

Sentiment Analysis of Top Stock Market Companies

Abeer Kheder Alghamdi¹, Bushra Mohammed Alsaadi², Hessah Almugati³, Shatha Al Asem⁴, Manal Abdullah⁵
¹Information Systems Department), Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
abeeralghamdi9@gmail.com
²Bmalsaadi.00@gmail.com
³hessahalmugati@gmail.com
⁴salqahtani1409@gmail.com
⁵maaabdullah@kau.edu.sa

ARTICLE INFO

ABSTRACT

Received: 18 Oct 2024
Revised: 15 Dec 2024
Accepted: 28 Dec 2024

In the current era, data grew in volume, variety, and velocity; therefore, handling the term big data. While conventional data type consists of formatted data, big data involve structured, semi structured and unstructured data for which enhanced tool is requisite for processing the data to pull out the value-added information. The characteristics of big data are briefly discussed in this paper particularly, the 5 Vs of big data which include Volume, Velocity, Variety, Veracity and Value as features used for the selection of datasets for analysis.

A set of 3 million tweets associated with five popular international enterprises, including Amazon, Apple, Google, Microsoft and Tesla was chosen to illustrate the characteristics of big data. This data set is both numeric and non-numeric; numeric data includes only tweet ID and metadata, while the non-numeric data includes the text of the tweet from 2015 to 2020. The study used this dataset to assess its applicability to the 5Vs framework and perform the descriptive statistical analysis to determine broader trends using Python.

The result of this study shows that the sentiment of all companies mainly neutral sentiments, followed by positive sentiments; a lesser quantity of tweets expressed negative sentiments. Through machine learning approaches, the best result was achieved by the SVM model with the correct rate of sentiment prediction at 96%, which is better than the other models like LR and Decision Tree. This analysis highlights the significance of real-time sentiment monitoring and spotlighting prescriptive recommendations, such as creating content strategies that actually engage audiences by successfully driving their interactions. By implementing these strategies, enterprises can strengthen their positions in the market and control their weaknesses more readily.

Keywords: Big Data, 5Vs, Twitter Dataset, Data Analysis, Structured Data, Unstructured Data, Sentiment Analysis, Apache Spark, Big Data Analytics, Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, and Prescriptive Analytics. Social Media Analytics

INTRODUCTION

Big data is a new pattern that came with the development of the information age and the tremendous growth of data in the information technology field. Big Data, defined by volume, variety, velocity, veracity and value, is a phenomenon that is transforming industries by introducing more sophisticated means of data management and analysis [1][2]. Big Data is generally referred to as the vast data sets that cannot be handled by conventional data management methodologies; it includes the two forms of Big Data: the structured data and the unstructured data, where the latter encompasses the information found in documents and on social networks [3].

Coordination of Big Data analytics own has emerged as a strategic success factor that organizations will need to cultivate if they are to achieve competitive edge. Sophisticated data processing methods together with computational methods are used to derive meaning from these structures for forecast and decision making [4]. It will be important to include social media content from sites such as Twitter as the means by which unstructured

data can be accessed and provide an insight into the trends about and opinions of the population. Twitter data are perhaps the best type of data to illustrate the nature and the potential of Big Data analytics: the data are highly voluminous, are characterized by high velocity and are heterogeneous [5].

Past studies demonstrate how highly favorable frameworks like Hadoop and Apache Spark have to be implemented to ensure effectiveness as well as efficiency in large datasets [6], [7]. These technologies allow processing in real time and are appropriate for such fields as uses of Twitter data for sentiment analysis and trends detection [8], [9].

In this paper, Big Data and its application to the context of Twitters are examined with specific emphasis on sentiment analysis and real-time trends assessment. Employing concept of big data and data mining, the study explores the possibility of social media analytics to deliver information that is helpful in decision making. With respect to the solutions proposed for dealing with issues of data integration and scalability, and the optimization of data processing in Big Data, this work advances knowledge in the fields of Big Data and social media.

LITERATURE REVIEW

A. Background

This section gives an overview of Natural Language Processing (NLP) and Sentiment Analysis that are vital tools in processing big data particularly from facility like Twitter. These technologies assist to extract insights embedded in textual data that has not been structured systematically.

There are some differences between traditional and big data in terms of the used data and the processing tools [4]. The former uses structured data stored in tables, and it depends on statistical analysis and query language such as SQL. While the latter uses huge amounts of data in different formats such as structured, semi-structured, and unstructured data. Big data requires more advanced analytical tools such as machine learning to extract patterns. Table 1 summarizes the main differences between traditional and big data [1].

Table 1. The main differences between traditional and big data [1]

Components	Big data	Traditional Data
Queries	Largely Abandoned SQL	Traditional SQL
Architecture	Distributed	Centralized
Data Types	Structured, Semi- Structured and Unstructured	Structured
Data Model	No Schema	Fixed Schema
Data Relationship	Unknown or complex	Known Relationship
Data volume	Petabytes or Exabytes	Terabytes
Data Traffic	More	Less
Data Integrity	Less	High

B. Natural Language Processing,

AI is an important field, which has a branch of study known as Natural Language Processing or NLP. The ultimate goal of the Natural Language Processing is to let the machines useful and meaningful and translate it to Human language. As the amount of textual data obtained from sources such as social networks, customer's reviews, and others continuously grows, the role of NLP to analyze such unstructured data intensifies [1].

NLP encompasses various tasks, such as:

- Text Classification: Classification of text into sets of known categories or classes (classification, for example, to spam or to a particular topic).
- Named Entity Recognition (NER): Extracting proper nouns like people's names, place's location or even an organization.
- Part-of-Speech Tagging: Labeling each word of an utterance, according to segments of grammatical categories being nouns or verbs and so on.
- Sentiment Analysis: Identifying the directions of the text and choosing the type of their emotions as positive, negative or neutral.

This has made the NLP to become complex because the task involves analyzing, various language patterns, dialects, idioms among others and there is also so much data to process. Some of the tasks are best handled with new advancement in Natural Language Processing, in which machine learning and deep learning models work efficiently. Scholars have also used framework like Hadoop and Spar for large scale NLT to provide better solution and deal with the volume problem [2],[5].

C. Sentiment Analysis

Specific to NLP, there is a subfield referred to as Sentiment Analysis that is concerned with extraction of opinionated data from text.

This technique is particularly applicable to large datasets of user opinions which are often posted on SNS such as Twitter. This mostly involves the classification of a text as positive, negative or neutral depending on people's perception in the public or customers.

Sentiment analysis typically involves several steps:

- Text Preprocessing: Now the text having cleaned and prepared for analysis, which means all the noise like stop words, punctuation, special characters etc.
- Feature Extraction: Finding out certain words and or phrases used for polarity-based sentiment analysis.
- Model Building: Filtering for sentiment annotation and using machine learning models – for instances Naive Bayes, Support Vector Machines or deep learning models such as LSTM on tagged posts.
- Evaluation: Measuring how well labeled sentiments agree with the sentiments that a program has predicted.

Real-time sentiment analysis tasks for social media data have also seen more mature big data platforms and tools such as Hadoop and Spark. It has also found application in sentiment analysis, to enhance accuracy when tackling complicated sentiment expressions of sentiments.

Sentiment analysis has wide applications across various industries:

- ☐ Brand Monitoring: Monitoring consumer perception on the company's products or services.
- ☐ Political Analysis: Measuring people's political literacy in terms of political awareness and views about a particular event or a personality.
- ☐ Market Research: Finally, analyzing customer needs and preferences as well as social taste.

Current studies in this area are more inclined towards enhancing the practicality and reliability of sentiment analysis by using it in the large data processing environments. For example, frameworks based on the architectural design of Hadoop have been applied to process and analyze incoming streaming Twitter data for real-time sentiment analysis [14]. Furthermore, LSTM deep learning models have been applied to overcome limitations, reduce the complexity of sentiment analysis and capture context data [11].

D. Big Data Overview

Apart from the two main discussed areas, there are other important sections is needed to improve your searching results on the big data and sentiment analysis as follows. To start with, the presentation of Overview of Big Data so as to establish deeper perception of the idea of big data; its characteristics inclusive of volume, velocity, variety, veracity and value-the 5Vs and relation with sentiment analysis. This section can also cover one or some of the technologies like Hadoop, Spark or cloud platforms that play an important role in handling large datasets in historical and real time mode. This overview will provide a good basis for the following discussions on big data technologies and its application in different industries such as healthcare, finance and industries, social media [3],[5].

The characteristics of big data evolved over time. In 2001, it was defined by three Vs: Volume, Velocity, and Variety [5]. The Volume stands for the massive amounts of data generated every second. While the Velocity highlights the rapid speed at which this data flows and must be processed. Finally, the Variety stands for the diverse formats and types of data collected.

After that in 2013 [6], a new 4th element was added, Veracity. It stands for the accuracy or reliability of the data. Later in 2014 [7], some experts defined a 5th element, Value, which refers to the usefulness or value that can be gained from the data (see Figure 1).

Later, more studies were conducted on big data characteristics and some experts reached 10 Vs: Volume; Velocity; Variety; Value; Validity, which refers to data authenticity; Variability, which refers to the consistency of data over time; Visualization, which refers to the graphical representation of the data; Vulnerability, which refers to the security breach of the data; and Volatility, the rate of data changes.

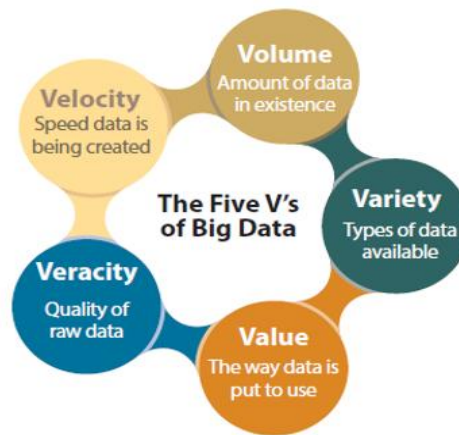


Figure 1 The Five Vs of Big Data

E. Big Data Analysis Problems

After that, the Challenges in Big Data Processing section should explain the nature, and the intrinsic problems, of handling big data sets. This also comprises data quality, which is dirty data commonly seen in the social media stream, what kind of and how many data can be processed and integrated from multiple heterogeneous sources. Data privacy and security are also another major issues, especially where personal data involves such as that gotten from Twitter and Facebook, where users' anonymity is of utmost importance. Overcoming these obstacles will provide insight into the misconceptions with regards to sentiment analysis on big data[7],[14].

F. Sentiment Analysis Techniques And Models

After that, the part devoted to Sentiment Analysis Techniques and Models would elaborate on particular techniques for sentiment analysis within textual data. It is still possible to build the traditional models of the machine learning like Naive Bayes, SVM or decision trees, but new complex solution like RNNs or LSTM networks are beginning to establish themselves for sentiment classification.

A detailed discussion of these methods will help highlight the evolution of sentiment analysis techniques and how they have adapted to the complexity of big data[11],[16].

G. Real-Time Sentiment Analysis Applications

RTSA Applications should discuss how sentiment analysis has been implemented for situations as they persist in the current circumstances. This can include tracking tweets and posts for brand monitoring, testing customers satisfaction feedback, to even rating the political sentiment during elections. Exemplifying these applications will establish that sentiment analysis is convenient in situations where big data analytics results are necessary in real time as it is common in fast moving markets [17][18].

H. Big Data Analysis in Social Networking Sites

More specifically, the opportunities and challenges of using social media as a source of Big Data Analytics about sentiments will be covered in the section called Big Data Analytics in Social Media. For example, Twitter offers highly informative content stream of users' activity; however, the data are usually load and extremely unstructured which call for elaborate pre-processing. This section may also examine how tools such as Hadoop, and the Spark framework can be utilized to analyze piles of social media information. Since present day information processing

relies heavily on avenues such as Twitter and Facebook, the role of these platforms for analyzing big data has to be studied in order for a work on sentiment analysis [9],[13].

I. Big Data Analytics Tools and Framework

In addition, it would be relevant to briefly go over the Big Data Analytics Tools and Frameworks that are generally employed to deploy sentiment analysis models. Due to their solid ecosystems, Hadoop and Apache Spark are used to manage big data processing tasks. Frontend technologies including but not limited to, Angular.js, React, Vue, Python frameworks like scikit-learn, TensorFlow, PyTorch, and Tableau and Power BI can help in the visualization of outcome. This section would have given the readers a perspective of the technical requirements needed in performing sentiment analysis at a large scale[14],[16].

J. Evaluating sentiment Analysis Models

Another section that would be useful for the topic would be the Evaluation of Sentiment Analysis Models, where you may describe how different metrics like accuracy, precision, the recall, and F1-score reflect the performance of sentiment analysis models. It will also be beneficial to learn model selection procedures for the improvement of these models including parameter search, feature extraction and validation. It is also possible to learn which methods better fit different sentiment analysis tasks by comparing classic machine learning algorithms with more complex deep learning techniques [11],[12].

K. Future Trends In Big Data and Sentiment Analysis

Future advancement in the Big Data and Sentiment Analysis would be further discussed in a section named Future Trends in Big Data and Sentiment Analysis. It can be discussed with examples including the combination of sentiment analysis with other AI fields like reinforcement learning or computer vision that can expand the potential of analyzing the users' sentiments in multimedia content. The issue of real time sentiment analysis also has a very bright future and it is especially important as more and more businesses and governments consider social media data for real time public opinion. Professional ethical considerations shall remain valid in sentiment analysis particularly in light of bias, privacy and misuse of information from social media content [1],[5].

L. Related work

This section presents and compares five relevant research studies on the use of sentiment analysis to tweets on leading firms. The study entitled "Speculator and Influencer Evaluation in Stock Market by Using Social Media" employed the identical dataset utilized in this study to evaluate speculators and influencers in the stock market by analyzing social media data and sentiment, utilizing the McDonald and Loughran sentiment analysis dictionary. Their findings highlight that there is no correlation between the number of tweets and the trading volume of companies [8]. Another related study [19] introduces a model that does sentiment analysis on tweets, employing a mix of supervised and unsupervised machine learning methods. The researchers developed several machine learning models to categorize tweets as reflecting positive, negative, or neutral sentiment. The performance of their model was assessed using cross-validation and F-score, showing its efficacy in directly mining text data from Twitter. Similarly, [20] applied sentiment analysis with machine learning and natural language processing to comprehend individuals' emotions and perspectives on social media. Their methodology involved preprocessing text data (converting to lowercase, removing stopwords, punctuation, URLs, etc.), tokenization and stemming, TF-IDF vectorization, calculating sentiment scores using R packages, and employing logistic regression, SVM, and Bernoulli Naive Bayes models for sentiment prediction. Whereas Tushar Rao indicated that Twitter sentiment analysis could be utilized to forecast changes in the stock market. The research employs Granger's Causality Analysis to investigate the correlation between Twitter conversations and stock price fluctuations, alongside the development and assessment of the "Expert Model Mining System" (EMMS) for predicting stock returns, utilizing R-squared and Maximum Absolute Percentage Error (MaxAPE) as performance indicators. The main finding indicates a strong correlation (up to 0.88) between stock prices and Twitter sentiment [21]. The last paper "Influence of Social Media over the Stock Market" employs a logit model to examine the impact of social media sentiment on stock markets. Furthermore, a fuzzy-set qualitative comparative analysis (fsQCA) will be employed to investigate the impact of investors' profiles on the association between social media and stock market risk [22]. This study will apply Spark for sentiment analysis, a method not employed in previous research, to evaluate individuals'

feelings and views about major companies. Assessing whether they are positive, negative, or neutral and the potential implications for their reputation and stock market performance.

METHODOLOGY

In this section, we present a systematic process for data analysis and decision-making (see figure 2).

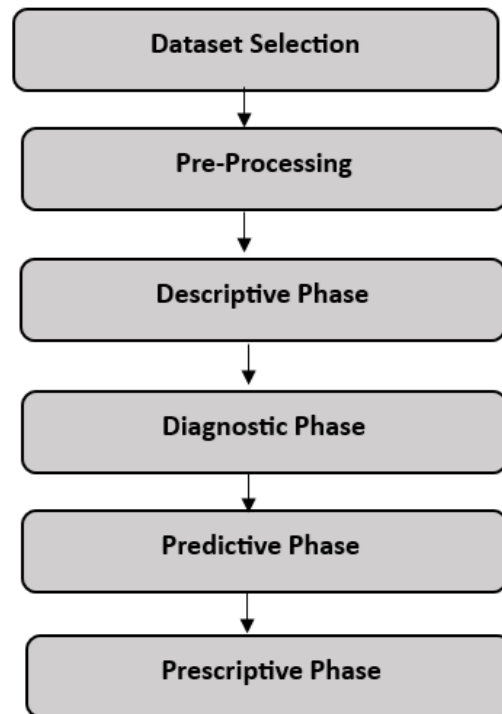


Figure 2 Big Dataset Methodology

M. Dataset Selection

The selected big dataset is real- data that is collected from Twitter about top global companies in stock markets. It covered the top 5 companies: Amazon, Apple, Google, Microsoft, and Tesla. It includes people's reviews about these companies, and it consists of over 3 million tweets for five years from 2015 to 2020.

The data was published in a research paper at an IEEE conference paper [1] and also shared via Kaggle [2]. The paper is entitled "Speculator and Influencer Evaluation in Stock Market by Using Social Media" and written by Mustafa Doğan; Ömer Metin; Elif Tek; Semih Yumuşak; Kasım Öztoprak. They used this dataset to conduct sentiment analysis about these companies using machine learning techniques.

It consists of two tables. Table 1 has 7 columns representing structured and unstructured data: tweet, tweet ID, tweet author, postdate, number of comments, likes, and retweets. The tweet is textual data, which is considered as unstructured data, and the rest are structured data. Table 2 contains the Tweet Id and the company that tweets talked about.

The table below gives an overview of the selected dataset.

Table 2 General Information of The Selected Dataset

Metadata	Description
Source	[1]
Size By Number of Rows	3326194
Size By File	772.34 MB

Columns	7 Columns: tweet, tweet ID, tweet author, postdate, number of comments, likes, and retweets
Structured Data	tweet ID, tweet author, postdate, number of comments, likes, and retweets
Unstructured Data	Tweets
Covered Companies	Amazon, Apple, Google, Microsoft, and Tesla
Period	5 years (2015 – 2020)

The Table below shows the table columns and their metadata.

Table 3 Columns Metadata

Column	Description	Type
Table 1		
Body	It contains the tweet text.	Object
Tweet ID	It contains the ID of the tweet that is assigned by the Twitter app.	int64
Writer	It contains the username of the person who wrote the tweet	Object
Post date	It contains the date on which the tweet was posted	int64
Comments Number	It contains the number of comments on each tweet	float64
Retweets Number	It contains the number of retweets of each tweet	float64
Likes Number	It contains the number of likes of each tweet	float64
Table 2		
Tweet ID	It contains the ID of the tweet that is assigned by the Twitter app.	int64
Company	It contains the name of the company the tweet is about	Object

N. Data Pre-Processing

Before starting the analysis processes, preprocessing techniques should be applied to prepare the data. We used different NLP techniques for pre-processing [28] as follows:

- **Tokenization:** splitting the sentence into a set of individual words called tokens, to process each word separately.
- **Stemming:** This process involves converting the words into their root. The aim is to group all the words with the same meaning together (e.g., took, taken will be converted to take).
- **Removing URL:** we removed all the URLs in the tweets, as some are related to advertisements and do not provide useful information.

-
- **Stop Words Removing:** remove common words that do not provide meaningful information such as 'the', 'and', 'is', 'are', and so on.
 - **Removing Special Characters:** all special characters including punctuations are removed, as they will not contribute to sentiment prediction.
 - **Dataset Sampling:** Due to the limitations of the computers used in handling big data in the analysis process, we randomly select a sample of 50K tweets, 10K for each company.

O. Descriptive Phase

The first level of Big Data Analytics is descriptive analytics and is used to extract data of past events to make future trends. Its primary focus is on the 'who, what when, where and why' model of enquiry as seeks to explain the 'what' of the occurrence. Organizations use big data to make strategic decisions based on large volume data to obtain an insight of the past performance of the firm. Techniques for performing descriptive analytics include data collection and the use of charts, graphs, dashboards and mean, total values, etc [23]. We used Python to perform this analysis, and Google Collab, leveraging libraries such as Pandas for data manipulation and Matplotlib for visualization.

P. Diagnostic Phase

Diagnostic analytics builds on discovery to encompass why certain things have happened in the course of time. This phase seeks to explain why particular events occurred or decisions were made during its period. While it may simply look for relationships, and correlations as well as try to determine factors giving rise to observed patterns, diagnostic analytics provides a broader endeavor. The approach used include data mining, correlation analysis and trend analysis for understanding how changes in one variable affects the results. Descriptive analytics just shows trends while Diagnostic analytics provides motives and realities that involve decision making by which patterns of future performances are diverged; thereby making prevention of undesired result formations a key objective of the diagnosis [24]. We used Apache Spark to determine the sentiment scores of a particular company during a specific year (drill-down). In addition to aggregate sentiment data by company (roll-up).

Q. Predictive Phase

predictive analytics is all about creating future trends from past performance and trends. While the information mining is geared at solutions to the question of 'What might occur?' Statistical models, a variety of machine learning algorithms, and other forms of data analysis can help organizations identify patterns, actions, and occurrences likely to happen in the future. For instance, using PA for customer churn it is possible to predict clients' behavior and when it comes to inventory, it is possible to know volumes that are required in the future or outing financial performance of the business given past data. The most frequently used methods include regression analysis, classification approach and time series analysis. This phase enables business to take early decisions, from potential course of action which may be an opportunity or threat [25].

R. Prescriptive Phase

Prescriptive analytics is the last stage in Big Data Analytics and deals with suggesting the right strategies in order to gain the planned goals. Descriptive and predictive analytics differ from prescriptive analytics where the latter is more about providing the solution to a given problem which can be embodied by the query 'what should be done on it'. This phase of planning involves suggesting the right strategies likely to be useful in dealing with a particular problem or exploiting an opportunity. That is even beyond picking out the probable occurrences in future than providing the strategies that would lead to the most desirable results. For instance, in supply chain operations, prescriptive analytics may suggest the best routes to use for product delivery or the right amount of stock to hold at a given time given real time data. In marketing, it could mean giving specific product suggestions for the customer, or varying the price depending on the time of day. Prescriptive analytics, therefore, unites data, analytics, and decision-making tools so that business can make the right strategic decisions and enhance operations' efficiency, profitability, and performance [23]. In this phase, we recommended actions or decisions based on the predictions to achieve desired outcomes.

S. Used Tools (Apache Spark and Others)

- **Apache Spark:** A distributional data processing tool that embeds in-memory computation for enhanced speed. Utilized for relating large datasets such as the Twitter actual time stream of information. Apache Spark is an open source distributed computing system developed mainly for real time large scale data processing and analytics. Perhaps one of the most distinctive of them is its ability to work with in-memory, thus freeing the data from having to be written to disc and improving processing speeds. This makes Spark particularly appropriate for works with big data and complicated data manipulation techniques. It is used across a plethora of use cases from simple data analytics to sophisticated machine learning and data marts, for purposes as simple as batch processing, data querying, and real-time stream processing. Spark is helpful when there is needed great data processing, such as, for example, the analysis of a large stream of tweets coming from Twitter. For example, information streams of millions of tweets per minute on Twitter can be consumed and analyzed by Spark in terms of emergence of trends, patterns or anomalies [24].
- **Hadoop:** An architecture for large scale sporadic data processing and storage. Appropriate for occasions that involve small batch processing of large datasets. Proving very helpful when working with the 3-million-tweets set for historic analysis [26].
- **Python:** The major involvement in the analysis taking advantage of functions in libraries such as:
 - **Pandas:** For data cleaning data transformation, and data massaging.
 - **Matplotlib:** For forming selective object visualizations, such as bar diagrams of engagement and frequency distribution.
 - **Google Colab:** With cloud environment provided for the dataset.

DATA ANALYSIS AND RESULT

T. Descriptive Analysis

The descriptive analysis provides an overview of the dataset by summarizing its main features, highlighting patterns, and identifying key statistical metrics. We investigated the dataset by conducting a descriptive analysis. We used Python to perform this analysis, and Google Collab, leveraging libraries such as Pandas for data manipulation and Matplotlib for visualization. Below is a detailed breakdown of the analysis process:

1) Data Exploration:

The dataset includes more than 3 million tweets focused on five main companies: Amazon, Apple, Google, Microsoft, and Tesla for the period between the years 2015 and 2020. These tweets specify structured data (tweet ID, author, postdate, likes, retweets, and comments) and unstructured data (tweet text).

In this step, we will investigate and explore the data set to show the main characteristics and structure.

The Table below shows the statistical analysis of the number of likes, comments, and retweets.

Table 4 Statistical analysis of the dataset

Statistic	Likes	Retweets	Comments
Min	0	0	0
Max	654	974	153
Mean	0.4	0.4	0.1
Total	191390	170528	48872

Figure 3 shows the general information of the table including the number of rows and columns, and the data type of each column.

```
df.info();

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1021454 entries, 0 to 1021453
Data columns (total 7 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   tweet_id        1021454 non-null  int64
1   writer          1012396 non-null  object
2   post_date       1021454 non-null  int64
3   body            1021454 non-null  object
4   comment_num     1021453 non-null  float64
5   retweet_num     1021453 non-null  float64
6   like_num        1021453 non-null  float64
dtypes: float64(3), int64(2), object(2)
memory usage: 54.6+ MB
```

Figure 3 Dataset General Information

Figure 4 shows a sample of the table 1, by showing the first five rows.

```
df.head()
```

	tweet_id	writer	post_date	body	comment_num	retweet_num	like_num
0	55044150917543456	VisualStockRSRC	1420070457	b21 made \$10,000 on \$AAPL - Check it out! htt...	0.0	0.0	1.0
1	550441672312512512	KeralaGuy77	1420070496	Insanity of today weirdo massive selling. Saap...	0.0	0.0	0.0
2	550441732014223360	DozenStocks	1420070510	\$BP100 \$Stocks Performance \$HD \$LOW \$SBUX \$TGT...	0.0	0.0	0.0
3	550442977802207232	ShowDreamCar	1420070807	\$GM \$TSLA: Volkswagen Pushes 2014 Record Recal...	0.0	0.0	1.0
4	550443807834402816	i_Know_First	1420071005	Swing Trading: Up To 8.91% Return In 14 Days h...	0.0	0.0	1.0

Figure 4 First five rows of the dataset

Figure 5 shows a sample of table 1, by showing the last five rows.

```
df.tail()
```

	tweet_id	writer	post_date	body	comment_num	retweet_num	like_num
1021449	73643130341924452	it_cmsuting	1462963368	#Google to ban payday lending ads: calling ind...	0.0	0.0	0.0
1021450	736431391262648801	CredBakLA	1462963389	Maybe \$AMZN should buy #STAPLES @halloweenRapor...	0.0	0.0	0.0
1021451	736431447434548792	TradingGuru	1462963402	RT \$TSLA HFT Algor triggered SELL in \$GMA-X...	0.0	0.0	0.0
1021452	73643145672504054	it_cmsuting	1462963405	#Google Translate now works inside any app on ...	0.0	0.0	0.0
1021453	736431458746054096	computer_huare	1462963405	#Apple suppliers are seeing strange things hap...	NaN	NaN	NaN

Figure 5 Last five rows of the dataset

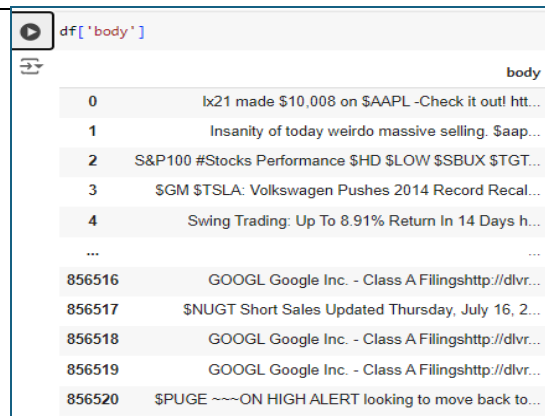
Figure 6 shows a sample of the table 2, by showing the first five rows.

```
df.head()
```

	tweet_id	ticker_symbol
0	550803612197457920	AAPL
1	550803610825928706	AAPL
2	550803225113157632	AAPL
3	550802957370159104	AAPL
4	550802855129382912	AAPL

Figure 6 First five rows of table 2

Figure 7 shows a sample of the unstructured data (tweet text), and figure 8 shows the structured data.



df['body']

	body
0	lx21 made \$10,008 on \$AAPL -Check it out! htt...
1	Insanity of today weirdo massive selling. \$aap...
2	S&P100 #Stocks Performance \$HD \$LOW \$SBUX \$TGT...
3	\$GM \$TSLA: Volkswagen Pushes 2014 Record Recal...
4	Swing Trading: Up To 8.91% Return In 14 Days h...
...	...
856516	GOOGL Google Inc. - Class A Filingshttp://dlvr...
856517	\$NUGT Short Sales Updated Thursday, July 16, 2...
856518	GOOGL Google Inc. - Class A Filingshttp://dlvr...
856519	GOOGL Google Inc. - Class A Filingshttp://dlvr...
856520	\$PUGE ~~~ON HIGH ALERT looking to move back to...

Figure 7 Sample of Unstructured Data

	tweet_id	writer	post_date
0	550441509175443456	VisualStockRSRC	1420070457
1	550441672312512512	KeralaGuy77	1420070496
2	550441732014223360	DozenStocks	1420070510
3	550442977802207232	ShowDreamCar	1420070807
4	550443807834402816	i_Know_First	1420071005

(a)

	comment_num	retweet_num	like_num
	0.0	0.0	1.0
	0.0	0.0	0.0
	0.0	0.0	0.0
	0.0	0.0	1.0
	0.0	0.0	1.0

(b)

Figure 8 Sample of Structured Data

Figure 9 shows the columns' names.



df.columns

```
Index(['tweet_id', 'writer', 'post_date', 'body', 'comment_num', 'retweet_num', 'like_num'], dtype='object')
```

Figure 9 List of Columns Names

2) Statistical Summary:

The statistical analysis focused on the number of likes, retweets, and comments:

- The minimum value of the metrics was 0 in all cases.
- Maximum Values: Reached the highest of 654 for likes, 974 for retweets, and 153 for comments.
- The mean values of likes and retweets averaged at 0.4 and 0.1 for comments, respectively.
- Total Engagement: In the dataset, the comments and likes have been registered to be 191,390 and 48,872, whereas the retweets recorded are 170,528.

3) Visual Representation:

Graphical plots were produced to show:

- The frequency of tweets in the case of each company.
- Engagement metrics (likes, retweets, and comments) of the top authors.

4) Data Insights:

- Authors: The dataset contains tweets from more than one million distinct authors, with the top five contributors making up a sizeable share of the total number of tweets. Figure 10 shows the total number of unique tweet authors.

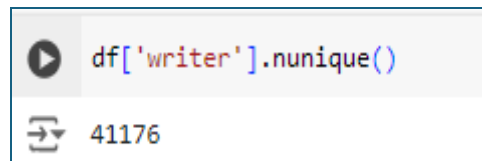


Figure 10 Number of unique authors

Figure 11 shows the top 5 authors by number of Tweets.

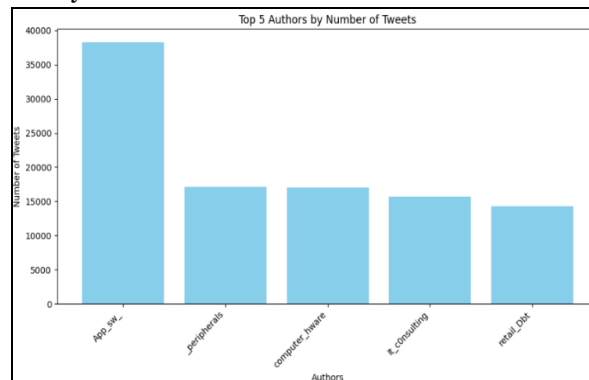


Figure 11 Top 5 authors by number of Tweets

Figure 12 shows the top 5 authors by number of Likes.

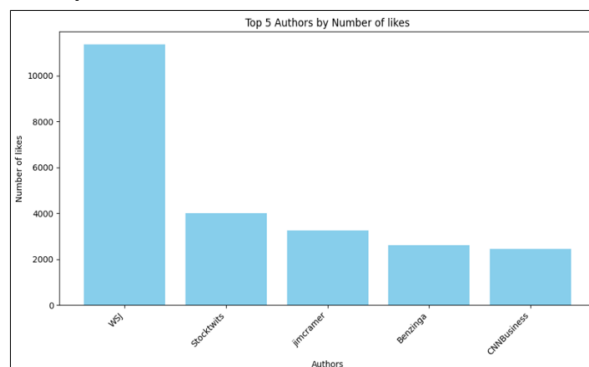


Figure 12 Top 5 authors by number of Likes

Figure 13 shows the top 5 authors by number of Comments.

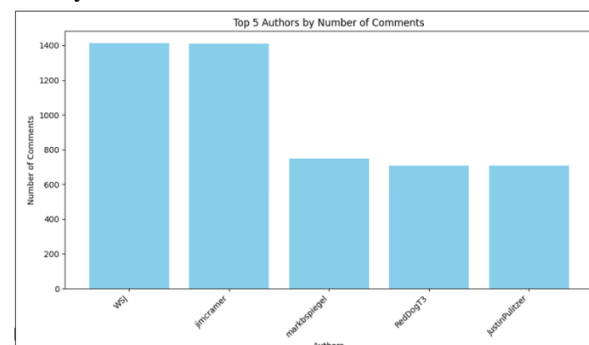


Figure 13 Top 5 authors by number of Comments

- **Company Distribution:** The dataset is evenly divided among the five companies, with each company having a representation of hundreds of thousands of tweets. Figure 14 shows the frequency of the companies by number of tweets.

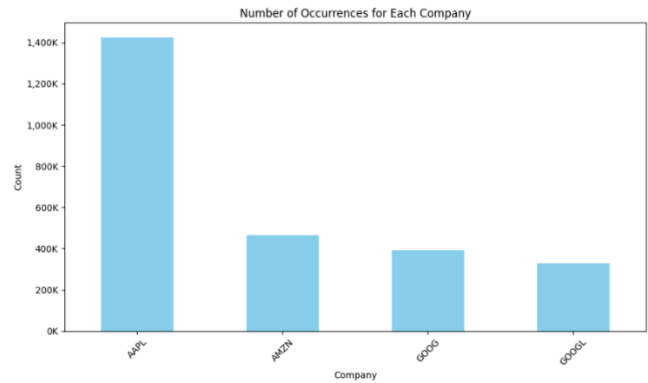


Figure 14 Frequency of the companies by number of tweets

This descriptive analysis provides insights into the data of the volume, variety, and engagement patterns of the dataset, thus making the ground for the next stage of diagnostic, predictive, and prescriptive analytics.

U. *Diagnostic Analysis*

During the diagnostic analysis phase, sentiment analysis of tweets regarding the big companies will be conducted by using Apache Spark and TextBlob. TextBlob is an open-source library in Python designed for the processing of textual data. It offers an intuitive API for executing several natural language processing (NLP) activities, including Sentiment Analysis to evaluate the emotional tone of text and categorizing it as positive, negative, or neutral [27]. First, the drill-down analysis was carried out to determine the sentiment scores of each firm, as seen in Figures 15-19. It demonstrates consistent patterns across all datasets. The majority frequency was for Neutral sentiment, indicating that most tweets lack an emotional tone. Nevertheless, the second frequency pertains to the Positive tweets. These tweets most likely convey confidence over firms' market news or performance. The least represented sentiment is Negative, which underscores worries over firms' products, stock markets, or overall dissatisfaction.

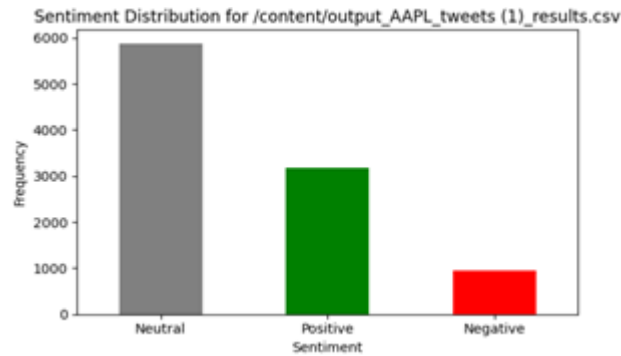


Figure 15 Sentiment Analysis for Apple

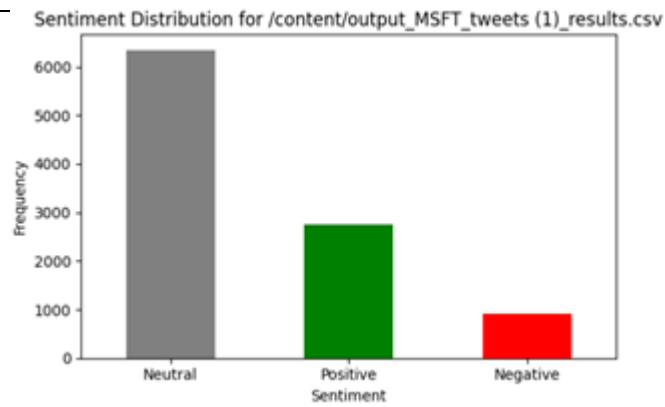


Figure 16 Sentiment Analysis for Microsoft

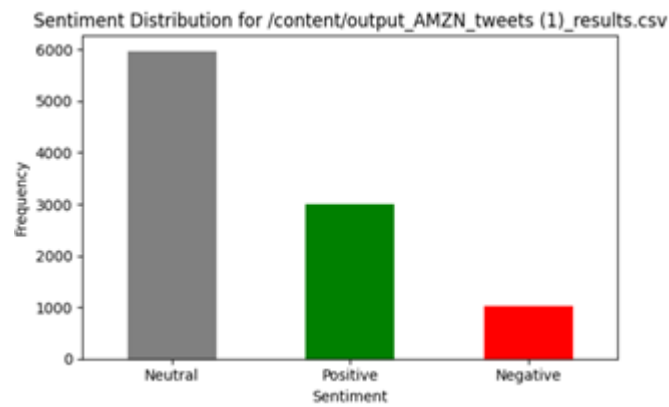


Figure 17 Sentiment Analysis for Amazon

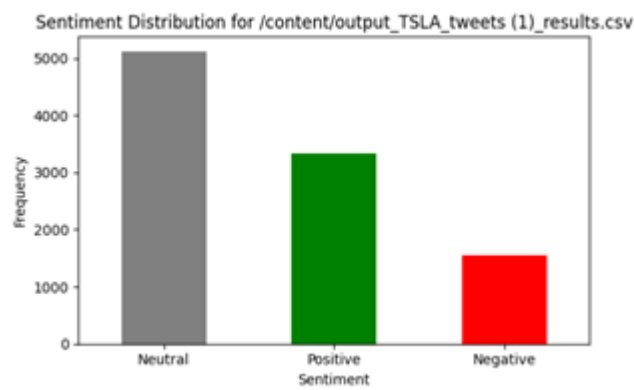


Figure 18 Sentiment Analysis for Tesla

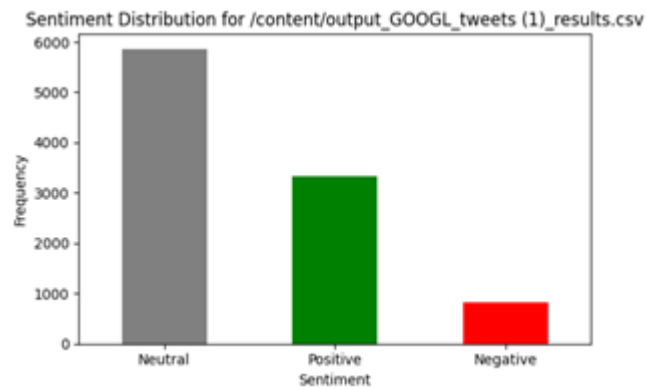


Figure 19 Sentiment Analysis for Google

Second, the roll-up analysis for all big companies will be

performed to aggregate sentiment data. Figure 20 depicts the sentiment distribution of tweets from the merged_tweets.csv dataset, classified into Neutral, Positive, and Negative opinions. The main frequency of tweets is Neutral, with an occurrence of around 30,000. Positive tweets rank as the second most frequency, totaling roughly 15,000. Negative tweets have the lowest frequency, totaling less than 10,000. The distribution indicates that the majority of provides insight in the sample are neutral or balanced, with positive sentiment far exceeding negative sentiment. This study may assist companies or stakeholders in comprehending public perception, monitoring audience sentiment patterns, or highlighting areas for enhancement.

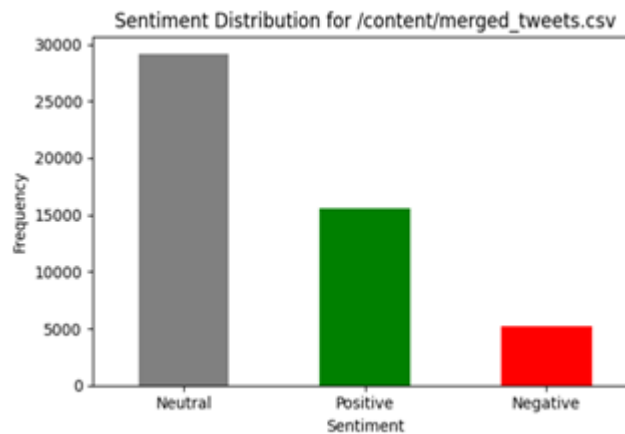


Figure 20 Sentiment Analysis for All Big Companies

V. Predictive Analysis

In this phase, we developed a machine-learning model that can predict the sentiment of new data. It is based on supervised learning, which will be learned from previously labeled data [29]. It is also based on a multi-class classification problem, which means the model will classify the text into one of three classes Positive, Neutral, and Negative.

To decide which ML algorithm to use, we select 3 popular algorithms and compare them: Logistic Regression (LR), Support Vector Machine (SVM), and Decision Tree (DT). Here are the definitions of these models [30]: LR model depends on logistic function find and represents the relationships between the features; the SVM model runs in n-dimensional space aiming to best split the data into related groups using hyperplane; and the DT model makes decisions by representing the data in a tree-like structure where the full dataset is in the root node, the attributes tests are in the middle nodes, and the final predication is in the leaf node.

To train and test the model, the data is split according to an 80-20 split. This means 20% of the data will be treated as a testing dataset, and 80% as a training dataset. The model was evaluated using 4 metrics: Precision, which measures from the samples that are predicted as positive, how many of them are really positive; Recall, from all data samples that are truly positive, how many of them are predicted as true? F1 score, measuring the trade-off between Perception and Recall. Table X shows the comparison of machine learning models results.

Table 1. Summarization of Machine Learning Models

Model	Precision	Recall	F1-score
Logistic Regression (LR)	0.95	0.89	0.92
Support Vector Machine (SVM)	0.96	0.95	0.96
Decision Tree (DT)	0.95	0.94	0.94

As shown in the table, the SVM model got the highest accuracy rate at 96% in predicting the sentiment analysis. Followed by DT (94%), and LR (92%). Figure X shows the Confusion Matrix for SVM model

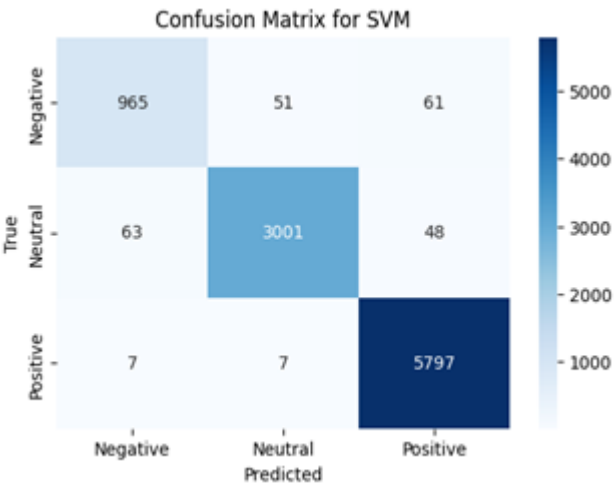


Figure 21 Confusion Matrix for SVM

W. Prescriptive Analysis

The prescriptive phase will give useful suggestions derived from insights gained throughout the descriptive, diagnostic, and predictive analytical phases. Provided below are the actionable recommendations:

- 1. Sentiment-Driven Approach
 - Employ predictive sentiment analysis with 96% accuracy for SVM to generate tailored messages for each company's audience, effectively addressing both positive and negative trends.
 - Implement real-time sentiment analysis to respond to any crisis or change in public opinion.
- 2. Enhanced Analytics Integration
 - Incorporate the SVM model into a continuous monitoring framework for optimal insights.
 - Utilizing NLP technology to enhance the analysis of unstructured data, such as Twitter text, can provide a superior understanding of user comments.
- 3. Content Strategy Informed by Data
 - Examine prevailing motifs in popular tweets for use in forthcoming content development.
 - Tailor content to each firm, mirroring the distinct interests and sentiment patterns of their audience.
- 4. Opportunities for Collaboration
 - Identify and evaluate shared audience interests among the five organizations to develop joint campaigns, such as environmental or innovation efforts.

These proposals seek to enhance the value obtained from the dataset via improved interaction, sentiment analysis, and strategic decision-making.

CONCLUSION

This study utilized big data analytics for sentiment analysis, employing a dataset of three million tweets related to big companies: Amazon, Apple, Google, Microsoft, and Tesla. In this context, descriptive, diagnostic, predictive, and prescriptive analyses were conducted. Descriptive and diagnostic analytics revealed significant trends in

public opinion toward the selected firms, mainly neutral sentiments, followed by positive sentiments; a lesser quantity of tweets expressed negative sentiments. Predictive analytics was conducted with machine learning models, with the Support Vector Machine yielding the highest accuracy, achieving an F1 score of 96%. These findings provided the foundation for prescriptive suggestions designed to improve brand perception and market performance through focused actions. This research shows that technologies such as Apache Spark and TextBlob are highly successful in managing and analyzing big unstructured data, making them particularly important for real-time sentiment analysis. Furthermore, the offered approaches can aid companies and stakeholders in analyzing sentiment to anticipate changes in markets and adjust to new difficulties. In Future research to improve the depth of sentiment analysis, more advanced deep learning models and real-time data streams may employ. In addition, ethical concerns on data privacy and potential biases in study must be handled to guarantee the proper use of social media data.

REFERENCES

- [1] Yassine Benlachmi, M.L.H., Current State and Challenges of Big Data. Advanced Intelligent Systems for Sustainable Development (AI2SD'2019), 2020.
- [2] Oracle. The Evolution of Big Data and the Future of the Data Platform. 2022; Available from: <https://www.oracle.com/a/ocom/docs/big-data/big-data-evolution.pdf>.
- [3] Tiao, S. What Is Big Data? 2024; Available from: <https://www.oracle.com/sa/big-data/what-is-big-data/>.
- [4] Tim Mucci, C.S. What is big data analytics? 2024; Available from: <https://www.ibm.com/topics/big-data-analytics>.
- [5] Stephen Kaisler, F.A., J.Alberto Espinosa and Wolliam Money Big Data: Issues and Challenges Moving Forward. Hawaii International Conference on System Sciences 46th, 2013.
- [6] Arockia Panimalar, V.S., Veneshia Kathrine, The 17 V's Of Big Data. International Research Journal of Engineering and Technology (IRJET), 2017.
- [7] Oguntimilehin, A. and E.-O. Ademola, A review of big data management, benefits and challenges. A Review of Big Data Management, Benefits and Challenges, 2014. 5(6): p. 1-7.
- [8] Doğan, M., et al. Speculator and influencer evaluation in stock market by using social media. in 2020 IEEE International Conference on Big Data (Big Data). 2020. IEEE.
- [9] Dogan, Ö.M.M. Tweets about the Top Companies from 2015 to 2020. 2020; Available from: <https://www.kaggle.com/datasets/omermetinn/tweets-about-the-top-companies-from-2015-to-2020>.
- [10] Hewage, T.N., et al., Big data techniques of Google, Amazon, Facebook and Twitter. J. Commun., 2018. 13(2): p. 94-100.
- [11] Vanam, H. Sentiment Analysis of Twitter Data Using Big Data Analytics and Deep Learning Model. in 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF). 2023. IEEE.
- [12] Amen, B., S. Faiz, and T.-T. Do, Big data directed acyclic graph model for real-time COVID-19 twitter stream detection. Pattern Recognition, 2022. 123: p. 108404.
- [13] Lin, J. and D. Ryaboy, Scaling big data mining infrastructure: the twitter experience. Acm SIGKDD Explorations Newsletter, 2013. 14(2): p. 6-19.
- [14] Demirbaga, U., HTwitt: a hadoop-based platform for analysis and visualization of streaming Twitter data. Neural Computing and Applications, 2023. 35(33): p. 23893-23908.
- [15] Nodarakis, N., et al., Using hadoop for large scale analysis on twitter: A technical report. arXiv preprint arXiv:1602.01248, 2016.
- [16] Ingle, A., et al., Sentiment analysis of twitter data using hadoop. International Journal of Engineering Research and General Science, 2015. 3(6): p. 144-147.
- [17] Alomari, E., I. Katib, and R. Mehmood, Iktishaf: A big data road-traffic event detection tool using Twitter and spark machine learning. Mobile Networks and Applications, 2023. 28(2): p. 603-618.
- [18] Rodrigues, A.P., et al., Real-Time Twitter Trend Analysis Using Big Data Analytics and Machine Learning Techniques. Wireless Communications and Mobile Computing, 2021. 2021(1): p. 3920325.
- [19] F. C. A. F. C. M. C. Neri, "Sentiment analysis on social media," in 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, 2012.
- [20] J. Yadav, "Sentiment Analysis on Social Media," Qeios, 2023.

-
- [21] S. S. Tushar Rao, "Twitter Sentiment Analysis: How to Hedge Your Bets in the Stock Markets," in *State of the Art Applications of Social Network Analysis*, Cham (Location of the publisher), Springer International Publishing, 2014, p. 227–247.
 - [22] M. V.-G. A. M. P.-P. Juan Piñeiro-Chousa, "Influence of Social Media over the Stock Market," *Psychology & Marketing*, vol. 34, no. 1, pp. 101-108, 2017.
 - [23] D. P. A. C. A. T. A. L. M., *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*, 2nd ed., O'Reilly Media, 2022.
 - [24] A. R. Patel and S. Kumar, "Predictive Analytics in Retail: Forecasting Future Trends Using Big Data," *Journal of Predictive Analytics*, vol. 34, no. 6, pp. 401-418, 2019.
 - [25] H. L. Wilson and P. G. Williams, "Prescriptive Analytics: From Insights to Actionable Strategies," *International Journal of Business Analytics*, vol. 14, no. 1, pp. 35-50, 2020.
 - [26] D. J. McKinsey and T. S. Thompson, "The Role of Descriptive Analytics in Big Data Decision Making," *Journal of Business Intelligence*, vol. 19, no. 3, pp. 210-225, 2018.
 - [27] S. Loria, "TextBlob: Simplified Text Processing," Technical Documentation, 2018.
 - [28] Silva, M.D. *Preprocessing Steps for Natural Language Processing (NLP): A Beginner's Guide*. 2023; Available from: <https://medium.com/@maleeshadesilva21/preprocessing-steps-for-natural-language-processing-nlp-a-beginners-guide-d6d9bf7689c9>.
 - [29] Cunningham, P., M. Cord, and S.J. Delany, *Supervised learning*, in *Machine learning techniques for multimedia: case studies on organization and retrieval*. 2008, Springer. p. 21-49.
 - [30] Mahesh, B., *Machine learning algorithms-a review*. International Journal of Science and Research (IJSR).[Internet], 2020. 9(1): p. 381-386.