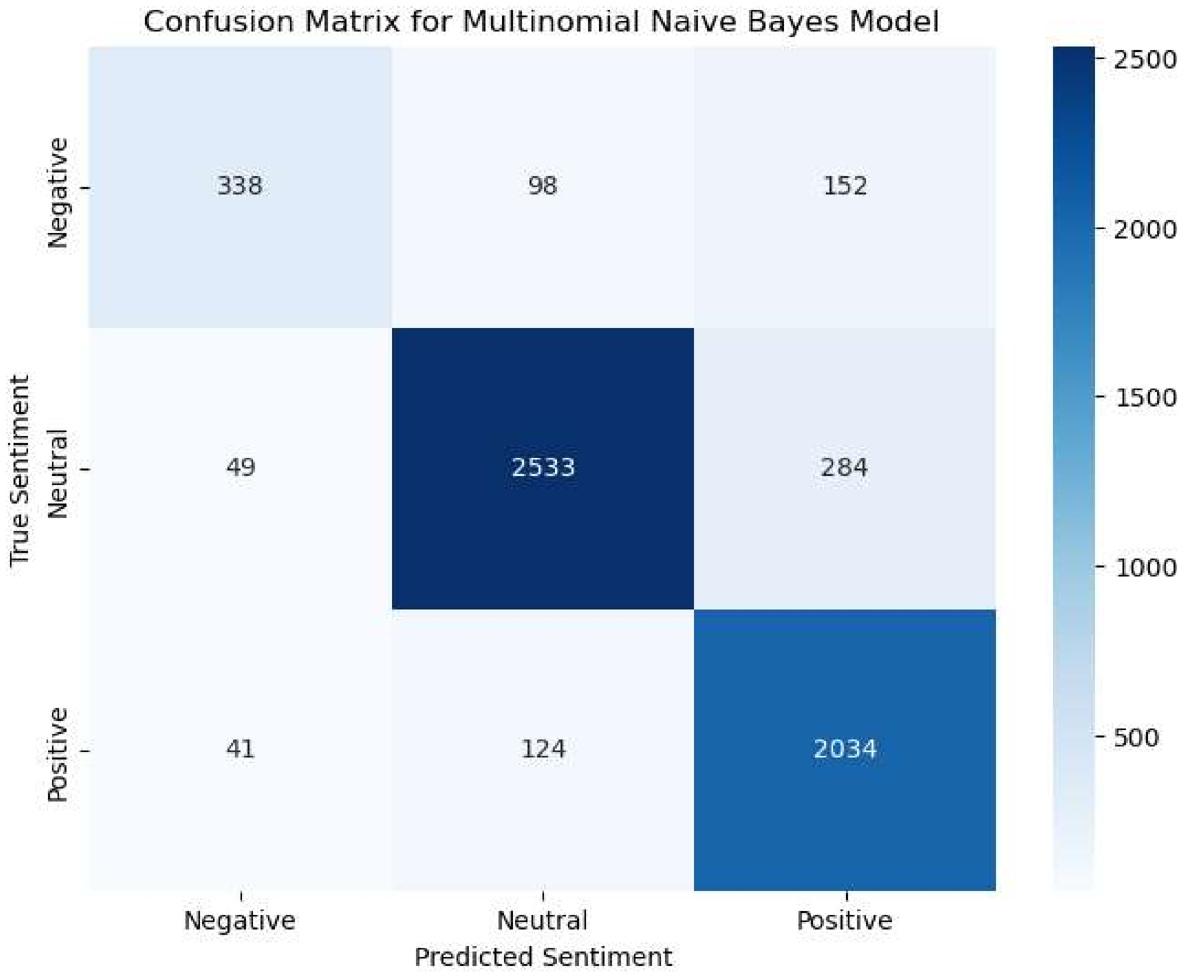




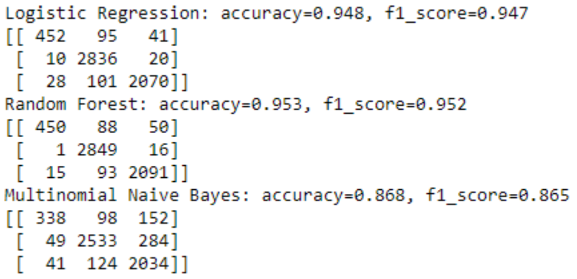
MULTINOMINAL NAÏVE BAYES MODEL

Here I performed training and testing a Multinomial Naive Bayes (NB) model on the pre-processed and vectorized text data. Multinomial NB is a probabilistic model that is commonly used for text classification tasks, and it assumes that the features (i.e., word counts) are generated from a multinomial distribution. The code first instantiates an object of the Multinomial class, fits the model on the training data using the fit () method, and then predicts the sentiment of the test data using the predict () method. The accuracy of the model is evaluated using the accuracy_score() function, and the F1 score (which is the harmonic mean of precision and recall) is calculated using the f1_score() function with the weighted average option. The confusion matrix (which is a table that shows the number of true positive, true negative, false positive, and false negative predictions) is calculated using the confusion_matrix () function.


Accuracy of Multinomial Naive Bayes Model
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Multinomial Naive Bayes



COMPARISON OF MACHINE LEARNING MODELS



```
Logistic Regression: accuracy=0.948, f1_score=0.947
[[ 452   95   41]
 [  10 2836   20]
 [  28  101 2070]]
Random Forest: accuracy=0.953, f1_score=0.952
[[ 450   88   50]
 [   1 2849   16]
 [  15   93 2091]]
Multinomial Naive Bayes: accuracy=0.868, f1_score=0.865
[[ 338   98  152]
 [  49 2533  284]
 [  41  124 2034]]
```

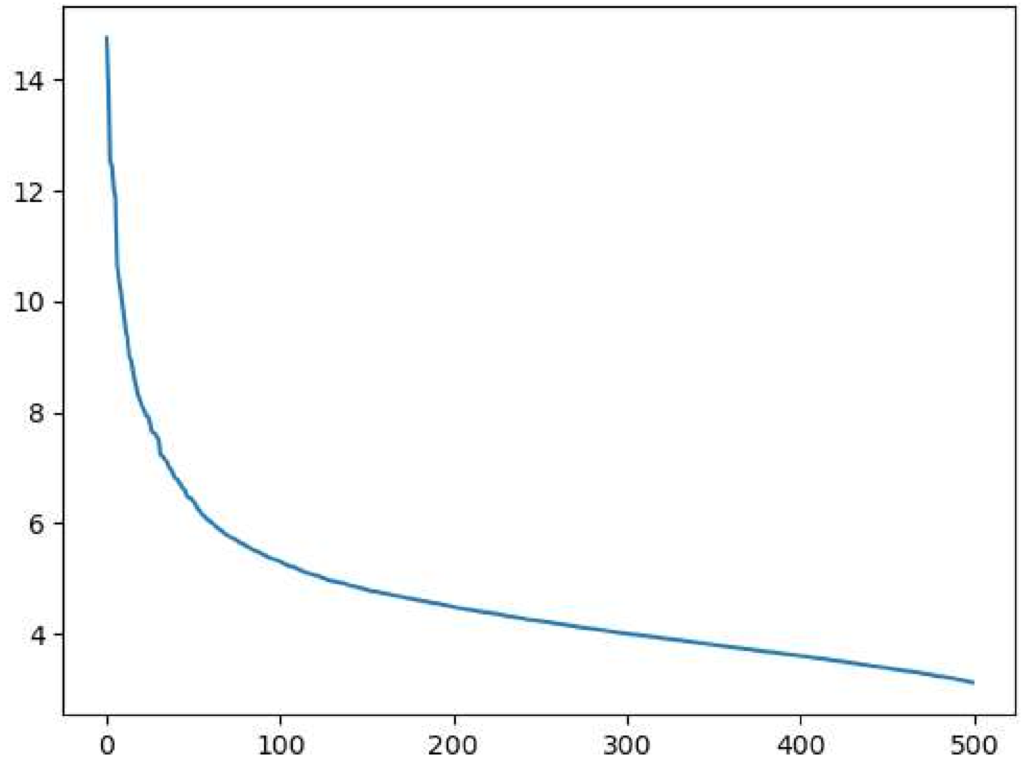
Hence, from the above result Random Forest has performed the best. According to my understanding, whenever there is complex relationship in the data and involves a lot of features, random forest usually performs the best.

K-MEANS CLUSTERING

DIMENSIONALITY REDUCTION

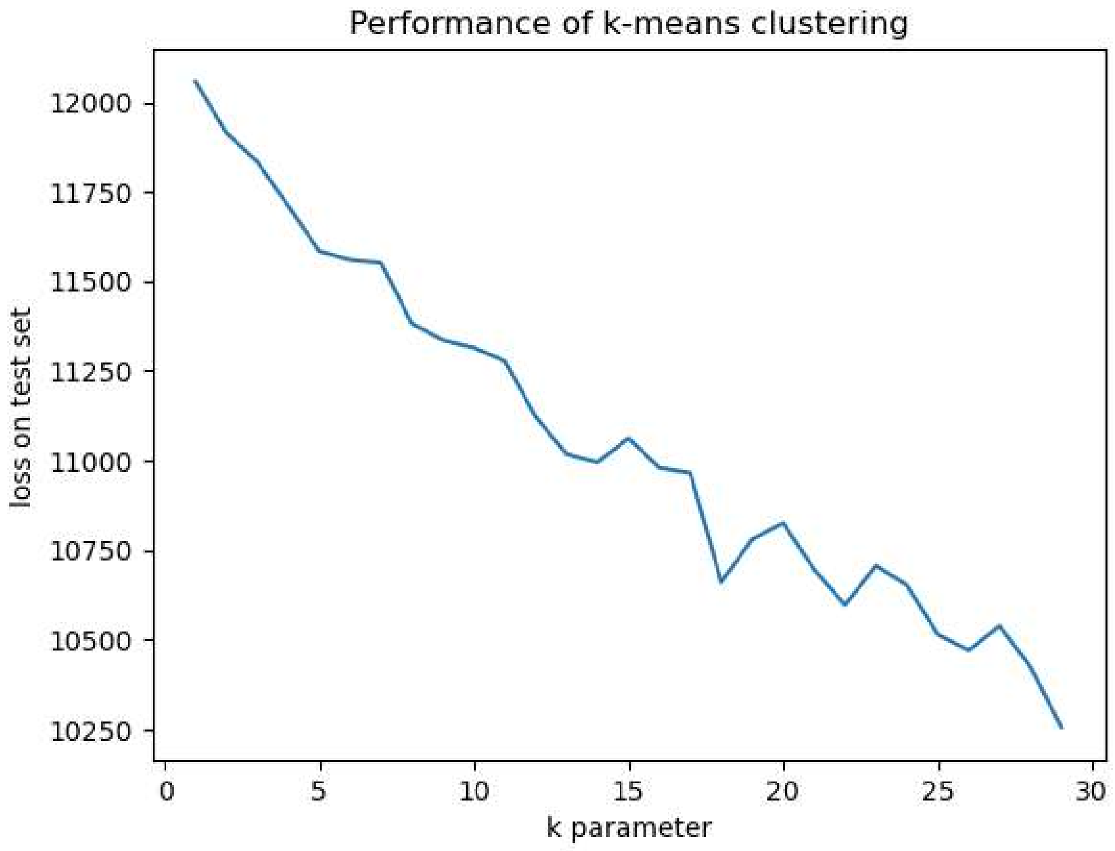
At first, I performed dimensionality reduction on the input data X using TruncatedSVD (Singular Value Decomposition) from the Scikit-learn library. The TruncatedSVD model is initialized with n_components=500, indicating that the transformed data should have 500 components. The fit_transform method is then called on the input data X, which applies dimensionality reduction and returns a new array X_pca with 500 components.

Next, the singular_values_ attribute of the pca object is plotted using lineplot. The singular values are the diagonal values of the singular value matrix obtained during the SVD decomposition, and they represent the strength of each component in explaining the variance in the data. The line plot shows how the strength of each component decreases as we move from the first component to the last component, indicating how much information is retained by each component.



K PARAMETER SELECTION USING ELBOW METHOD-

I performed k-means clustering on the reduced feature set of data X_reduced with different values of the k parameter, where k represents the number of clusters to be formed. I wrote a function that computes the sum of squared distances between the samples and the cluster centers (inertia) for each k value and appends the negative of the score to the losses list. Then, I plot the k parameter values on the x-axis and the corresponding loss values on the y-axis to visualize the performance of the k-means clustering model. My ultimate goal is to find the optimal k value that minimizes the loss and produces the best clustering result.



According to the graph above, =14 is the best.

I performed k-means clustering on a reduced dataset X_reduced with 14 clusters, maximum of 6000 iterations, and using 4 CPUs in parallel to speed up the computation.

CLUSTER ANALYSIS

Cluster analysis is a statistical technique used in data analysis to group similar objects or observations together into clusters based on their characteristics or attributes. The goal of cluster analysis is to identify groups of tweets that are more similar to each other than to those in other groups.

I performed cluster analysis using k-means clustering. K-means clustering involves partitioning the data into k clusters, where k is a user-defined parameter, and assigning each object to the cluster with the nearest mean.

Cluster analysis has numerous applications, including customer segmentation, image segmentation, gene expression analysis, and social network analysis. It can also be used as a pre-processing step for other data analysis techniques, such as classification and regression.

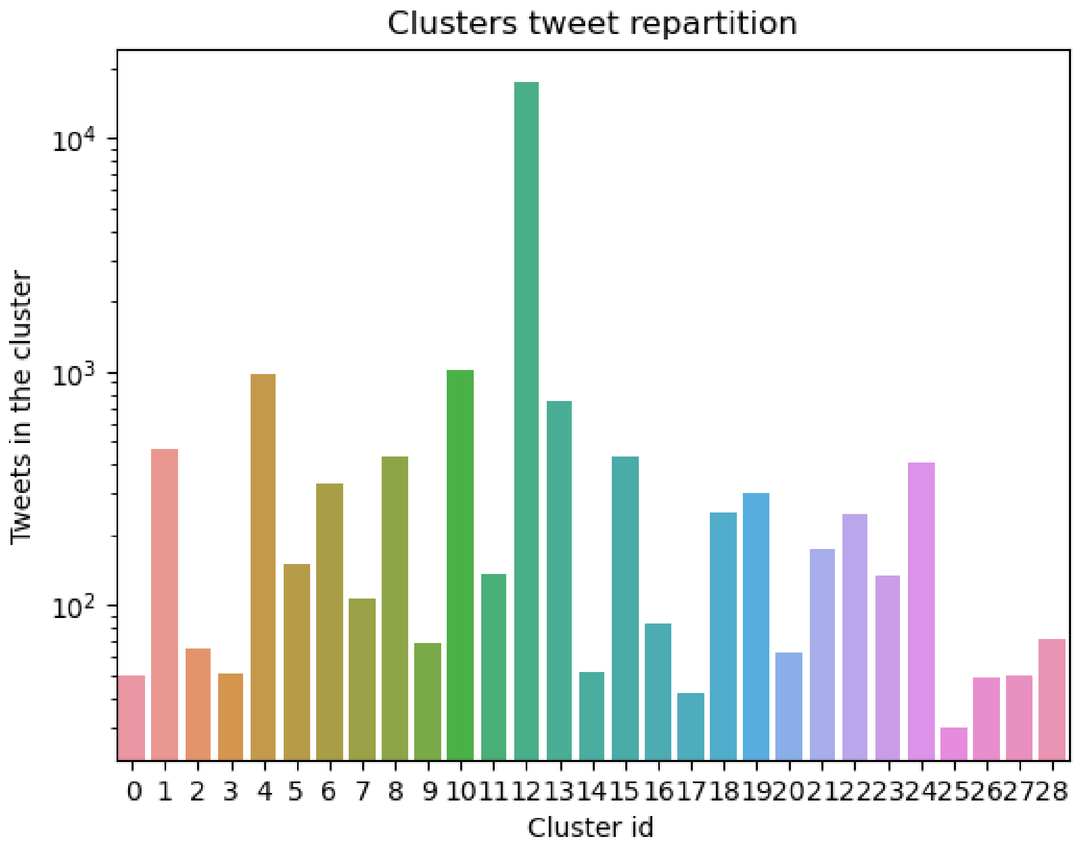
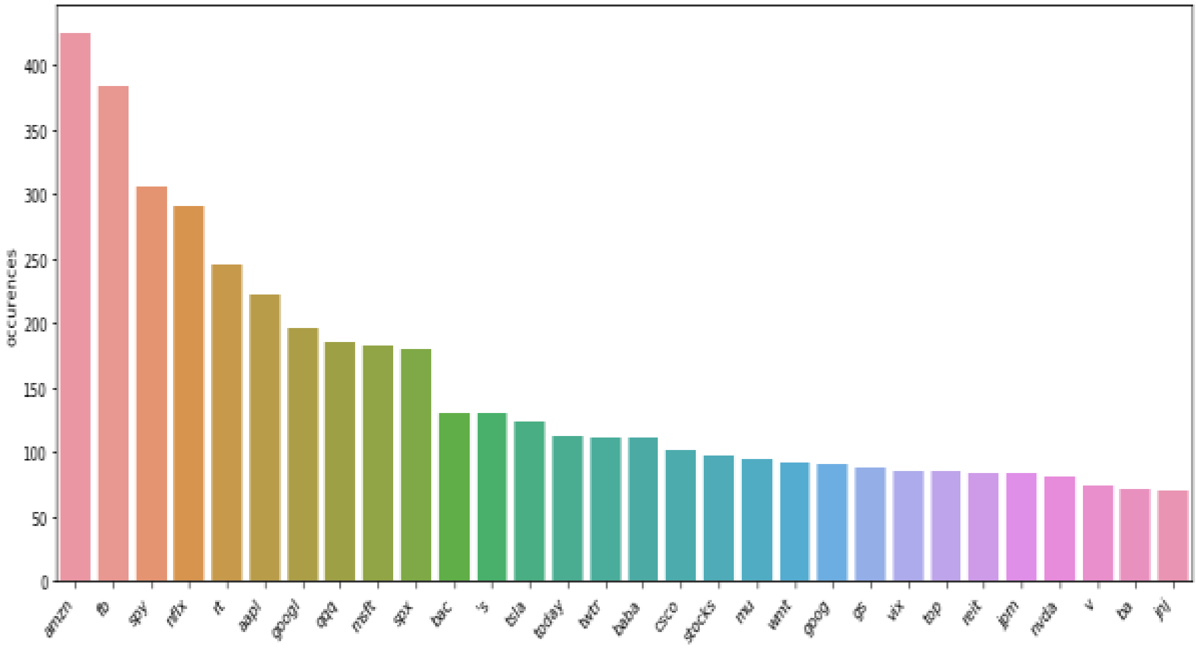


Fig: Looking at the cluster repartition, one big cluster with more than 10,000 tweets and a few smaller ones and we don't expect this large cluster to be especially meaningful, so we'll look more into the others.

Financial news about the technology market

Upon examining the most frequent tokens in cluster 7, it was found that there are many symbols or mentions of well-known American tech companies such as Amazon, Facebook, Netflix, Apple, Google, Microsoft, Tesla, Cisco, Alibaba. This indicates that the tweets in this cluster are mostly related to the technology sector and specifically to innovative technology companies. It is possible that the tweets in this cluster discuss recent developments, news, or opinions related to these companies, or their products and services. This information can provide insights into the topics and trends that are relevant to the technology sector, and can be useful for businesses, investors, or researchers who are interested in this domain.

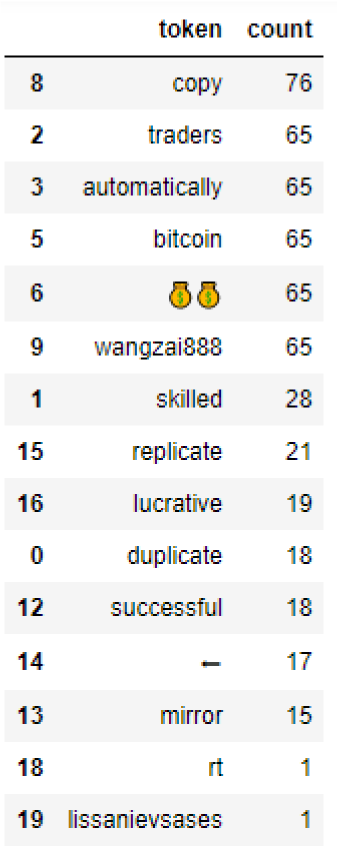
For example, if a company wanted to know what people were saying about a particular technology product or service, they could use clustering and analysis techniques like the ones used in this context to identify clusters of tweets that are most relevant to their product or service. By examining the most frequent tokens in these clusters, they could gain insights into the sentiment and opinions of Twitter users about their product or service, as well as competitors in the same market.



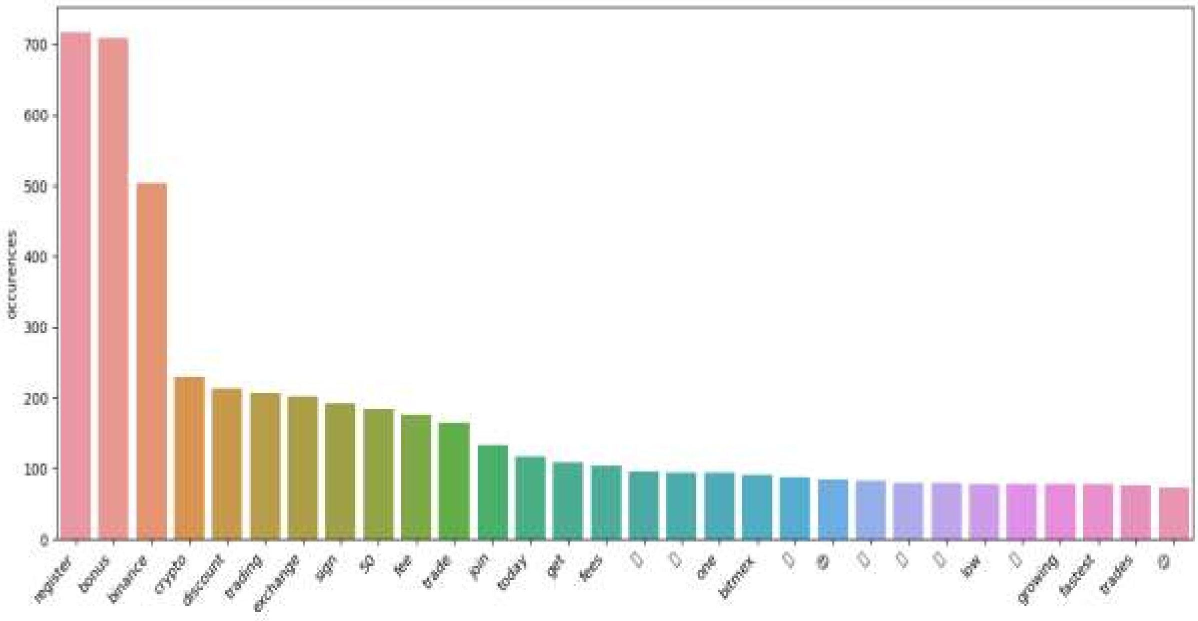
Promotional tweets & spam detection

In the next section, I analyzed the most frequent tokens in tweets that belong to cluster number 2. The tokens suggest that the tweets are related to promotional activities for cryptocurrency trading platforms, and not directly related to financial news. The presence of marketing terms, emojis implying wealth and success, and references to cryptocurrencies and trading platforms support this conclusion. I have put a figure below which shows the count of different tokens.

The statement further suggests that in certain applications, such tweets may not be desirable and can be filtered out through better data collection. This means that the model can be used to predict the cluster membership of new tweets, and those that belong to cluster 2 can be excluded from the analysis. This approach can help to ensure that the data used for financial analysis is relevant and accurate.

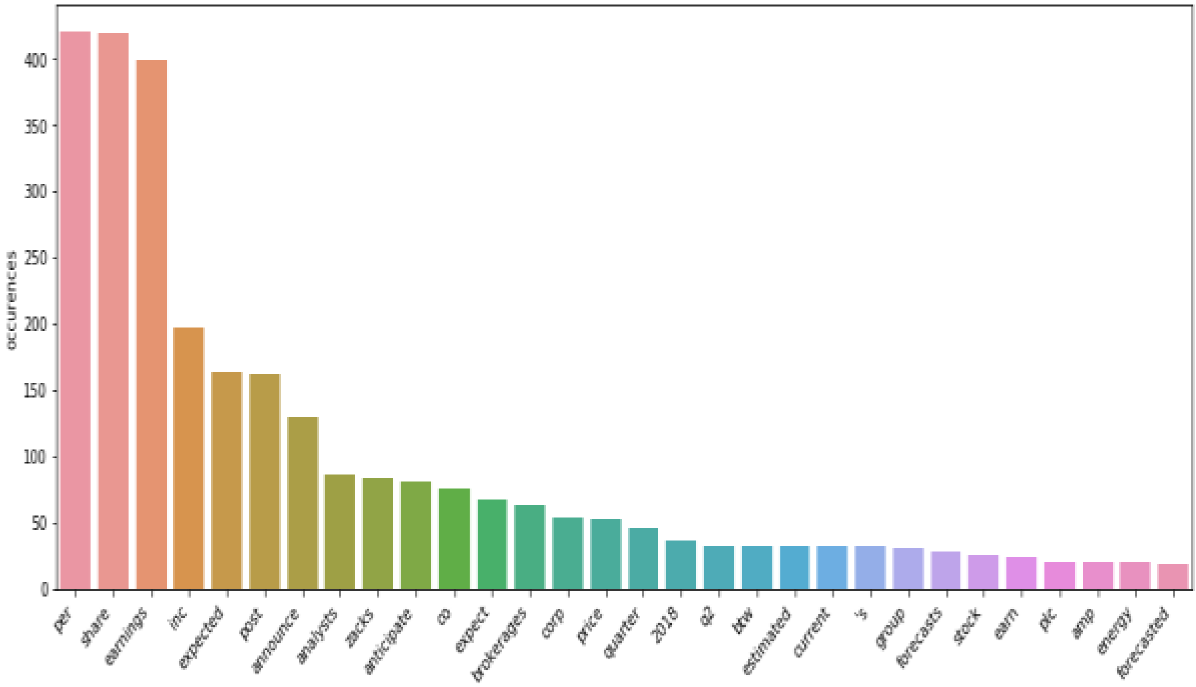


	token	count
8	copy	76
2	traders	65
3	automatically	65
5	bitcoin	65
6	💰💰	65
9	wangzai888	65
1	skilled	28
15	replicate	21
16	lucrative	19
0	duplicate	18
12	successful	18
14	➖	17
13	mirror	15
18	rt	1
19	lissanievsases	1



TWEETS ABOUT FORECASTING THE MARKET.

In the next section, Cluster 13, which has been identified based on a clustering model applied to a dataset of tweets related to stocks. The tokens that appear most frequently in this cluster include words that are related to predicting the future performance of stocks, such as "expected," "post," "announce," "anticipate," "expect," "estimated," "forecasts," and "forecasted." Tweets in this cluster are likely to contain predictions or forecasts about the future price movements of different stocks. Such tweets are commonly posted by stock market analysts and investors who are interested in predicting the future performance of stocks based on various factors, such as market trends, company news, and financial data. By identifying this cluster, analysts and investors can focus on tweets related to stock price forecasting, allowing them to stay informed about the latest predictions and forecasts in the market.



CONCLUSION-

During the data exploration, the clustering model was able to identify a cluster (Cluster 7) that contained tweets related to news in the technology sector and another cluster (Cluster 13) that contained tweets related to market predictions. This shows that the model was able to successfully group tweets based on their content and can be used to find relevant tweets about these specific topics.

However, the data collection also showed a flaw in the dataset. It was discovered that there were many tweets in a cluster (Cluster 2) that contained promotional content for cryptocurrency exchanges. This indicates that the data collection process was not perfect and allowed irrelevant tweets to be included in the dataset. This issue can be addressed by improving the data collection process and filtering out irrelevant tweets before building a model.