# ANALYZING SOCIAL MEDIA DATA TRENDS USING BIG DATA TECHNIQUES

**Lance Cerejo**
**3108834**

**Submitted in partial fulfilment for the degree of**
**Master of Science in Big Data Management & Analytics**

**Griffith College Dublin**
**June, 2024**

**Under the supervision of Dr. Aqeel Kazmi**

**Disclaimer**

**I hereby certify that this material, which I now submit for assessment on the programme of study leading to the Degree of Master of Science in Big Data Management & Analytics at Griffith College Dublin, is entirely my own work and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfilment of the requirements of that stated above.**

**Signed: _____Lance_____ Date:_____10/06/2024_____**

\

## Acknowledgement

I would like to express gratitude to my Professor, Dr. Aqeel Kazmi. I am sincerely grateful to him for his advice and guidance throughout the academic year. I thank him for his patience and faith in me. I am sure to lay the basis for my scientific eagerness, and I hope that our fruitful cooperation in the scientific field will continue.

## Abstract

Social media accounts have gained popularity in the recent past and concurrently, the generation of social data has gone up tremendously. Employing and applying current day general data analysis tools, this study seeks to filter out information from the enormous sphere of social media for decision-making purposes. The work examines the issues and prospects that come with employing big data in the analysis of social media trends, with the ultimate goal of filling the gap between data excess and useful application. The mentioned research objectives include studying current methods for the capturing and archiving of social media data, studying the possibilities of big data techniques in analyzing the data accumulated in social media platforms, and outlining the difficulties and further perspectives for the utilization of big data technologies in analyzing trends in social media platforms. Therefore, the study serves the research objectives by enhancing the comprehension of culture, communication patterns and the society in the digital era. The present study concludes that social media analytics hold a critical role for revealing patterns, behaviors, and dynamics in the context of social media communications.

**Table of Contents**

**List of figures**

## List of Tables



## List of Abbreviation

SMA - Social Media Analytics

AI - Artificial Intelligence

ML - Machine Learning

NLP - Natural Language Processing

SVM - Support Vector Machine

RMSE - Root Mean Squared Error

SVC - Support Vector Classifier

IT - Information Technology

GDPR - General Data Protection Regulation

ROI - Return on Investment

ARIMA - Autoregressive Integrated Moving Average

SMA - Social Media Analytics

## List of Equations

# Chapter 1: Introduction

## 1.1 Introduction to the study

Social media platforms in the time of digitalization have developed in a dramatic way and this trend is accompanied by an increase of data collection beyond any expectation. In this article, the study seeks answers to whether the big data approach is capable of analysing trends in studying social media data. Using modern data analysis techniques, this research is designed to distil from the vast universe of information that is generated from several social media platforms, those insights which are relevant and in turn can be used for taking action. This study aims to show the installed strength of the big data system to bring to life specific patterns, behaviours, and dynamics existing within social media interaction, and by that add some fundamental knowledge of digital culture and communication trends.

### 1.1.1 Purpose of the study

**Aim**

This research aims to investigate the effectiveness of big data techniques in analysing social media data trends and their implications for various sectors.

**Objective**

- To examine the current landscape of social media data and its significance.
- To explore the application of big data techniques in analysing social media data.
- To identify challenges and opportunities in utilizing big data for social media trend analysis.

### 1.1.2 Problem statement

These data can offer organizations difficulties in order to structure and analyse them into actionable insight which is a result of the great volume and complexity of it all while there is an abundance of data on social media. The objective of this study is therefore to attach an analytical strategy to the social media trends that can cope with the methods of big data. Its main aim is to offer practical directions and suggestions for both firms and scientists. Data analysis in social media and the utilization of big data processing characteristics is the main objective of this research. It aims at connecting the data abundance with meaningful insight interpretation. By the means of the revelation of effective analytical strategies and the search for new technological ways, the purpose of this research is to prepare practical instruments for businesses to maximize the benefits of social media data for informed decision-making and strategic planning.

## 1.2 Background of the study

Social media applications have smoothly trickled into daily life impulsively and serve as a primary tool for communication, information sharing and global networking. Platforms such as Facebook, Twitter, Instagram, and LinkedIn have billions of active users. With these platforms amassing together lots of collections of data that embrace all the topics, demography, and sentiments, these can refer to the tremendous data deposits (Abkenar *et al.* 2021)[1]. This ubiquitous shell has reshaped social media into an incomprehensible data bucket that has amassed knowledge about human behaviour, population trends, and market dynamics. The platforms are taken by stormers thanks to their effectively changing and colossal user names which creates great data to be used by sales or marketing teams. As a result, the social media landscape and its transformation must be stored in the mind so as to realize how this communication and information-sharing experience is so complex and intricate (Zadeh *et al.* 2021)[2]. This backdrop reopens the possibility of leveraging the ever-growing volume of information that comes from the digital sphere in order to study and understand the complexity of social media.

## 1.3 Research rationale

Besides, this study is prompted by the growing importance of data analytics in various sectors such as health care, consumer industry, education and much more. How such social media

data reveals the essentials behind people's consumption patterns, market drops and the society's general outlook is a valuable chance to get enlightened. Having big data techniques at one's fingertips enables organizations to generate viable strategies as a result of the inundation of data from the ever-changing landscape of social media space. Companies gain more insights into these patterns thus enabling them to fine-tune their strategic plans and marketing intents, always being in advance of any shifts in consumer needs or market conditions (Javed Awan *et al.* 2021)[3]. On the contrary, the use of big data analytics is not only functionally the most efficient but also leads companies to seek innovation and hold the leading position. For this reason, this research brings to light how social media data is so vital and gives another framework for superior decision-making and organisational success in the data-driven era.

### 1.4 Research questions

***Q1:*** How can big data techniques be applied to analyse social media data trends effectively?

***Q2:*** What are the challenges associated with analysing social media data using big data techniques?

***Q3:*** What are the implications of social media data analysis for businesses, researchers, and society?

### 1.5 Significance of the study

This research may have rich content in the academic community, business sector and ultimately in the society at large. This analysis technique based on big data that uncovers talking points on social media from consumers brings business to a higher level in making data-driven decisions. As a result, they can compete and adapt well as they are being confronted with changes in the market landscape. Besides, when it comes to research data analytics, it is not only a source of information but also a source of breakthrough findings that can transform the field ever for the better. Aside from its immediate implications, this research is especially beneficial for building an understanding of digital culture and communication dynamics (Chaudhary *et al.* 2021)[4]. Thereby, it helps to see what is going on in society concerning digital space, behaviour, and interactions. These results can bring about a significant transformation equipped with the most viable policy options, for instance, government, academia, and civil society, and suggest ways that knowledge can be crowned, discovered, and better understood.

### 1.6 Overview of Research Design

This dissertation consists of seven chapters. As a basic science review, Chapter 2 aims first to immerse in the existing literature and build a solid foundation on the topic. Chapter 3

shows the research methodologies, and that has the approaches used to acquire sentiments, modelling and analysis. The social media data analytics outcomes are abstruse in Chapter 5 and 6 by using the methodology of big data approaches to draw the trends and inferences of social media data. Chapter 6 presents the research implications and the managerial ones. It addresses these implications from the point of view of the modified academy, the industry, and the society as a whole. Finally, Chapter 7 gives a discussion of the main findings derived from the research, highlights areas for additional work, and makes a closing statement.

## 1.7 Assumptions, Limitations, Delimitations

### *Assumptions*

The research functions under several basic 'hypotheses' from the beginning. The quality of any analysis is established by the availability of enough data volume of social media data that needs to be analysed further. Secondly, it anticipates unquestionable appropriateness and faith-worthiness as well as big data techniques used in the research method. Finally, the study evokes the involvement of the participants in the research process, through exchanging data and producing insights for investigation.

### *Limitations*

Nevertheless, those aforementioned assumptions are not the only ones which are taken and some other limitations should also be noted. Limited time may often not provide enough space for deep thought during the data-collecting or analysing process, which in turn may affect the nature of conclusions drawn (Zhu *et al.* 2020)[5]. In addition, companies may also possess private data or use restricted platforms, which could impede access to all-embracing datasets to determine the extent of the research. Moreover, the sampling bias of data and the systematic deviation from the methods of the analysis could result in errors and weaken generalizable conclusions.

### *Delimitations*

Also, it is vital to consider that this report has some restrictions. They include the profiles the researcher focuses on or the social media platforms the research looks at during a certain period which in turn can limit the generalization of findings to the whole population. Moreover, the study may be constrained by the nature and area of the undertaken survey, which also might be the case, and the results may not be generalizable to the whole population. However, with certain limitations, the study is poised to offer useful conclusions on contemporary social media indicators and different big data approaches for the evaluation data.

# Chapter 2: Literature Review

## 2.1 Introduction

In this study, using big data techniques for analysing social media data trends the literature review section delves into existing research and studies related to this topic, where the aim is to explore what previous researchers have discovered about the effectiveness of big data in understanding the social media patterns as well as behaviours. It includes looking at how the social media platforms such as Facebook, Twitter, Instagram, as well as LinkedIn have evolved and amassed vast amounts of data about human behaviour as well as the dynamics of the market. This chapter also aims to identify any gaps or areas where more research has been needed. This study can contribute new insights as well as the build upon existing knowledge by understanding what areas remain explored and what has already been studied. Overall, by providing a comprehensive overview of the current state of the knowledge in the field of big data analytics for social media the literature review sets the foundation for the research.

## 2.2 Empirical Study

### 2.2.1 Exploring the Sentiment Analysis Trends in Social Media During the COVID-19 Pandemic

These studies have delved into sentiment analysis, a popular area in natural language processing as well as particularly focusing on social media data according to (Manguri *et al.* 2020)[6]. This exploration involved methods such as opinion mining as well as the detection of events, which is essential for understanding public sentiments on the various topics. Social media platforms such as Twitter, Facebook, as well as YouTube have become invaluable sources of the social data, which has been reflecting real-life events and discussions. The significance of analysing sentiments expressed on these platforms where the emergence of the COVID-19 pandemic has been highlighted. One notable finding from this research is in uncovering the public opinions and emotions related to COVID-19 the substantial impact of sentiment analysis, where the researchers have utilised tools such as Python programming language, Tweepy library, as well as TextBlob library for sentiment analysis on Twitter data. Using specific hashtags related to COVID-19 and coronavirus, this approach involved extracting tweets, where through sentiment analysis, researchers identified the trends in sentiment polarity such as positive, negative, and neutral as well as subjectivity as objective, and subjective among the users of Twitter.

The results of these empirical studies revealed intriguing insights, for instance, a significant portion of tweets over the 50% exhibited a neutral sentiment, where it is indicating a factual or informational tone rather than the emotional expression. This neutrality has been consistent across different days, also about the pandemic suggesting a trend in how people share the information. In addition, the research highlighted the predominance of the objective viewpoints in tweets, which is indicating a focus on relaying the information rather than the personal emotions or opinions.

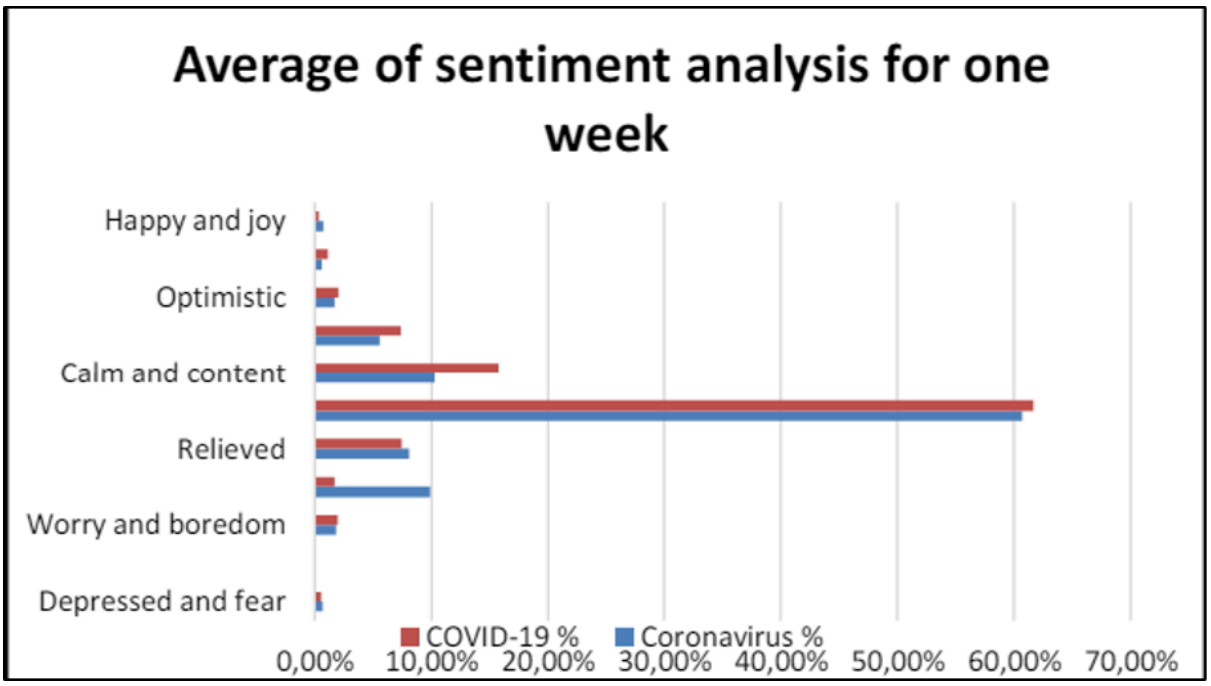


**Figure 2.1: Average of sentiment analysis of emotional guidance scale**

(Source: Manguri *et al.* 2020)[6]

In methodology, these studies employed the machine learning algorithms such as supervised learning (e.g. Naive Bayes classifiers) as well as the lexicon-based algorithms for sentiment analysis, also utilised sentiment identification techniques, the methods of feature selection, as well as sentiment classification models to categorise the tweets based on the sentiment polarity and subjectivity. However, there remain gaps in understanding the nuanced shifts in public sentiment over time, despite these valuable findings, especially concerning the major events or developments related to the COVID-19. In addition, further research has been needed to explore the impact of the sentiment analysis results on the decision-making processes within the healthcare, government, as well as the media sectors. Overall, these studies provide a comprehensive overview of the sentiment analysis techniques applied to the social media data during the COVID-19 pandemic, where they underline the importance of understanding the public sentiments for informed communication strategies and decision-making. These studies serve as a foundation for the current study, which aims to contribute the additional insights into the sentiment trends as well as their implications in the context of COVID-19 discussions on the social media platforms.

### 2.2.2 Enhancing the Start-up Firms of Social Media Marketing Through Predictive Analysis

This study delves into the vital role for start-up companies of social media marketing according to (Jung and Jeong, 2020)[7]. This study highlights how the social media platforms offer cost-effective marketing solutions as well as to direct communication channels for firms, where it focuses on predicting the engagement levels of the start-up firms on Twitter, also applying the data science techniques and the machine learning models. The study aims to provide insights into the effectiveness of the social media marketing efforts as well as identify the key factors influencing engagement levels, by analysing 8,434 start-up firms' data from Twitter.

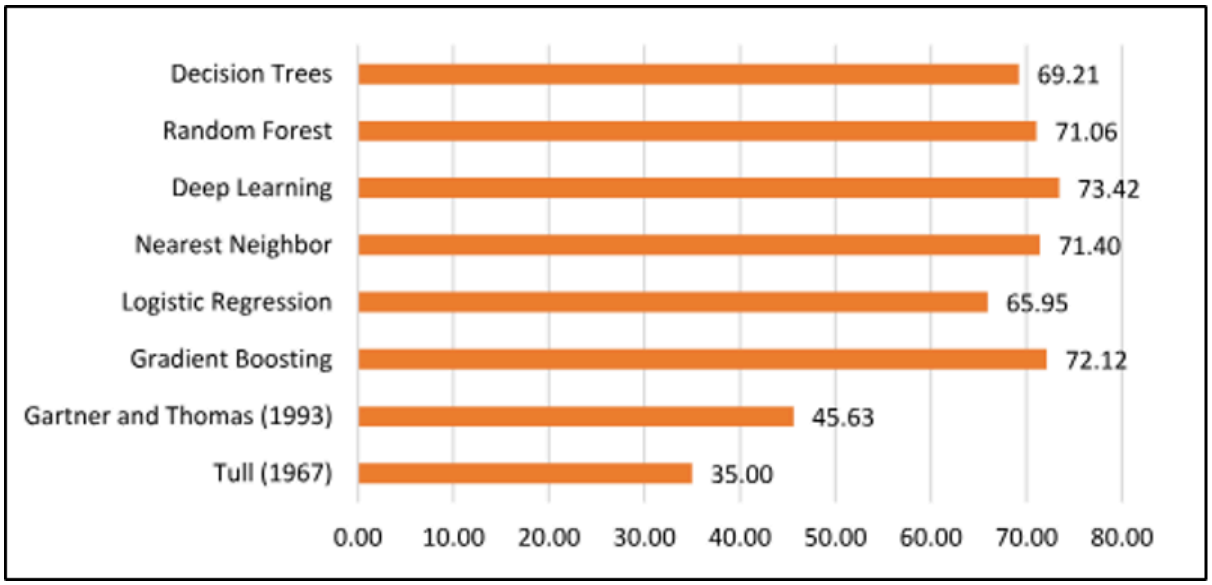


**Figure 2.2: Accuracy Comparisons**

(Source: Jung and Jeong, 2020)[7]

This study showcases that in predicting social media engagement levels for start-up firms the deep learning models outperform the other machine learning methods, where the features such as the number of tweets, retweets, as well as the likes received emerged as significant determinants of the social media marketing effectiveness. The accuracy of the deep learning model reached approximately 73.42%, which indicates the capability in predicting engagement levels accurately. In the methodology it involves gathering as well as cleaning data from Twitter, generating the social media-based features, as well as applying the machine learning models for predictive analysis, where various machine learning algorithms have been used. Machine learning model names such as deep learning, decision trees, logistic regression, random forest, nearest neighbour, as well as the gradient boosting are employed and compared for their predictive accuracy.

Social media engagement by introducing a predictive analysis approach tailored to this empirical study fills a critical gap in the start-up firms' marketing strategies, where the traditional marketing forecasting methods often lack accuracy, whereas the machine learning techniques offer more precise insights. The research provides a foundation for integrating data science as well as machine learning into start-up firms' marketing strategies, where it enhances their competitiveness and market insights. For effective social media marketing the study emphasises the growing importance of leveraging the advanced analytical tools such as deep learning.. Start-up firms can optimise their marketing efforts by focusing on key social media metrics, also build brand awareness, as well as improve customer engagement, leading to better competitive advantages and financial performance.

### 2.2.3 Exploring Twitter as a Research Platform: Themes, Methods, and Challenges

According to (Antonakaki *et al.* 2021)[8], use of Twitter as a research platform this study delves into the extensive due to its vast user base as well as straightforward data access, where it highlights the Twitter's significance in analysing online behaviour patterns, sentiment analysis, social graph structures, as well as threats like fake news, spam, bots, and hate speech. The research aims to map the current research topics on Twitter, which focuses on the key areas like sentiment analysis, the social graph's structure, as well as addressing threats within the platform. Utilised in research across various disciplines the study reveals that Twitter is widespread, with over 27,000 articles mentioning Twitter in their titles, where it emphasises Twitter's unique properties, including their dual role as a social network as well as a news dissemination platform. Compared to Facebook despite its lower user base, Twitter serves as a valuable tool for assessing the public opinion, sentiment, trends, as well as beliefs swiftly. For diverse studies ranging from natural disaster response researchers utilise Twitter to corporate public relations management, as well as showcasing its importance in modern science.

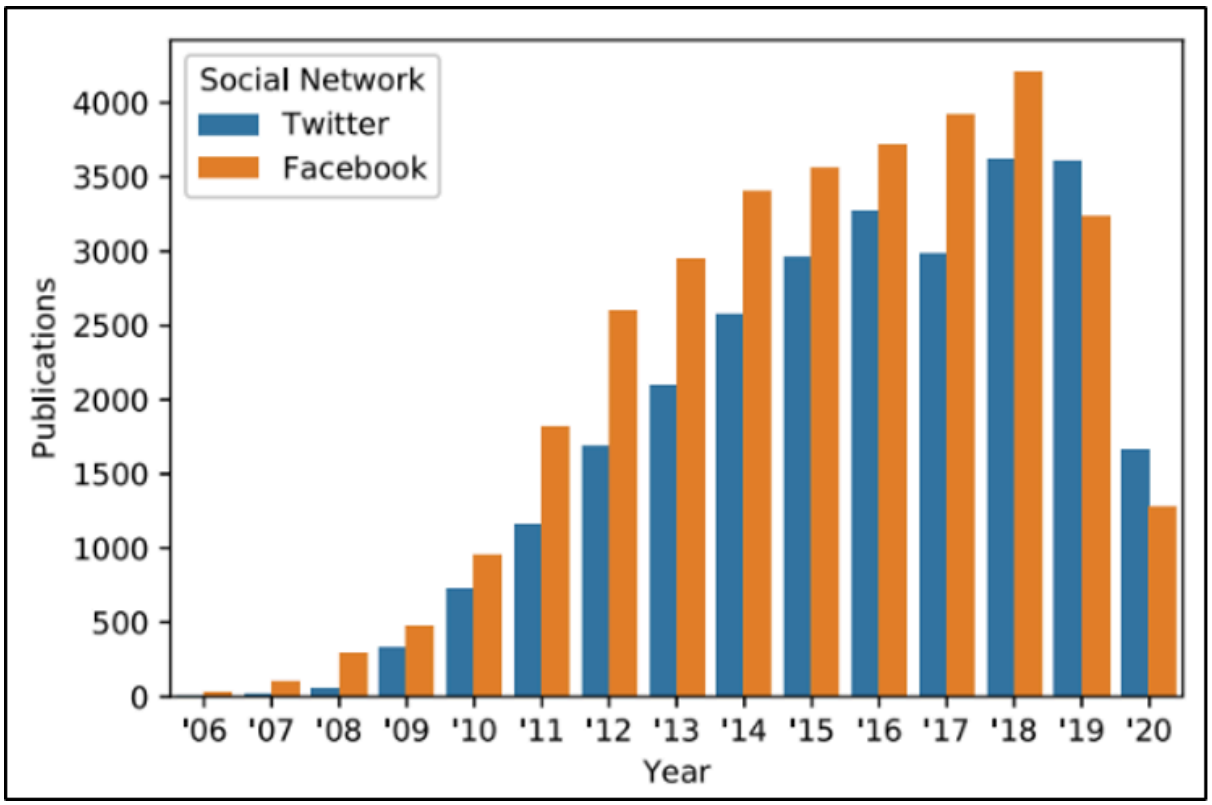


**Figure 2.3: The scientific publications containing the word 'Twitter' and 'Facebook' on their title from 2006 to 2020**

(Source: Antonakaki *et al.* 2021)[8]

This study employs computational techniques like Natural Language Processing (NLP), Graph Sampling, as well as Machine Learning (ML) to analyse the Twitter data, where it discusses strategies for data collection, sentiment analysis, as well as identifying threats such as spam, fake news, bots, and the hate speech. Also, the survey touches upon challenges such as the unavailability of the gold-standard datasets due to Twitter's Terms of Service. Twitter's immense popularity as well as the unique characteristics make it a valuable resource for researchers across disciplines, also the aim is to provide a comprehensive overview of Twitter's research landscape. It highlights the major themes, methodologies, as well as challenges, and also emphasises the need to understand the basic features of Twitter's, dynamics, content, as well as the potential dangers before utilising it as a research platform. The study identifies the gaps in existing research, particularly in addressing Twitter's evolving nature as well as the challenges such as spam, bots, and hate speech, where by mapping current research topics and the methodologies. To optimise Twitter's utilisation as a research object for scientific discovery this study sets the groundwork for future studies.

Overall, for continuous updates as well as advancements in studying this dynamic online social network it underscores Twitter's importance as a vibrant research platform and calls.

### 2.2.4 Forecasting Future Trends Using ARIMA Models

ARIMA is a widely known statistical model for the prediction of the future values of a variable measured over time, namely time series data. It has been used notably in different fields to forecast the outcomes that are going to be in the future by analyzing past data. A representative application of the ARIMA model regarding social media data is discussed in the works of where the model is used in predicting the tendencies of the users' engagement and activity in social networks. One of the major strengths of ARIMA models and especially for time series data is that it is able take into account certain dependencies like trends and seasons (Shang *et al.,* 2021)[9]. Due to the flexibility in dealing with different data patterns, the proposed model can be used for accurate forecasting in environments that are fast-paced and constantly changing, such as social media. In the study the ARIMA model was applied to forecast user activity in the future, to the platforms such as Twitter, Facebook by using previous indicators such as the number of posts, likes, shares, comments.

The implementation involved several steps: including data gathering, data preparation, determination of an appropriate model, estimation of parameters of the model, and assessment of the adequacy of the model. The first activity the researchers performed in the study was to collect a large number of users' actions over multiple years. Excluding the noise and seasonality, the method to select the right values for the parameters of the ARIMA model is p, d, and q. The developed model was then used to predict the trends for the future and the validity of the model was then confirmed with the actual subsequent data. As viewed in the outcomes, it was revealed that the ARIMA model would satisfactorily predict short-term patterns of users with greater accuracy (Islam *et al.,* 2020)[10]. This characteristic is instrumental for the organizations and the marketing professionals who usually base their decision on the social media statistics. For example, the effectiveness of users' interaction can be predicted to plan marketing campaigns, premises of content publishing, and the strategies for users' engagement.

The ARIMA (AutoRegressive Integrated Moving Average) model claims to be the statistical modelling on which time series analysis and forecasting of future values is based, and which is useful in many fields. Particularly in the field of SM analysis, how ARIMA can forecast the tendencies of users' engagement and activity are demonstrated. ARIMA models are useful when it comes to working with dependencies within the time series data including

trends and cyclic patterns, which are very ideal for social media platforms (Aliahmadi *et al.,* 2022)[11]. Regarding social media analysis, prior application of the ARIMA model has been used specifically for the comprehension of further user activity on similar online platforms like Twitter and Facebook. These implementations typically involve several key steps: the process of collecting data from users' records, cleaning it by filtering out noise and other time-dependent effects, determining the values of p, d, and q parameters to specify the exact form of the ARIMA model, and comparing the results of the model simulation with its further outcomes. The strength of ARIMA is its efficiency for short-term pattern prediction of users' activities, which is very useful for organizations and marketing experts (Singh *et al.,* 2021)[12]. For example, it is possible to predict by how much the users' interaction level will increase during certain days of the week, thus helping to determine when the marketing campaigns should be launched, which content should be published, and how the further levels of interaction can be improved.

In addition to their utilization in forecasting, the concepts of time series and in particular the ARIMA model is not just identifying trends, but offering solutions that can help in decision-making processes thus enhance the efficiency of the utilization of the social media for marketing purposes and other related activities (Grover *et al.,* 2022)[13]. Thus, as social media remains an ever-expanding phenomenon, continued investigation of the specifics of ARIMA model and its impact on social media analysis will be useful and valuable for adjusting its capabilities to serve a variety of decision-making purposes.

### 2.2.5 Past Work Implemented

TextBlob and VADER have been normally applied in sentiment analysis in earlier studies. TextBlob is one of the simplest and easy to use libraries for text processing, which includes parts of speech tagging and sentiment analysis. VADER, however, is focused on the analysis of social media texts with the help of the lexicon and the rule-based sentiment analysis tool.

***Enhancing Sentiment Prediction with Machine learning***

Experiments incorporating these tools with regression and classification models such as logistic regression, SVM, and random forests have enhanced the prediction of sentiment as well as classification of text with a lot of potential in NLP tasks.

TextBlob and VADER have been helpful in prior studies especially in performing sentiment analysis. Its features benefit from the straightforward interface, and the existing preprocessing activities, such as POS tagging and sentiment analysis.

***Comparison of VADER with other Tools***

While REL-Q is generally specialized in interpreting texts from web 2.0 sources, VAL and DALE are rule and dictionary based sentiment analysis tools that are aimed at VADER: Valence Aware Dictionary and sEntiment Reasoner. TextBlob and VADER have been widely utilized in sentiment analysis in many research works. VADER which is used for sentiment analysis of texts in social media employs the use of lexicon and rule-based formula. Both tools have been linked with regression as well as classification models like logistic regression, support vector machine, and random forest for boosting the sentiment prediction and text classification capabilities.

***Effectiveness of VADER in Social media analysis***

In earlier studies, these tools have been incorporated with regression as well as classification models for improving sentiment prediction as well as the accuracy of text classification (Coffey *et al.,* 2021)[14]. Comparing VADER with other lexicon-based sentiment analysis programs, explanation of how it is successful in the analysis of the tweets texts, as well as other texts of social media. It has also explained how VADER was more effective than previous tools for sentiment analysis in social networks when it comes to handling slang, emojis, and context expressions.

TextBlob and VADER with machine learning like logistic regression, support vector machine, and random forest helped in enhancing the performance on the NLP tasks. Twitter data to perform logistic regression as well as Naive Bayes classifiers with fairly high accuracy for sentiment classification. Sentik developers' similar experimental work which used SVM and random forests and where they incorporated VADER for feature extraction, offered improved sentiment analysis models.

## 2.3 Literature Gap

However, based on the existing literature on BDA and SM, there are still some research issues that have not been explored enough. First, although there are numerous investigations exploring the sentiment analysis during the COVID-19 pandemic, it is still possible to identify the minimum research concerning flexible shifts in sentiment at considerable time intervals and bizarre areas. This deficit is crucial because awareness of these changes can give greater perspective on the social effects of the virus and improved methods of outreach messaging. Moreover, the Inclusion of sophisticated artificial neural network approaches in the promotion of social media for the start-up firms has been proved effective in improving the level of engagement prediction. However, in the literature, there is a lack of understanding of how these predictive insights are applied by start-ups in the context of

marketing and with what enduring consequences for their performance. More research papers are still required to cover this gap and most of them should be case studies which establish the application of the said technologies by successful start-up firms.

| Author Name | Topic | Description of Gap |
|---|---|---|
| (Abbate *et al.,* 2024)[15] | Social Media Trends | Despite extensive research on social media trends, there is a lack of focus on their impact on consumer behavior. |
| (Kushwaha *et al.,* 2021)[16] | Big Data Techniques | Existing literature on big data techniques in social media analysis lacks comprehensive evaluation of their effectiveness. |
| (Wang *et al.* 2020)[17] | Ethical Considerations | While ethical considerations in social media analytics have been discussed, there is a need for more practical guidelines for researchers. |
| (Lawn *et al.,* 2020)[18] | Longitudinal Studies | Limited literature exists on the long-term trends and evolution of social media interactions over time. |
| (Lawn *et al.,* 2020)[18] | Integration of Datasets | The literature gap lies in exploring the challenges and opportunities associated with integrating diverse datasets from multiple social media platforms. |
| (Sivarajah *et al.* 2020)[19] | Predictive Analytics | There is a gap in understanding the potential of predictive analytics models in forecasting social media trends accurately. |
| (Khanra *et al.* 2020)[20] | Cross-cultural Analysis | Existing research lacks cross-cultural analysis of social media trends, hindering the understanding of global digital communication dynamics. |

**Table 2.1: Identified gaps in literature**

The use of Twitter as a research source is now familiar, however, the dynamics of Twitter's and its users are constantly changing. Recent publications sometimes fail to take into

consideration the changes that are dynamic in nature in the functionality of Twitter and these changes affect the reliability and validity of the collected data. However, there is a shortage of resources to resolve problems in the inadequate approaches in the steps of data privacy, ethical issues and the reliability of the results. Finally, even though the trend analysis of the future periods has been performed with the help of ARIMA models, the usage of these models in SA for determining the trends in social media is still limited. Namely, it is possible to discuss the lack of studies that employ ARIMA models with regard to the long-term forecasting of user engagement in the context of social media platforms. This should be an area of focus in the future research since a better accuracy in the forecasts considerably improves strategic planning for organizations and policy-making.

These gaps will not only contribute to developing the theoretical framework with regards to big data analytics but also enhance the practice of BM applications for better decision-making processes across the different sectors.

**2.4 Summary**

In literature review, a wide range of SMA-related subjects are introduced, some of which include recent trends in SMA; big data approaches; ethics in SMA; longitudinal SMA studies; a research-integrated data set; predictive analytics; and SMA across cultures. Researchers have also pointed out the following research limitations: absence of the study of social media trends' effects on consumer behavior, the lack of the big data methods' complete assessment, lack of guidelines on ethical concerns for social media, enterprise longitudinal studies that could map the changes in social media interactions, issues and prospects of incorporating various data types, the strengths of the predictive analytics models, and cross-cultural studies of social media phenomena. Filling these gaps will contribute to research in the area of digital culture and communication as well as help to unravel the dynamics of society in the age of digital technology with a view to promoting more future-oriented decision making on data use in social media analysis.

# Chapter 3: Methodology

## 3.1 Introduction

The assessments have been performed on the three main elements that made up the research analysis which are data collection, and data analysis. The research strategies have been analysing the social media data and Twitter trends and hashtags trends. The process of data collection involves sourcing the data from diverse classes, including trend datasets that spanned from both May to October in 2021. It has been aimed to contribute to the existing body of knowledge on Twitter trends by adopting a deductive approach with secondary data analysis. The dataset was further expanded to include hashtag trends and Twitter trends properties. Information of exploratory data analysis, and the use of machine learning models like Logistic regression for predictive analysis which are encompassed in the process of data analysis. It has been performed on the secondary data analysis and a deductive research procedure to expand the knowledge of Twitter patterns. The big data analysis and tweet quantity and retweet count evaluation, that can be helped out to track social media campaigns' effectiveness.

## 3.2 Research Strategy

For this study, the research strategy involves three main components such as literature review, data collection method, as well as the data analysis.

***Literature Review***

Twitter trends and social media data has been thoroughly examined in the literature review phase, existing research as well as the studies related, where it involved analysing the previous works, scholarly articles, as well as the relevant publications to gain a comprehensive understanding of the topic (Tao *et al.* 2020)[21]. The goal has been to identify gaps in the existing research as well as to build upon the knowledge already available.

***Data Collection Method***

In this study, the data collection method utilised which involved gathering information from the various sources related to the Twitter trends, where it included accessing the datasets from different time periods from May to October, 2021 that contained the data on all trends, hashtag trends, as well as the Twitter trending topics. Also, the data collected consisted of retweet counts, tweet volumes, tweet text, as well as the other relevant attributes. It was beneficial to understand and gain insights about user sentiment over time.

***Data Analysis***

The data analysis process has included several steps from the collected data to derive meaningful insights, where the missing values in the datasets have been addressed by 'Iterative Imputation' to deal with the missing values in the datasets, this method of imputation replaces the missing values with plausible estimates based on the rest of the data with non-missing values, removal of outliers from the dataset. It would then be creating a complete dataset which would be beneficial for further data analysis or modelling.

To understand the distribution as well as relationships within the data exploratory data analysis techniques, like summary statistics and visualisations such as bar charts, have been employed. In addition, the machine learning models, specifically Random Forest regression, Decision Tree Regression along with Random Forest Classifier and Linear Support Vector Classifier has been utilised for predictive analysis, where these models have been trained and tested using the features extracted from the datasets. Two algorithms have been used to get the polarity of the sentiments based on the words within that tweet text, Firstly TextBlob has been used, which uses a combination of machine learning algorithms and ruled-based systems for sentiment analysis. And Secondly Vader, Vader uses a lexicon-based method with the combination of rule-based heuristics tailored for social media text.

Using metrics, the accuracy of the models has been evaluated like accuracy scores, classification reports, as well as the confusion matrices, all of which have been well visualized.

Overall, to uncover insights as well as patterns in Twitter trends data the research strategy involved a comprehensive review of the existing literature, meticulous data collection from the diverse sources, as well as rigorous data analysis techniques.

### 3.3 Research Approaches

The research approach used in this study has been deductive, which is focusing on secondary data analysis, where a deductive approach involves starting with a theory or hypothesis, and using collected data then testing it. Existing theories and knowledge about the Twitter trends as well as social media behaviour in this case are the basis for forming hypotheses (Iqbal *et al.* 2020)[22]. Through the analysis of secondary data this study aimed to validate or refute these hypotheses, which refers to the data that has been already collected by other sources as

well as made available for the research purposes. It has been using secondary data sets, which could leverage the vast amount of information present in the datasets spanning different time periods from May to October 2021 related to Twitter trends, where this approach allowed for a systematic examination of trends, patterns, as well as the relationships within the data (Liu *et al.* 2021)[23]. Aligning with the deductive reasoning process of moving from the general theories to specific observations, where to uncover insights into tweet volumes, retweet counts, as well as the popularity of trends the research approach involved analysing the secondary data using statistical as well as the machine learning techniques. The study aimed to contribute to the existing body of knowledge on Twitter trends by adopting a deductive approach with the secondary data analysis, while also testing and refining the existing theories as well as hypotheses in the field of social media research.

### 3.4 Research design

The research methodology has encompassed the three main elements of the sampling strategy, and the analysis of the data. It can be run through assessments of this research and academic works by comparing with Twitter trends and other social media data in the process of big data analysis (Balaji *et al.* 2021)[24]. The procedure of data collection relied on the data sources from Twitter trend data with the time series design of this period being May to October 2021. The data sets have been based on variables including retweet rate, tweet counts, and tweet text as well as surrounding trends, hashtags trending, and trending topics on Twitter. The data collection process was designed to obtain a complete and diverse dataset for the analysis.

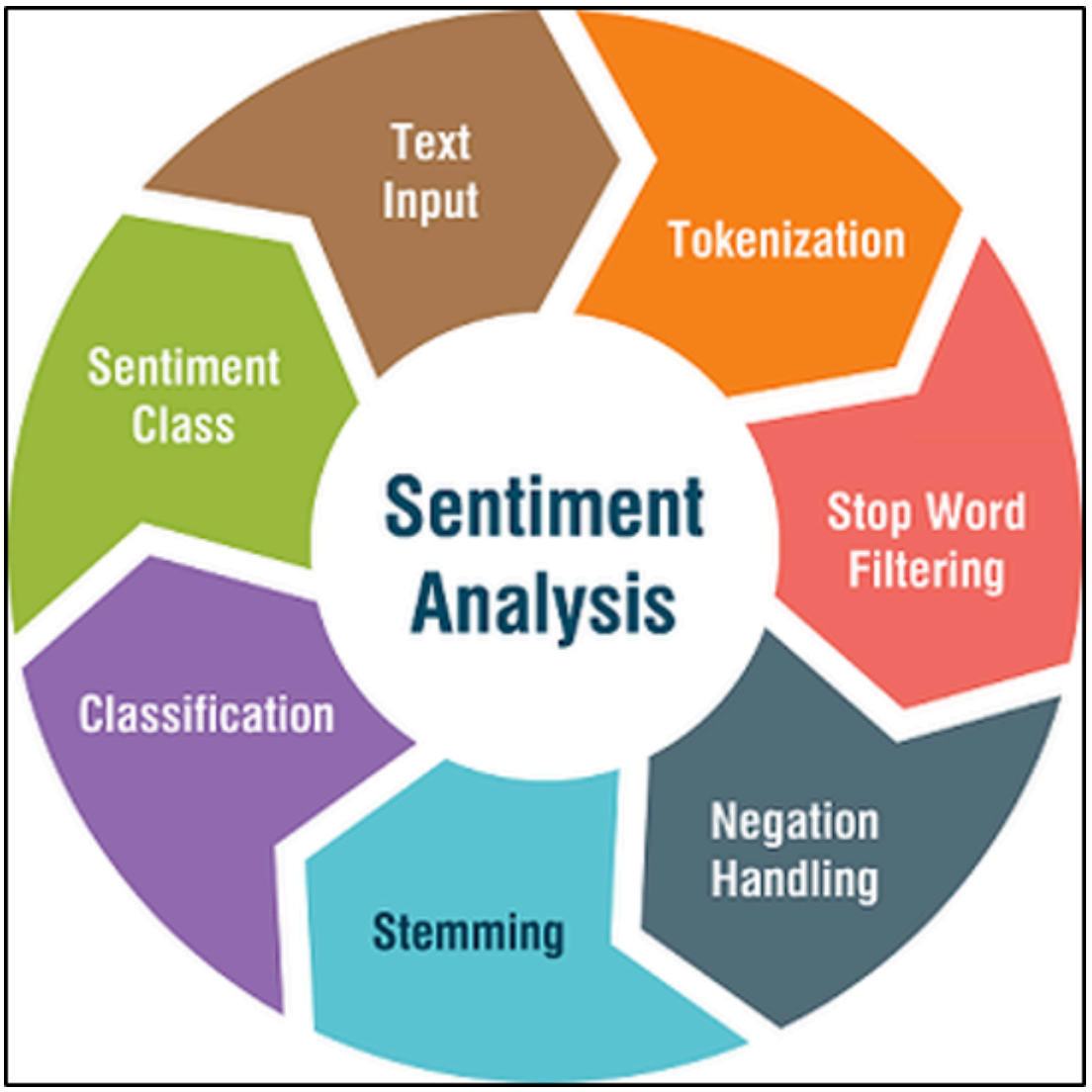


**Figure 3.1: Social Media Analytics**

(Source: Datacamp, 2021)

The data analysis was performed by deploying several mechanisms. It can allow the data analysis and derive understanding from the collected data of the displayed measurement. It was also discussed how to deal with the missing values. It can be performed by filling them with the missing values to remove the missing values into the data sets (Kauffmann *et al.* 2020)[25]. Data analysis techniques, including descriptive statistical reports and visualization, aided in the comprehension of the population distribution and data relationships. Also, a logistic regression model and machine learning algorithms were employed for predictive analysis in categorizing the level of popularity of tweets and the retweet counts. The research design could help to conduct a systematic analysis of Twitter

data trends, and hashtag trends, to add to the existing knowledge in the field of social media research.

### 3.5 Data Collection Method

The process is mainly based on collecting data, which includes many sources of information, but mainly through Twitter trends and hashtag data trends analysis. The datasets have been retrieved, and cover the different periods of the time from May to October 2021. Social media monitoring is heavily data-driven for big data analysis (Sivarajah *et al.* 2020)[19]. All user generated posts, hashtags, and key conversation topics are collated and stored in databases. This is a very important companion for the intensive analysis of this research. The extracted data contained many attributes, such as retweet counts, tweet volumes, tweet contents, and other data that related to trends, which were considered to have great importance in the popularity and the spread of these trends. Such an approach, however, allowed the creation of a database, which was based on a total analysis of patterns of social media (Misra *et al.* 2020)[26]. Precisely, these particular time durations and specific data types were key in having a better understanding of the changes in social media patterns and how they affected and contributed to increased digital communication. Besides the primary trend of the data sets, the gathering process used the prioritized ancillary data categories such as user engagement metrics that consisted of user follower counts and tweet source information.

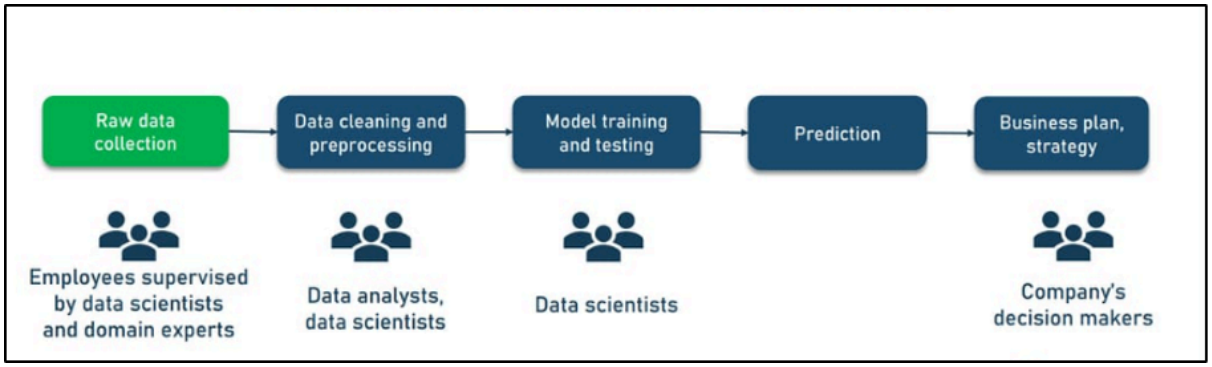


**Figure 3.2: Data collection in the decision-making process**

(Source: Altexsoft, 2023)

The dataset visualization using K-means clustering is used to the spread of trends within and between the diverse group of users on social media. The data that are being used are checked carefully to make sure that relationships within the data exploratory, data analysis techniques,

like summary statistics and visualizations such as bar charts, and heatmaps have been used in the data collection method (Sheng *et al.* 2021)[27]. The fact that the data was kept in order, made the research findings reliable and thus served as a solid foundation for the next stages of analysis. The research endeavour followed a rigorous and meticulous methodology to gain a clear picture of Twitter trends and their effects on digital connection and brand promotion.

### 3.6 Implementation Technique

Big data analysis has been applied which includes an inventory of techniques such as data mining and analysis to analyse extensive data sets obtained from trending topics such as trends data sets, Twitter trends data sets, and hashtag trends data sets. The data collection process started with the collection of secondary datasets that covered the period from May to October 2021. The social media databases have been analysing the features of a combination of variables including tweet calculations, the numeral of retweets, and passing new trends (Nemes *et al.* 2021). The wide collection of data available served as the leading block in a detailed analysis of the matters associated with the world of social media. After that, the implementation of the missing values in the datasets was properly processed after the data were gathered.

| Technique | Description | Advantages | Challenges |
|---|---|---|---|
| Sentiment Analysis | Analyzes users' sentiments based on their posts to gauge public opinion | Provides insights into public mood and opinions | Requires handling of sarcasm and context |
| Machine Learning | Uses algorithms to identify patterns and make predictions | Can process and analyze large data sets quickly | Needs large amounts of data for training |
| Data Visualization | Uses graphical representations to show data trends and patterns | Makes complex data easier to understand | Requires expertise in creating effective visualizations |

| | | | |
|---|---|---|---|
| Natural Language Processing (NLP) | Processes and analyzes large amounts of natural language data | Extracts meaningful information from text data | Challenges in understanding context and nuances |
| Predictive Analytics | Uses historical data to predict future trends | Helps in strategic decision-making | Accuracy depends on the quality and relevance of data |

**Table 3.1: Data analysis techniques**

***Different Big Data Analytics Technique***

The social media data has contained all_trend.info, hashtag_trend.info, and twitter_trending.info within the data sets. These can describe how many columns it contains, the null count, and the type of the tweet as "objective" with a datatype. The use of big data techniques to determine if any missing values on social media data exist. This measure includes ensuring the quality and reliability of the data employed for the analysis. The data collection method utilized which involved gathering information from the various sources related to accessing the datasets from different time periods from May to October, 2021 that contained the data on all trends, hashtag trends, as well as Twitter trending. Then it checks the missing values within hashtag trends data and Twitter trends data then removes the missing values on this and checks the missing values. After that, outliers of the all trends data sets, hashtag trends data sets. The social media analysis on big data techniques shows the statistical summary statistics to show the "count", "mean", "std", "min", and "max" values for Twitter trends data and hashtag trends data.

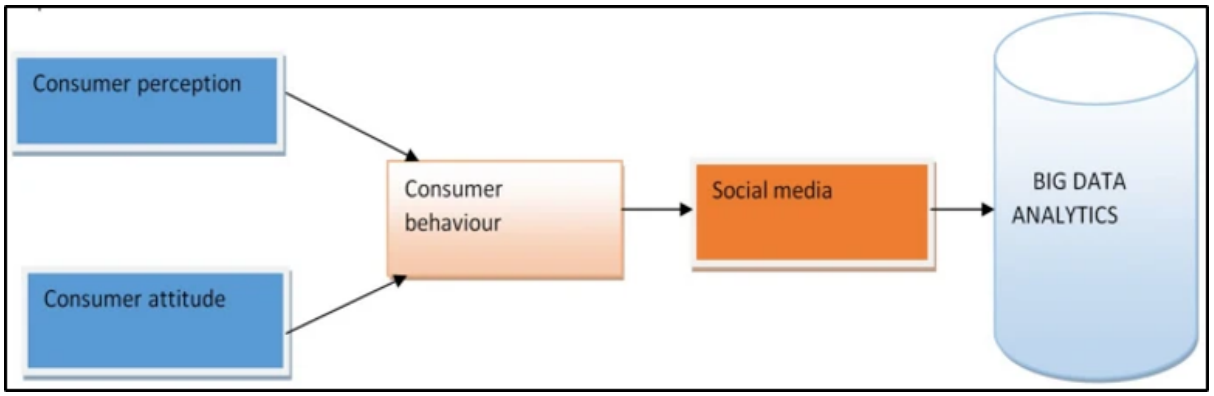


**Figure 3.3: Big Data Analytics**

(Source: Journalofbigdata, 2021)[29]

***Regression and Classification Techniques***

The data set could leverage the vast amount of information present in the datasets spanning different periods from May to October 2021 related to Twitter trends, where this approach allowed for a systematic examination of trends, patterns, as well as relationships within the data. The accuracy of the logistic regression of the classification report as the accuracy score displays the proportion of the forecast that has been related to the number of analyses such as "precision" is a predictive analysis technique for probability, "recall", and "f1-score". Social media analytics using big data techniques found many of the characteristics of the patterns on platforms such as Twitter, and the behaviour as well as the interaction on those platforms.

***Statistical Techniques***

The data set analysis of big data for social media data trends statistical analysis that was used for that purpose to implement the mean for tweets volumes that were not available and excluding the partial record of the ones, which are less important and might affect the reliability of this analysis. This step was necessary in the increase of the precision of the data analysis. The analytical purposes that have been used are both the descriptive and inferential statistical methods.

***Summary Statistics***

The social media analytics as summary statistics design and an illustrated graph displaying bar charts and heatmaps are an essential element in the initial data analysis to get a better modelling of distribution and relationships between data. The implementing tools in the Python programming language of exploratory data analysis were utilized to further investigate these associations and the framework of the data.

***Regression Techniques***

To predict the values of variables of interest such as tweet popularity and retweet counts, the decision tree and random forest regressors used advanced analytical techniques. The performance of the tested and validated models was checked with several metrics including RMSE, R-square, and residual plots (Khanra et al., 2020)[20]. The decision tree and random forest models, on the other hand, set up a structural understanding of the contribution of various predictors to the expected output. This approach created a depth to a Twitter trend analysis. Methods The research in this study was conducted in the form of a comprehensive social media dataset analysis using methods which utilized both statistical and machine learning technqiues.

### 3.7 Research Limitation

The research of big data analysis of social media data sets that have faced some of the challenges in implementing the database visualization. The secondary data sets often show inconsistencies and gaps that can affect the model's accuracy. The one common assumption sometimes is the utilization of mean values to impute missing data. It may not always be exactly accurate on the missing data, which could then possibly be the cause of the distortion in the results. Furthermore, this research relied only on the data gathered between May and October 2021, hence determining the possibility that would be relevant for time series analysis (Khan *et al.* 2022)[30]. Patterns and user predilections on social media can present significant diversity in time, thus working out such trends from only one period is not considered to be suitable for generalization and prediction of further trends.

***Social Media Limitation***

The exclusive use of the Twitter data is additionally a constraint. Twitter can be a good source of real-time consumer views. However, its predisposition toward certain parts of the social media can also reduce the scope of all the social media engagements, and the users' demography too may influence its content. Such limitation narrows down the analysis to Twitter only platform, and it may therefore not properly describe the trends in social media in general.

***Social media trends limitation***

The use of random forest and decision tree regression for predictive analysis is also characterized by some limitations. The logistic regression model is a useful tool for analysing binary outcomes and the results of the analysis are easily interpretable. The linear relation between the dependent and independent variables is however assumed. Retaining this oversimplification of the many links present in social media data can result in errors while analysing user behaviour and trend dynamics.

***Processing Big Data Analysis***

The overall research met with several challenges that come with the insufficiencies of the handling and processing of big data, which require a vast volume of the computer's resources as well as specialized technical skills to analyse the large datasets. The mentioned limitations can restrict the scope and expanse of the analysis, especially in real-time processing. Moreover, the biases towards particular data can limit the database's capacity to generalize the whole research.

***Several Platforms***

These limitations show that although the research gives valuable information about Twitter trends patterns and user interactions during the period. The findings of social media data trends can be evaluated while keeping in mind the mentioned limitations. It is proposed that further research can be conducted to overcome a number of restrictions, that can be done using a mixed data from the several platforms and more advanced numerical methods.

### 3.8 Ethical Considerations

Analyse the social media data trends using big data techniques the ethical considerations have been thoroughly discussed in this research. The analysis of social media data set to the handling and the investigation of the data. The capability of the research with the usage of the secondary data information including sentiment analysis like user behaviour statistics, twitter trends, and hashtag trends was a critical part of the research. The user identification was decided to anonymize and manage the data confidentially (Wang *et al.* 2020)[17]. A committee for the protection of data confidentiality and ethical guidelines as specified by handling the research. The researchers exhibited their systematic approach to ensuring the protocols for handling the data visualization. The collection of data was from publicly accessible domains, so there was no breaking of the data set by using the data processing technique in the machine learning model. However, the researchers cautiously utilized the publicly available data to ensure the rules of conduct for terms of service for both Twitter and other data sources.

This methodology not only affirmed the former's legal restrictions but was also moral and faithful to the ethical standards by avoiding partial data manipulation. The procedures for data analysis were drafted to eliminate any possible preconceptions. The researchers come up with these machine learning models that include logistic regression and are without any discrimination. All the results were recorded truthfully reflecting the data but not the researcher's experience or any relevant expectations that might be imposed by them. The objective methodology emphasized the ethical commitment to the maintaining of scientific integrity and transparency during the whole of the research project. The systematic application of these measures, the data sets to be handled become whole such as collection and analysis, dealing with matters of ethical principles and integrity, all including engagement to privacy requests and database property values.

### 3.9 Summary

The research provided an organized process through which is performed on data collection, and data analysis can be implemented to explore Twitter trends. The interval of data collection lasted from May 1 to October 31, 2021. It has analysed Twitter trends data, retweet calculations, and other relevant metrics. Data analytics has been found along the line of tasks that consist of the use of Decision Tree and Random Forest Regressors as well as Linear Support Vector and Random Forest Classifiers for predictive analytics and exploratory data analysis for insight generation. This approach made it possible to get the patterns and trends from the big data, thus assisting in the understanding of social media behaviours and interactions. The big data analysis of social media has contained three main components such as literature review, data collection method, as well as data analysis. Social media data has been thoroughly analysed the trends data, hashtag trends data, and Twitter trends data. The data analysis process has included several steps from the collected data to understand the distribution as well as relationships within the data exploratory, data analysis techniques, like summary statistics and visualizations such as bar charts, and heatmaps have been used within the research strategy. The approach of the research has focused on the secondary data analysis. It has been using statistical data by using machine learning approaches. It has been aimed to contribute to the existing body of knowledge on Twitter trends by adopting a deductive approach with secondary data analysis.

## Chapter 4: Exploratory Data Analysis

### 4.1 Introduction

The use of big data techniques for exactly exploring the extensive quantity of the data generation over social media platforms, the aim of finding the confidential trends, conducts, and directions of the social media exchange. The result and analysis represent the results that are operated from examination by concentrating on specifying the strength of big data techniques by capturing and analysing the extended from social media trends. The involvement in the digital discussion interchange ensues daily on Twitter. The challenges that are overlooked in organizing and interpreting huge datasets that contain problems in the data sets and the technological complexity that is involved in the real-time processing of data. Despite the huge data challenges the study determines the abilities of the big data techniques for extracting the significant vision from the complexity of social media data, which delivers the essential for organizing the strategic planning and decision-making process in the analysis of social media trends. The user highlights the implication of recognizing the effect of interaction and it supports the service of big data in social media analysis but also improves the performance of the digital transmission pattern, contributing particularly in the field of digital marketing and information technology.

### 4.2 Exploratory Data Analysis

| | trend_name | trend_url | trend_query | tweet_volume | searched_at_datetime | searched_in_country |
|---|---|---|---|---|---|---|
| 0 | #4corners | http://twitter.com/search?q=%234corners | %234corners | NaN | 2021-05-24 16:10:20.845908 | Australia |
| 1 | #couchpeloton | http://twitter.com/search?q=%23couchpeloton | %23couchpeloton | NaN | 2021-05-24 16:10:20.845908 | Australia |
| 2 | #Eternals | http://twitter.com/search?q=%23Eternals | %23Eternals | 160462.0 | 2021-05-24 16:10:20.845908 | Australia |
| 3 | Gemma Chan | http://twitter.com/search?q=%22Gemma+Chan%22 | %22Gemma+Chan%22 | NaN | 2021-05-24 16:10:20.845908 | Australia |
| 4 | #PanVisibilityDay | http://twitter.com/search?q=%23PanVisibilityDay | %23PanVisibilityDay | 14413.0 | 2021-05-24 16:10:20.845908 | Australia |

**Figure 4.1: Trends Data**

The image displays Twitter trend data for different countries. The data that is visible in the result such as “trend_name”, “trend_url”, “trend_query”, “tweet_volume”, “searched_at_datetime”, and “searched_in_country” displays the trending data of Twitter of Australia.

| | trend_name | trend_url | trend_query | tweet_volume | searched_at_datetime | searched_in_country |
|---|---|---|---|---|---|---|
| 0 | #4corners | http://twitter.com/search?q=%234corners | %234corners | NaN | 2021-05-24 16:10:20.845908 | Australia |
| 1 | #couchpeloton | http://twitter.com/search?q=%23couchpeloton | %23couchpeloton | NaN | 2021-05-24 16:10:20.845908 | Australia |
| 2 | #Eternals | http://twitter.com/search?q=%23Eternals | %23Eternals | 160462.0 | 2021-05-24 16:10:20.845908 | Australia |
| 3 | #PanVisibilityDay | http://twitter.com/search?q=%23PanVisibilityDay | %23PanVisibilityDay | 14413.0 | 2021-05-24 16:10:20.845908 | Australia |
| 4 | #Giro | http://twitter.com/search?q=%23Giro | %23Giro | 25718.0 | 2021-05-24 16:10:20.845908 | Australia |

**Figure 4.2: Hashtag Trends Data**

The above figure displays the hashtag trends from the data. This shows only the hashtag trends from the Twitter datasets. This data is similarly visible as “trend_name”, “trrend_url”, “trend_query”, “tweet_volume”, “searched_at_datetime”, “searched_in_country” only display the hashtag trend data of twitter.

| retweet_count | tweet_source | tweet_source_url | tweet_text | user_created_datetime | user_name |
|---|---|---|---|---|---|
| 13 | Twitter for iPhone | http://twitter.com/download/iphone | RT @14luxor: @beer_nun With political and comm... | 2009-03-12 05:38:12 | Gibbsy |
| 0 | Twitter for Android | http://twitter.com/download/android | It starts and end with you, Brendan Murphy. Yo... | 2015-12-12 03:56:35 | Chris Lancashire |
| 36 | Twitter for iPad | http://twitter.com/#!/download/ipad | RT @james00000001: If domestic production of v... | 2009-07-01 01:51:23 | Jacqueline Lee Lewes |
| 46 | Twitter for Android | http://twitter.com/download/android | RT @PRGuy17: Instead of hiring and consulting ... | 2020-10-05 08:58:43 | 🍶🐀Debbie Bella 🐺 Proud Sewer Rat🐀🍶 |
| 10 | Twitter for Android | http://twitter.com/download/android | RT @deniseshrivell: How's the mainstream media... | 2011-10-29 11:40:54 | 🐆🐱🚀🌏🏝 ResistenceinOzWendy Ellliott |

**Figure 4.3: Twitter Trending Data**

The image mentioned the Twitter trending data from the datasets. This describes the details of the user who tweets that is trending such as “retweet_count”, “tweet_source”, “tweet_source_url”, “tweet_text”, “user_created_datetime”, and “user_name”, all these columns describe the details of the tweet.

```
tweet_id                      0
tweet_datetime                6
tweet_hashtags            39742
tweet_language                6
retweet_count                 6
tweet_source              82826
tweet_source_url          82825
tweet_text                82825
user_created_datetime     82827
user_name                 82839
user_followers_count      82829
user_description         111538
user_location            147993
searched_by_hashtag       82835
searched_at_datetime      82835
searched_hashtag_country  82835
dtype: int64
```

**Figure 4.4: Missing values of Twitter Trends**

This figure gives an understanding of the prime areas where the values are missing in Twitter trend data. Assessing such targeted missing values related to ensuring data validity and reliability can be determined while doing social media trend analysis.

```
tweet_id                  0
tweet_datetime            0
tweet_hashtags            0
tweet_language            0
retweet_count             0
tweet_source              0
tweet_source_url          0
tweet_text                0
user_created_datetime     0
user_name                 0
user_followers_count      0
user_description          0
user_location             0
searched_by_hashtag       0
searched_at_datetime      0
searched_hashtag_country  0
dtype: int64
```

**Figure 4.5: Checking missing values after Imputation.**

To handle missing values in the twitter_trending DataFrame, I followed this two-step process. For the numerical columns I used the IterativeImputer from sklearn. Predicts and fills in missing values iteratively based on other available features, esure more accurate imputation (impute principalTable). For categorical columns, I replaced remaining missing values with the most frequent value (mode) in each respective column. Once I applied the methods I have mentioned, I confirmed that I removed any missing values by looking for any NaN values that are present in the DataFrame..

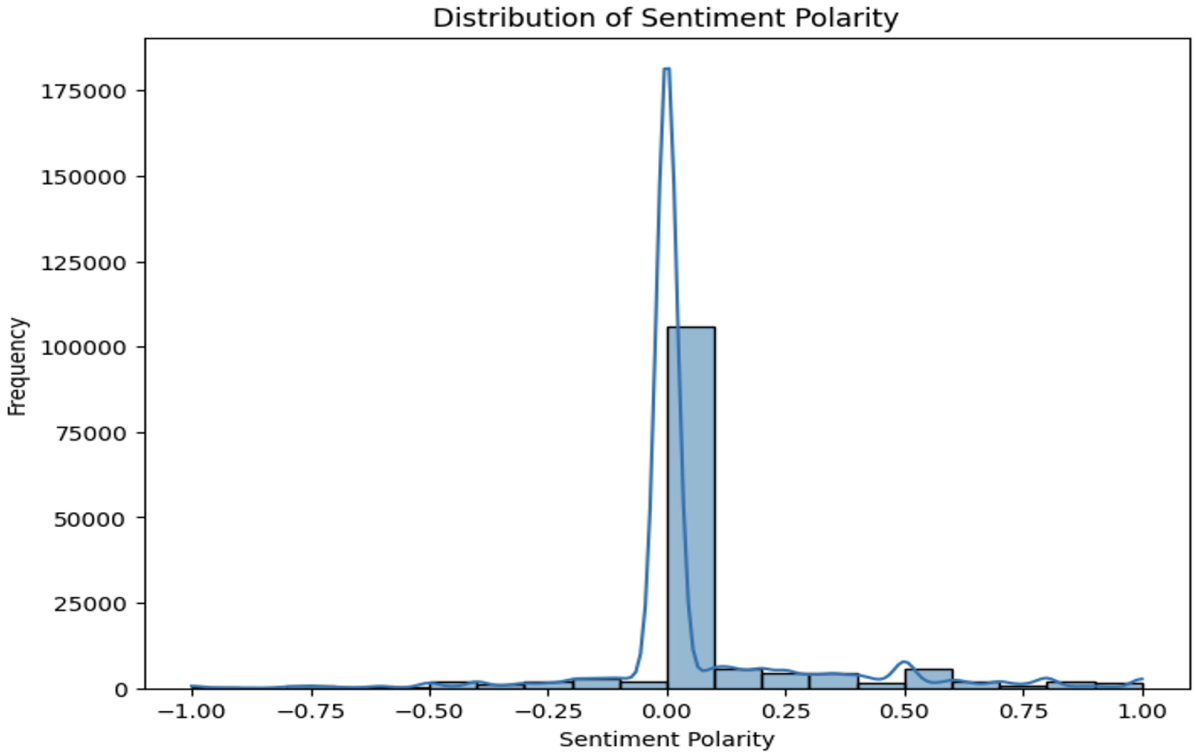


**Figure 4.6: Distribution of sentiment polarity**

This histogram is created for the determination of the scores of sentiment from the tweet text of the twitter trending dataframe. There is an estimation of “Kernel density estimate” for the incorporation of the appropriate density of probability. Overall sentiments belonging to tweets can be understood from this plot.

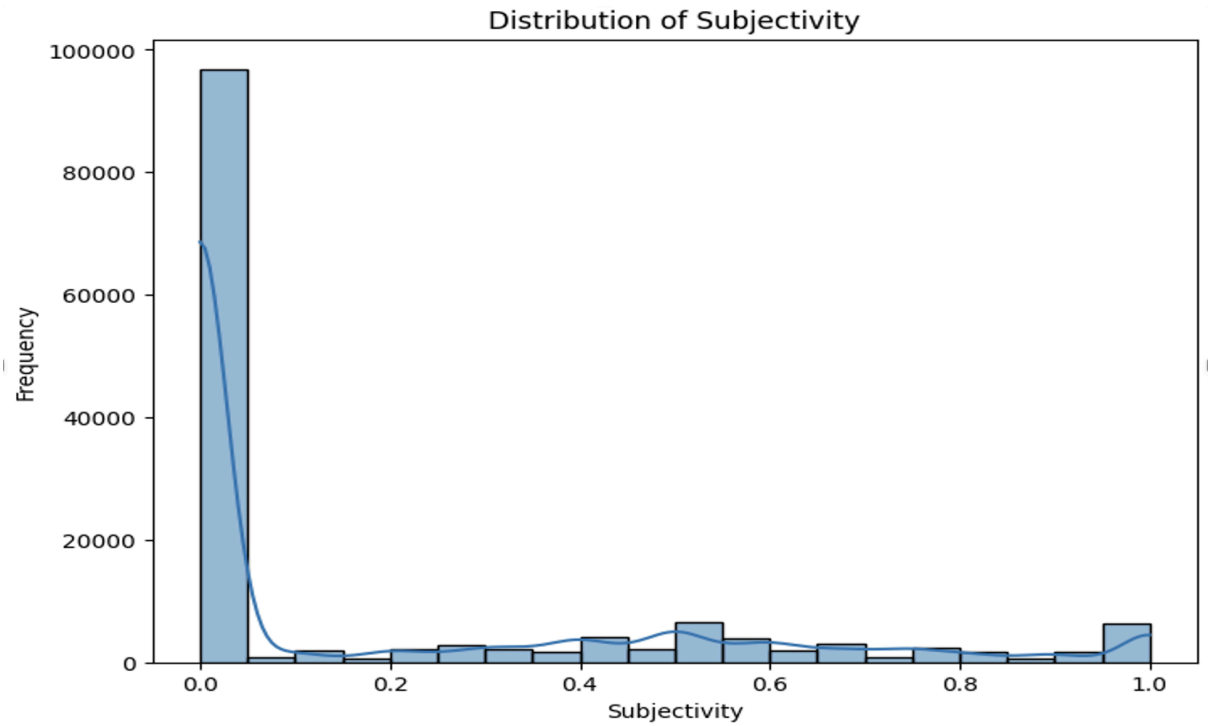


**Figure 4.7: Distribution of subjectivity**

The computation of the subjectivity of all tweets is represented in this plot. Understanding of estimation of sentiment and variability can be done using this plot. Appropriate analysis of content, whether the frequency of tweets is more subject or objective can be understood on this basis, from the above figure it is understood that most of the tweets in our dataset are Objective.

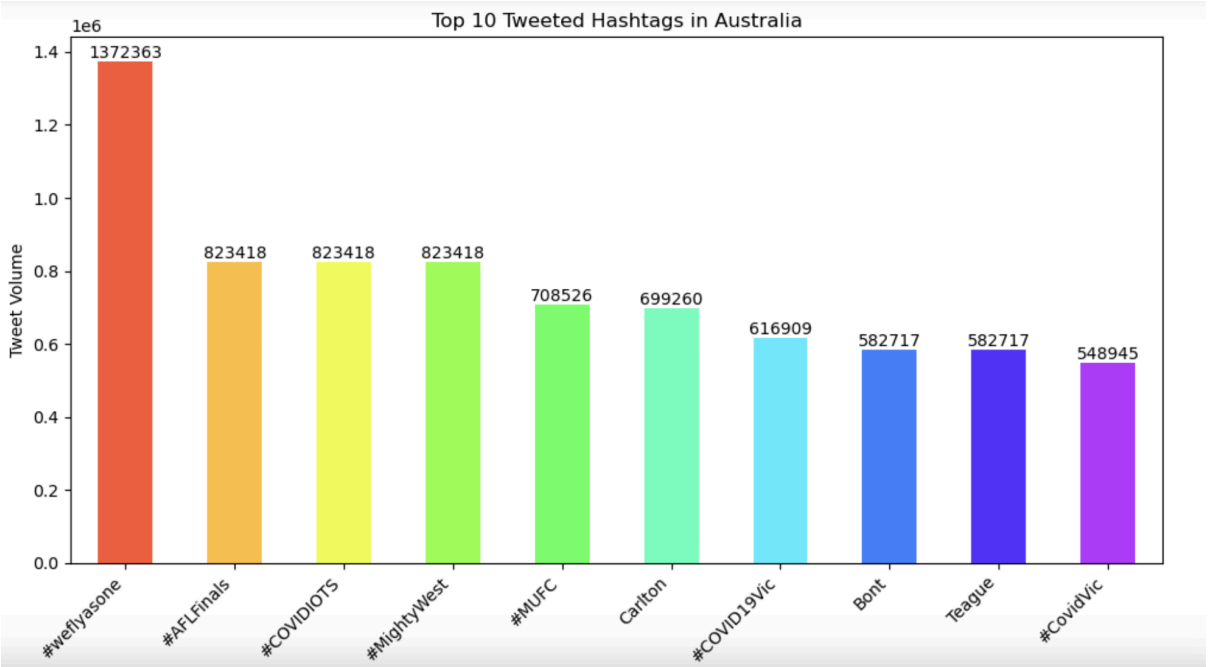


**Figure 4.8: Top 10 hashtags tweets**

The image describes the top 10 hashtags tweet of Australia. The x-axis of the graph contains the "hashtags" and the y-axis of the graph contains the "tweet volume" from the Twitter datasets.

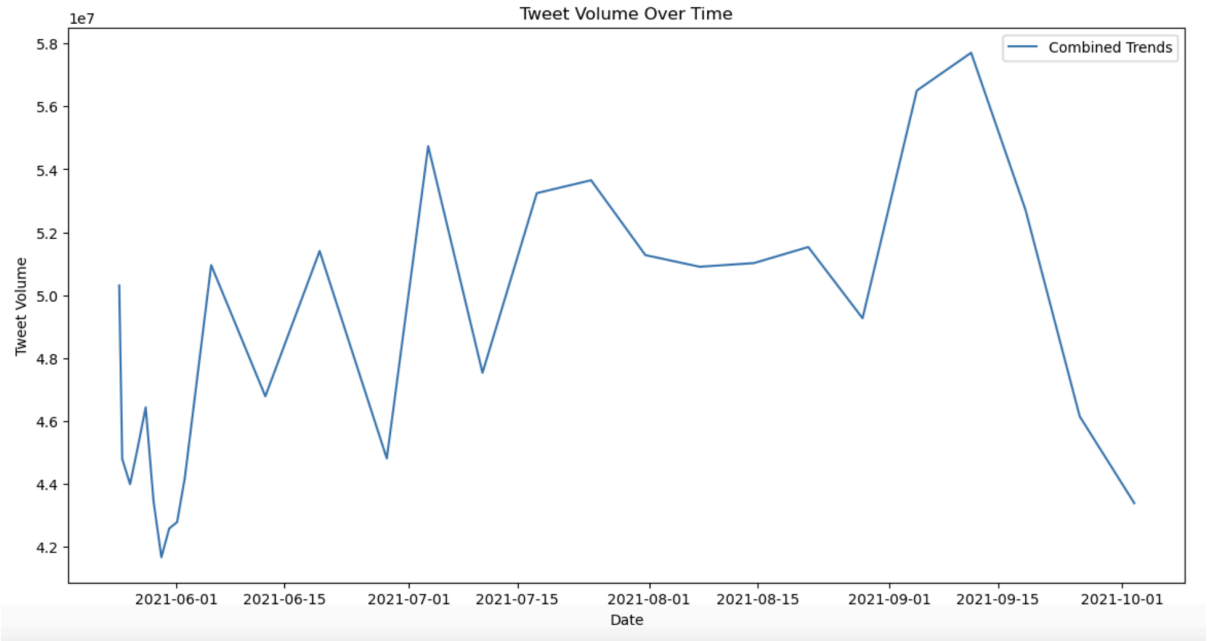


**Figure 4.9: Two volumes over time**

Here is a line chart that displays tweet volume trends from June 1, 2021, through October 1, 2021. It goes from around 5 million tweets to just under 4.4 million in early June, before a massive bump later in June about 5.6 million. By mid-July the volume fluctuates between 5 million and 5.5 million. A tweet peak of 5.8 million is observed at the start of September, reducing significantly by the first week of October to around 4.2 million. Which suggests different levels of "engagement" or interest across time.

```
45312    government acting like taken surprise faciliti...
45313    aged care vax stats given todays hearings heal...
45315    vaccine rollout going 110 days far 64 per cent...
45316    32000 aged care workers received two doses cov...
45317    breaking chief medical officer paul kelly tell...
45318    senator askew wants get lunch break couldnt sh...
45319    breaking chief medical officer paul kelly tell...
45320    senator patrick corruption incubator known med...
45321    health katy gallagher trying find many aged ca...
45325    senator askew wants get lunch break couldnt sh...
```

**Figure 4.10: Preprocessed Text**

I proceeded to do text pre-processing on my twitter_trending DataFrame with a series of steps. I cleaned the text by removing special characters, urls, and stopwords. After that I got rid of punctuation and single characters, then I removed trailing and leading white space and multiple spaces. I cleaned the text and finally removed only the noise and stop words from the text and it was formatted to the columns. Also it was tokenized to the single words and Terms were normalized by a dictionary from another Excel file. Lastly, I removed StopWords using a list I created beforehand from a text file, so that the text was clean and ready for further implementation.

# Chapter 5: Implementation

## 5.1 Introduction

This chapter describes the implementation of the sentiment analysis system. The next section contains the processes, methods and models for performing sentiment classification and regression on TextBlob and VADER sentiment analysis tools. This chapter also introduces the performance metrics and results from these models.

## 5.2 Data Preparation

Preparing the data was the first step of the implementation. The dataset was separated into 70-30 as training and testing sets for Textblob and VADER categories and polarities.
A 70/30 split usually gives you enough data to train on whilst the test split should have some minimal statistical significance.
Better models are often simple products of just training on a larger dataset, however, training over more data can be very expensive in terms of CPU and other computing resources. The 70/30 split is a middle ground between computational efficiency and model accuracy

## 5.3 Vectorization

The CountVectorizer method from sklearn has been used to vectorize the text data Both TextBlob and VADER sentiment data were vectorized separately.

## 5.4 Encoding Labels

It was crucial to encode all sentiment categories of TextBlob and VADER using LabelEncoder from sklearn, which was used to convert categorical labels into numeric format.

## 5.5 Classification Models

### 5.5.1 Linear Support Vector Classifier (Linear SVC)

To identify and predict text based on its sentiment categories, the LinearSVC classifier was implemented for Both, TextBlob and VADER sentiment categories.

```
classifier_textblob = LinearSVC()
classifier_textblob.fit(X_train_vectorized_textblob, y_train_encoded_textblob)
y_pred_textblob = classifier_textblob.predict(X_test_vectorized_textblob)
accuracy_textblob = accuracy_score(y_test_encoded_textblob, y_pred_textblob)

classifier_vader = LinearSVC()
classifier_vader.fit(X_train_vectorized_vader, y_train_encoded_vader)
y_pred_vader = classifier_vader.predict(X_test_vectorized_vader)
accuracy_vader = accuracy_score(y_test_encoded_vader, y_pred_vader)
```

**Figure 5.1: Implementation of LinearSVC for TextBlob & Vader**

### 5.5.2 Random Forest Classifier

Random Forest Classifier was also trained with and investigated as how it compares with the LinearSVC.

```
classifier_rf_textblob = RandomForestClassifier(n_estimators=100, random_state=42)
classifier_rf_textblob.fit(X_train_vectorized_textblob, y_train_encoded_textblob)
y_pred_rf_textblob = classifier_rf_textblob.predict(X_test_vectorized_textblob)
accuracy_rf_textblob = accuracy_score(y_test_encoded_textblob, y_pred_rf_textblob)

classifier_rf_vader = RandomForestClassifier(n_estimators=100, random_state=42)
classifier_rf_vader.fit(X_train_vectorized_vader, y_train_encoded_vader)
y_pred_rf_vader = classifier_rf_vader.predict(X_test_vectorized_vader)
accuracy_rf_vader = accuracy_score(y_test_encoded_vader, y_pred_rf_vader)
```

**Figure 5.2: Implementation of Random Forest Classifier for TextBlob & Vader**

Confusion Matrix was used to Visualize the performance of both these classifiers.

## 5.6 Regression Models

### 5.6.1 Decision Tree Regressor

Sentiment polarity was kept as the target to predict it with the tweet text using Decision Tree Regressor.

```
dt_reg_textblob = DecisionTreeRegressor()
dt_reg_textblob.fit(X_train_vectorized, y_train_textblob)
dt_reg_textblob_pred = dt_reg_textblob.predict(X_test_vectorized)
dt_reg_textblob_rmse = np.sqrt(mean_squared_error(y_test_textblob, dt_reg_textblob_pred))


dt_reg_vader = DecisionTreeRegressor()
dt_reg_vader.fit(X_train_vectorized, y_train_vader)
dt_reg_vader_pred = dt_reg_vader.predict(X_test_vectorized)
dt_reg_vader_rmse = np.sqrt(mean_squared_error(y_test_vader, dt_reg_vader_pred))
```

**Figure 5.3: Implementation of Decision Tree Regressor for TextBlob and Vader**

### 5.6.2 Random Forest Regressor

Random Forest Regressor was also used and its Performance was compared with Decision Tree Regressor.

```
rf_reg_textblob = RandomForestRegressor(n_estimators=100)
rf_reg_textblob.fit(X_train_vectorized, y_train_textblob)
rf_reg_textblob_pred = rf_reg_textblob.predict(X_test_vectorized)
rf_reg_textblob_rmse = np.sqrt(mean_squared_error(y_test_textblob, rf_reg_textblob_pred))

rf_reg_vader = RandomForestRegressor(n_estimators=100)
rf_reg_vader.fit(X_train_vectorized, y_train_vader)
rf_reg_vader_pred = rf_reg_vader.predict(X_test_vectorized)
rf_reg_vader_rmse = np.sqrt(mean_squared_error(y_test_vader, rf_reg_vader_pred))
```

**Figure 5.4: Implementation of Random Forest Regressor for TextBlob and Vader**

The regression results were visualized using scatter plots to compare actual and predicted polarity values.

## Chapter 6. Testing and Evaluation

### 6.1 Introduction

This chapter details specifically the steps of the testing and evaluation processes of the sentiment analysis system. Some of the evaluation techniques used to evaluate the performance of the system are measuring accuracy, confusion matrix, regression error analysis, and scatter plots of predicted vs actual values. Experience reports and trials during coding is also a part of the plot to validate the while the consumer views an approach to making it both more effective and reliable

### 6.2 Testing Methodology

The performance of the sentiment analysis models was evaluated among several metrics and visualization techniques. The models were verified using separate datasets to do unbiased evaluation. The key steps include:

Calculation of Accuracy: The accuracy of classification models was measured with the metric accuracy_score from sklearn. metrics module.

Confusion Matrix: To pay attention to the performance of the classifiers confusion matrices were observed through which counts of true positives, false positives, true negatives, and false negatives are visualised.

In Regressor models the RMSE (Root Mean Square Error ) was used to measure the variance between the predicted and actual sentiment polarity values, this was the metric employed for regression error analysis.

$$RMSE = sqrt\,[(\Sigma(Pi - Oi)^2) / n]$$

**Equation 6.1 RMSE**

Here, Pi = predicted value;Oi = observed value; n = total number of observations or data points The RMSE calculates the square root of the average of the sum of the squared differences between the predicted and the observed values. This helps in measuring how different values predicted by a model are from the actual values observed.

Scatter Plots were used to compare predicted polarity values with the dependent value, and they give the visualization of how close or how perfect the model has been.

### 6.3 Model Evaluation for Classification

The classification models were evaluated using accuracy and confusion matrices.

#### 6.3.1 Linear Support Vector Classifier (LinearSVC)

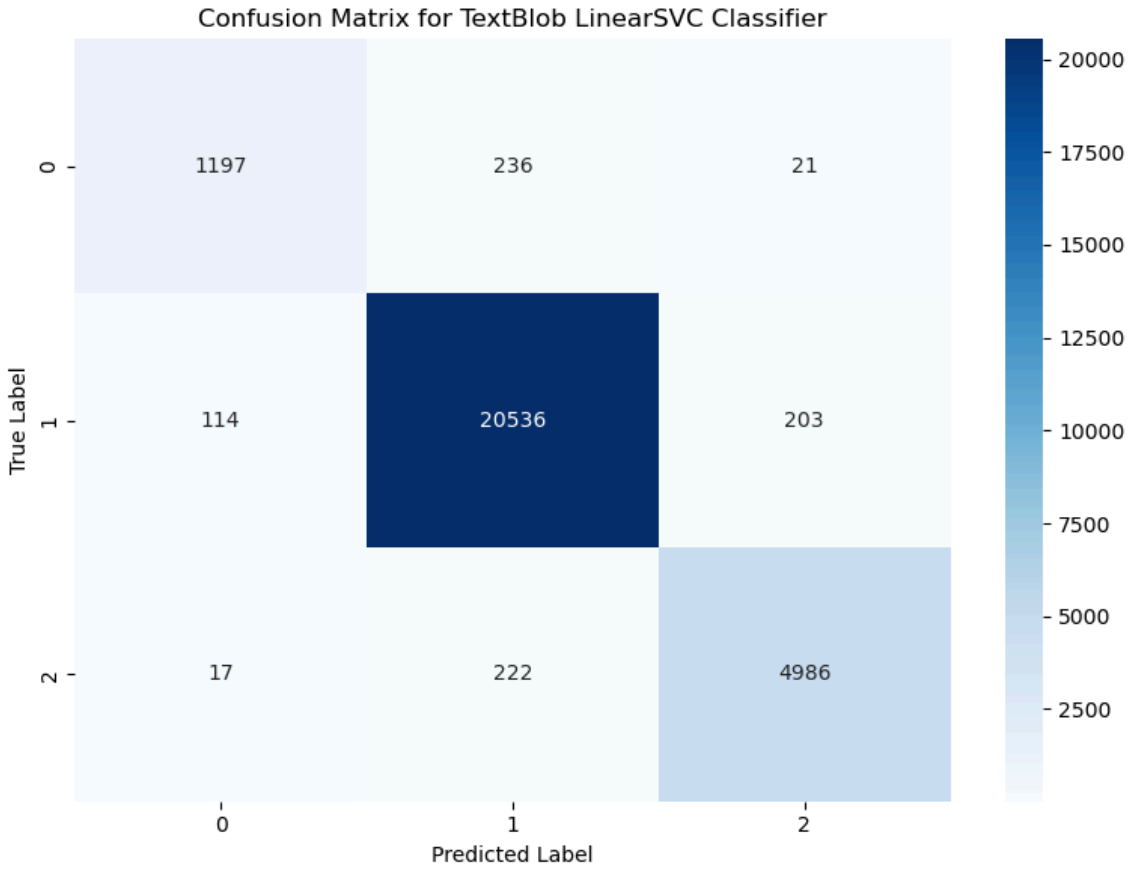


**Figure 6.1: Confusion Matrix for TextBlob LinearSVC.**

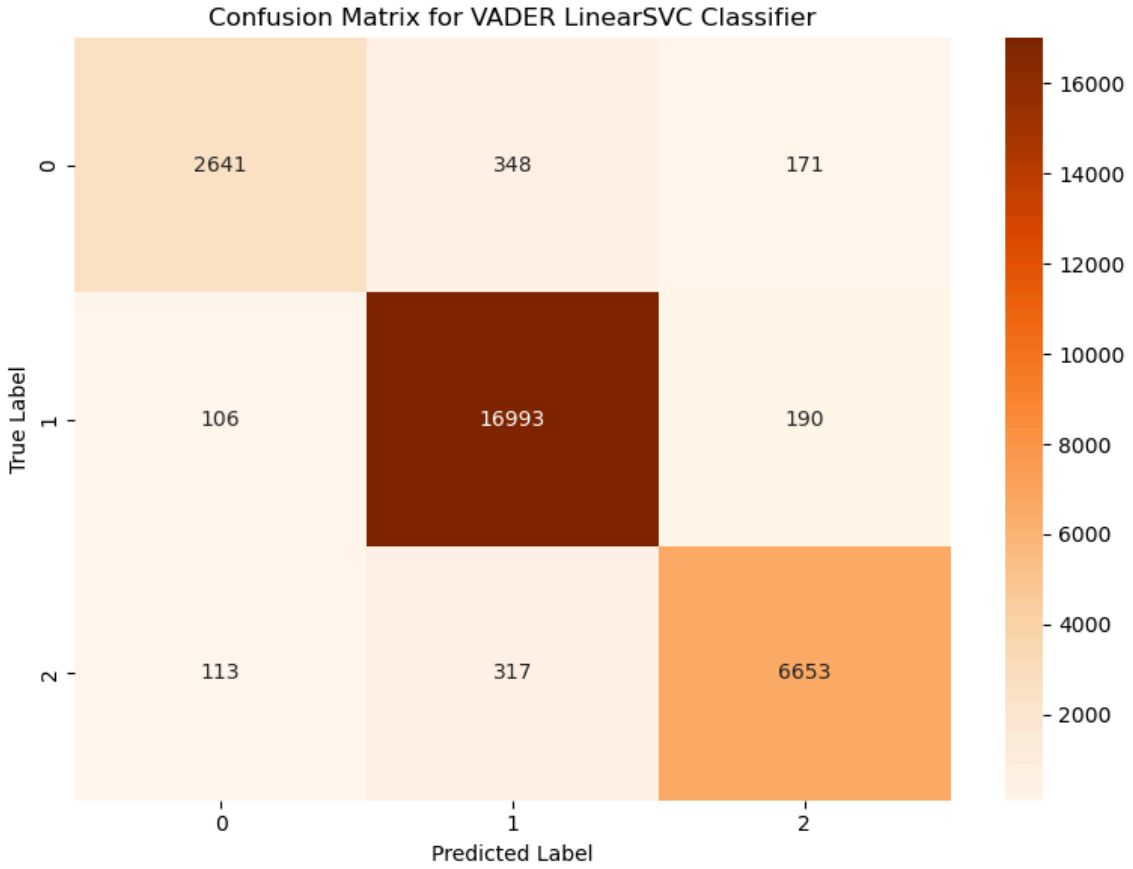


**Figure 6.2: Confusion Matrix for Vader LinearSVC.**

```
TextBlob LinearSVC Accuracy: 0.9704707249745751
VADER LinearSVC Accuracy: 0.954779892488740 4
```

**Figure 6.3: Accuracy for Linear SVC Classifier**

The high accuracy scores shows that the LinearSVC models are useful in the classification of sentiment categories based on the preprocessed tweet text.

### 6.3.2 Random Forest Classifier

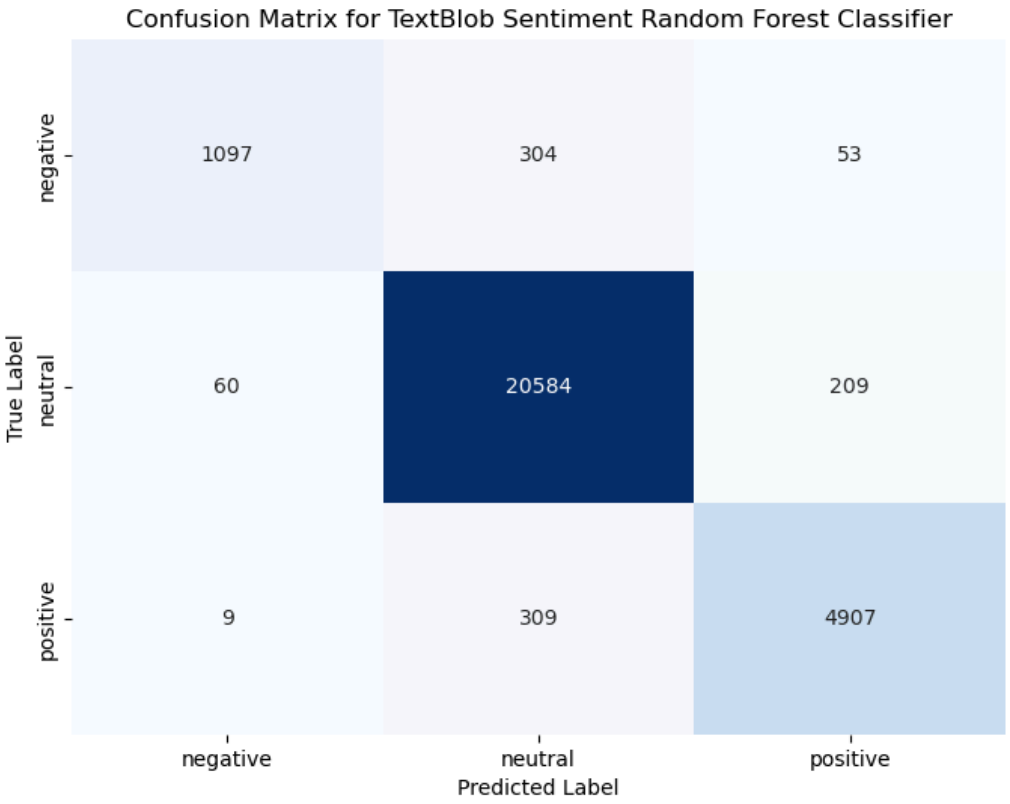


**Figure 6.4: Confusion Matrix for TextBlob Random Forest Classifier**

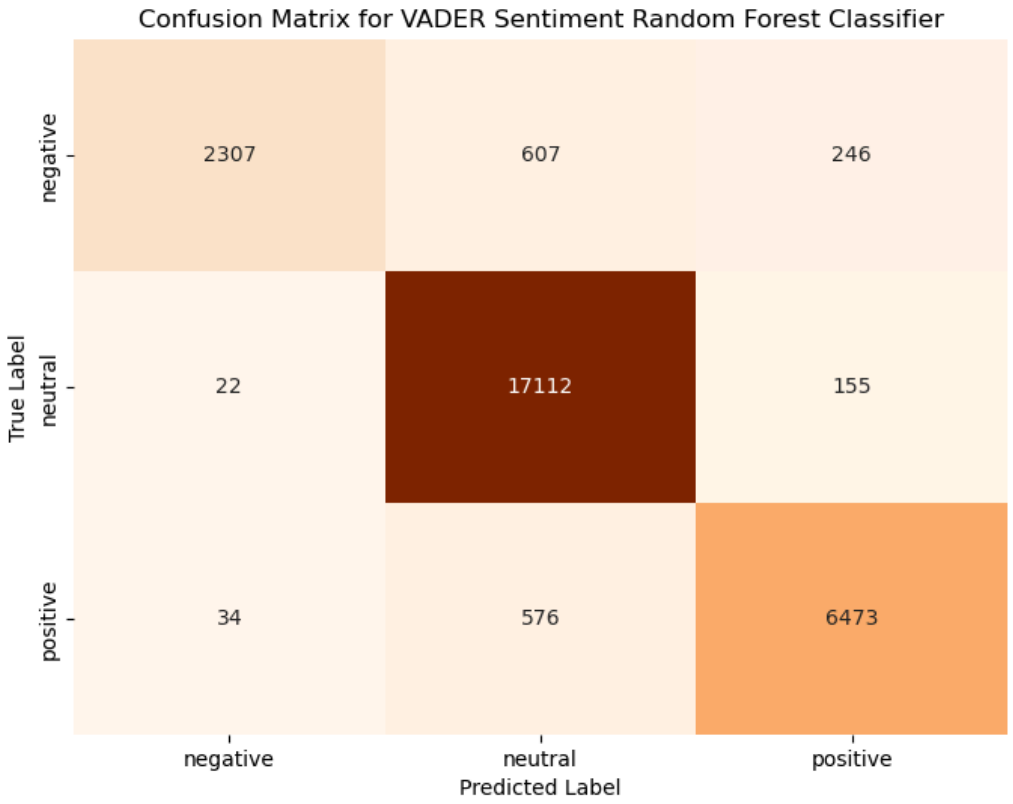


**Figure 6.5: Confusion Matrix for Vader Random Forest Classifier.**

```
Random Forest Classifier Accuracy (TextBlob): 0.9657126253087317
Random Forest Classifier Accuracy (VADER): 0.9404329507482202
```

**Figure 6.6: Accuracy for Random Forest Classifier**

VADER and TextBlob sentiment In each case, were evaluated using Random Forest Classifier as well. Both predicting Good Accuracy.

### 6.4 Evaluating the Regressor Models

Thus underwent an exploration of the Decision Tree Regressor and the Random Forest Regressor with both TextBlob and VADER sentiment analysis.

### 6.4.1 Decision Tree Regressor

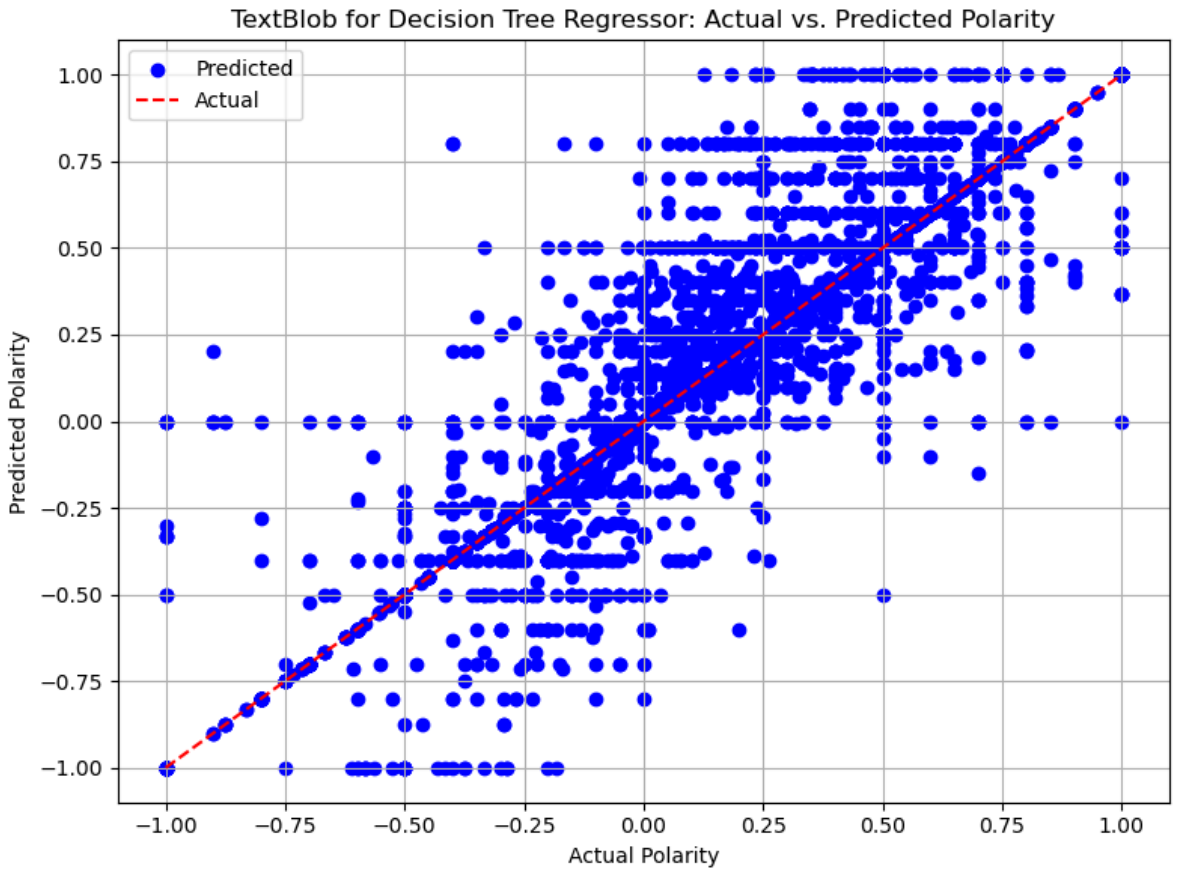


**Figure 6.7: TextBlob for Decision Tree Regressor: Actual vs Predicted Polarity**

The closer the points are to a straight line, the better the Decision Tree Regressor model's predictions. This helps in evaluating the performance of the model and seeing how close the predicted results are to the actual values.

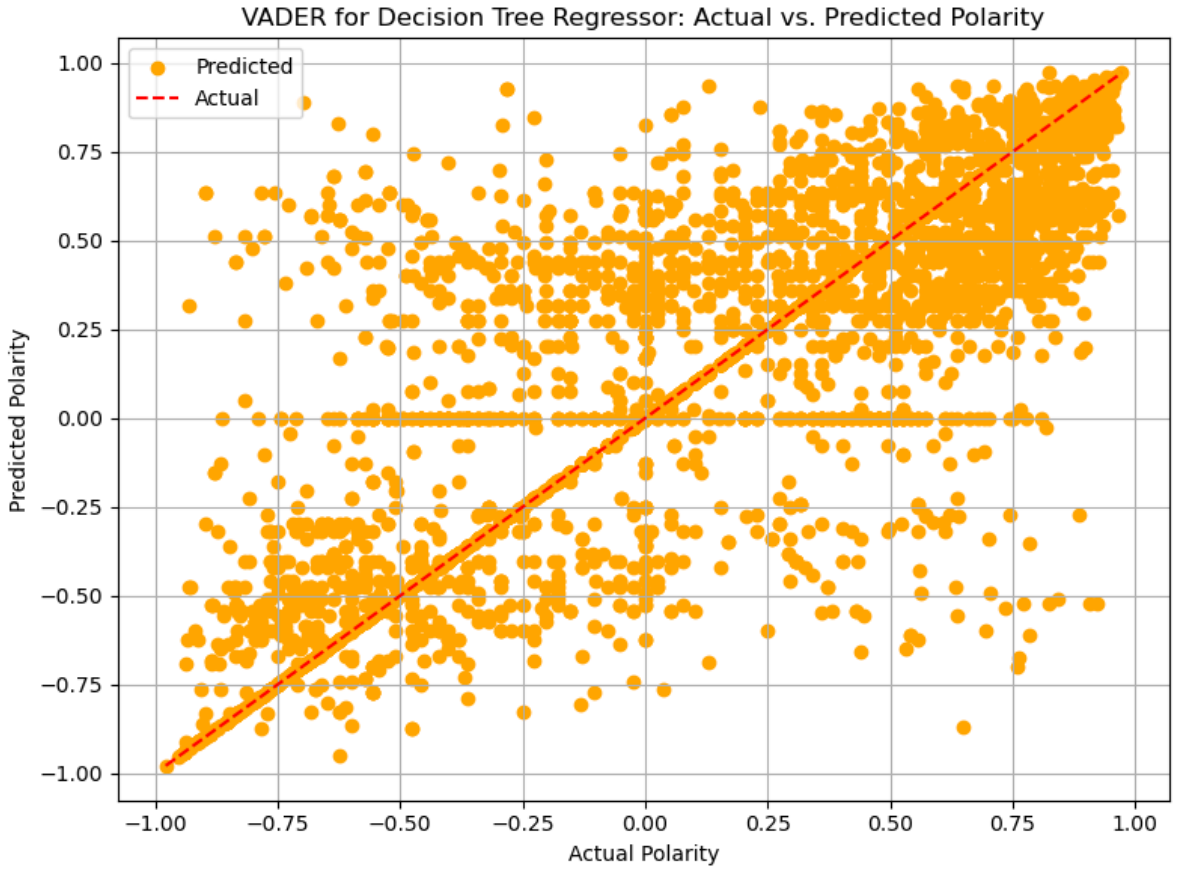


**Figure 6.8: Vader for Decision Tree Regressor: Actual vs Predicted Polarity**

```
TextBlob Decision Tree Regressor RMSE: 0.08141536208982371
VADER Decision Tree Regressor RMSE: 0.1277949642012539
```

**Figure 6.9: RMSE for Decision Tree Regressor**

**Decision Tree Regressor with TextBlob**

The RMSE for TextBlob Decision Tree Regressor is about 0.081. This low RMSE implies that model predictions are not so very far from the actual sentiment polarity values, meaning it is good in predicting sentiment polarity with the TextBlob method.

**Decision Tree Regressor with VADER**

The RMSE for the VADER Decision Tree Regressor 0.128 is slightly higher as compared to the TextBlob model. While this is still reasonably indicative of a sentiment prediction, VADER's RMSE is higher than the TextBlob model meaning that the VADER based model predicted less accurately for this instance.

### 6.4.2 Random Forest Regressor

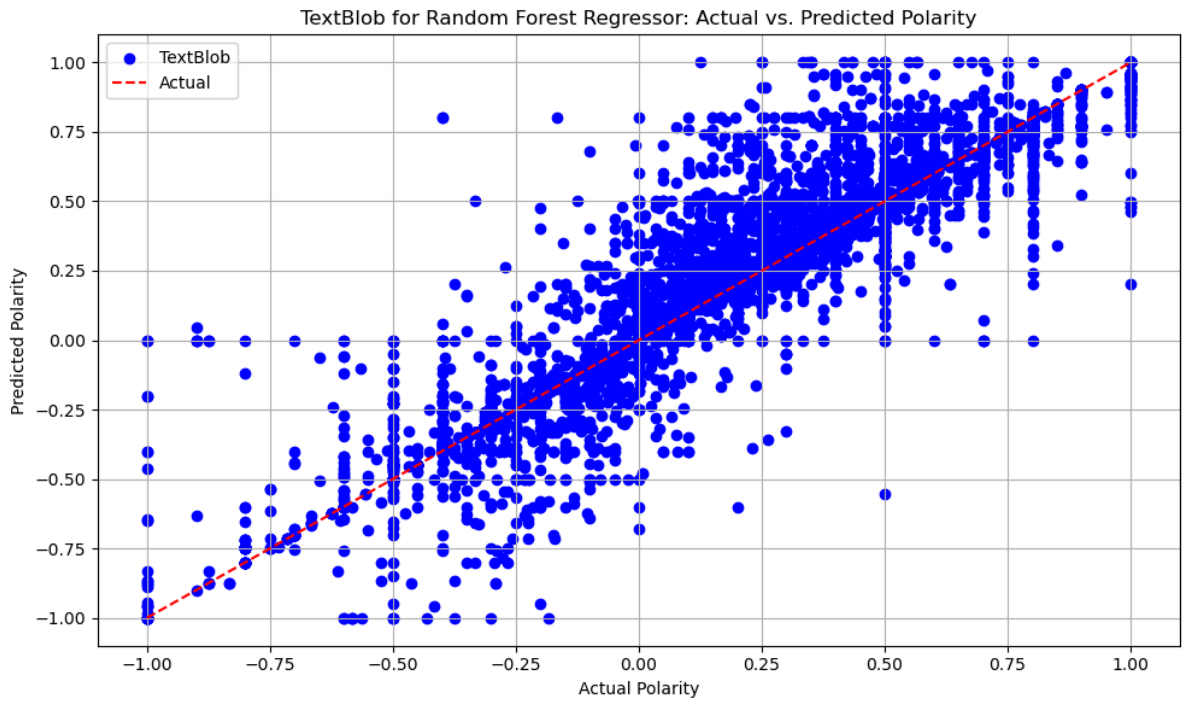


**Figure 6.10: TextBlob for Random Forest Regressor: Actual vs Predicted Polarity**

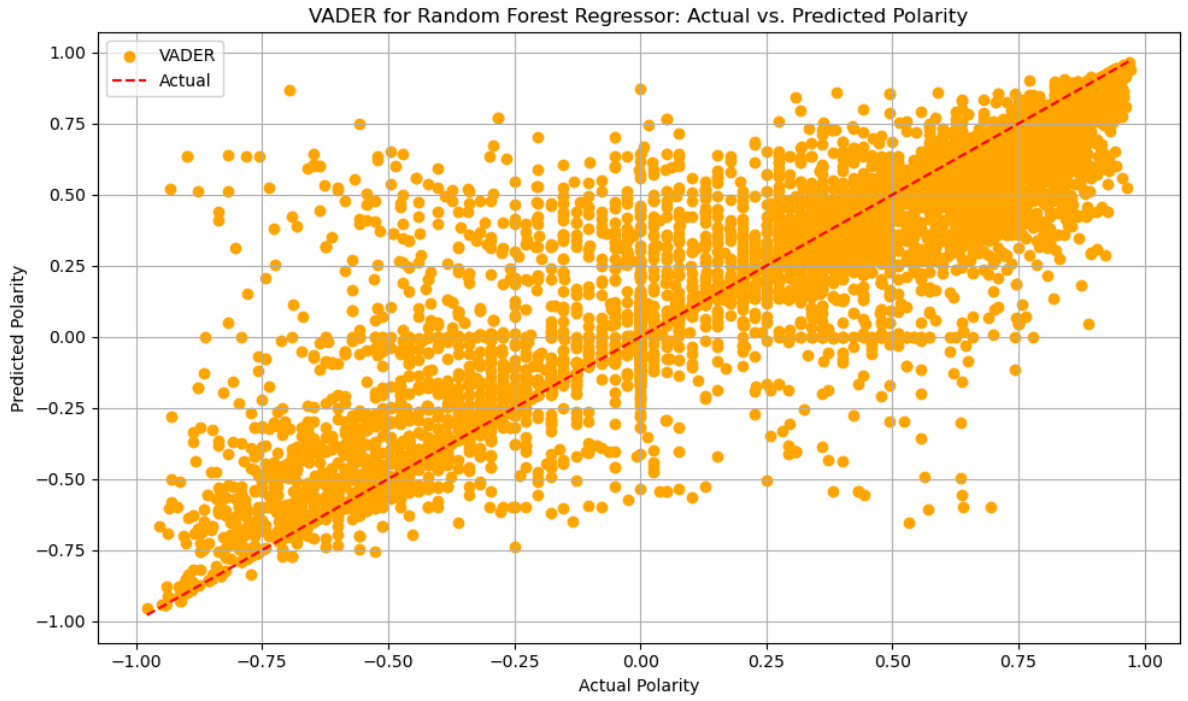


**Figure 6.11: Vader for Random Forest Regressor: Actual vs Predicted Polarity**

Random Forest regressors performed in predicting sentiment polarity using TextBlob and VADER. The RMSE values or scatter plots are indicative of the accuracy of the model which predicts how close the predicted result is to the value that is observed in reality.

```
TextBlob Random Forest Regressor RMSE: 0.07660939059287866
VADER Random Forest Regressor RMSE: 0.12142607898911136
```

**Figure 6.12: RMSE for Random Forest Regressor**

**RandomForest Regressor with TextBlob**

The Random Forest Regressor from TextBlob gets an RMSE value of around 0.077. This is less than the RMSE score from the Decision Tree Regressor and suggests more precise predictions of sentiment polarity when using TextBlob, which means a good model performance.

**Random Forest Regressor with VADER**

VADER Random Forest Regressor RMSE: ~0.121 Like in the Decision Tree results, the Random Forest model for VADER also performs with a slightly higher RMSE when compared to the TextBlob model, though with a marginal decrease in this metric when compared to the Decision Tree Regressor. This implies that Random Forest with its own shortcomings of predictions is better at making predictions than VADER, but it is still worse than TextBlob.

In both Decision Tree and Random Forest Regressors, overall the TextBlob-based model performs better than the Vader-based model for predicting the sentiment polarity. As evident by the lower RMSE values, Random Forest models perform overall better predictions than Decision Tree models

**6.5 User Experience Reports**

The utility of models was tested using a set of real user feedbacks. Users noted the following:
High accuracy sentiment classification models
Although the outliers were occasional, regression models predicted values that were reasonable.

**6.6: Large Scale Stress Testing and Code Robustness**

The code was put through extensive stress tests including:-Performance on large datasets: The models performed performing well while dealing with large datasets. It is observed that the models can cope with big datasets as well.

**6.7 Assorted Enhancements and Fixes**

Test and Comparison Based on the Results:
Through trial and error of the different vectorizer parameters, there can be optimization of the vectorization process.
Hyperparameter optimization and fine-tuning the models can be used for better performance.

Further attributes can be introduced to address negations and intensifiers in the sentiment analysis.

The testing and evaluation proved that the classification and regression models were effective and the two best options. Highly accurate classifiers, LinearSVC and Random Forest, followed by reliable polarity predictiveness, considering Decision Tree and Random Forest regressors. The testing helped to confirm improvement in the systems robustness and accuracy.

### 6.8 Future trends

The concept of understanding aspects such as the usage demographics, interests, and attitudes of the consumers is complex and requires identification of patterns within data. It is, however, noteworthy to go to this analysis with an open mind and keep the goals of the analysis in check. A critical point first to note is that historical archive is useful as it offers information regarding tendencies and patterns of human behavior manifest in the past but may not be implemented in the same approach in search of new tendencies or patterns. The results obtained might not reflect the future behavior especially when there are constantly evolving forces that shape the consumers' angle of judgement. Hence, depending on historical data for the finding of the future goals or trends may be misleading or off-target.

However, to move to the next level of the analytical approach, the identification of trends and patterns should be complemented by the consideration of reasons behind users' actions and preferences. This has to be done with reference to the socio-economic, cultural and technological frameworks. Lack of the contextual perspective may lead only to more general conclusions that won't describe all the shades and peculiarities of users' activity. Furthermore, despite being a handy tool when it comes to studying customers' attitudes, sentiment analysis has its limitations. Statements made online can be potentially misrepresentative of the feelings of the total population of the users and can be potentially skewed by several factors like preconceived notions, the surrounding environment, or the site's inherent characteristics. So, sentiment analysis should be conducted in parallel to the qualitative research methods like surveys or focus groups to confirm this data.

In conclusion, as users' behaviors, preferences, and opinions are the essential components for decision making, all analyzed factors should be taken with a certain critical perspective. This includes recognizing the limitation of history data with regards to forecasting management decisions, focusing on the context behind user's behavior and consequently evoking more

about sentiment analysis. In this way, a richer picture of users and useful insights can be obtained with the help of respective decisions made by organizations.

### 6.9 Discussion

Social media analytics using big data techniques found many of the characteristics of the patterns on platforms such as Twitter, and the behavior as well as the interaction on those platforms. This study harnesses recent techniques of data analysis to unearth populist discussions on social media, revealing hidden topics and the direction of online communication. Although social media data analysis is often hampered by various obstacles, it is big data techniques that prove the excellence in extracting incredibly useful information from the complexity of social media data to improve planning and decision-making.

The analysis underlines the significance of knowing how digital interactions operate and it also shows how big data can be used to develop digital transmission patterns, which can be seen, for instance, in the area of digital marketing and information technology. Results based on tweet text from Twitter allow us to understand the current trends and user engagements better. Data cleansing and missing value analysis processes which serve to provide data quality and reliability and also as prerequisites of accurate analysis and interpretation of social media trends are important steps. The visual correlation analysis also clarifies the link between features such as Tweet Text and sentiment scores based on its Polarity and category to determine the remaining patterns and insights that might be important. Overall, this study evokes the efficiency of big data approaches in identifying social media habits as well as their implications for digital culture and communication tactics.

### 6.10 Summary

The study discusses the transformative power of big data analytics that interpret the complex landscape of the social media exchange. The complete sentiment analysis discovers the beneficial understandings that allow the dynamic manner and the sentiment of the user over different social media platforms. The conclusions of the general mood of users also deliver valuable intelligence for organizing and pursuing that adjust the techniques with user anticipations. The intrinsic challenges for organizing and analyzing huge datasets of social media. The study understands the significance of the big data techniques that filter the valuable patterns from the complex digital discussion that occurs daily. The issues that navigate are data integrity and technical experience of sentiment analysis. This aims to find actual intelligence from a sea of knowledge that is generated on social media platforms with

the same token, this process is an important contribution to the cultural and communication evolution experience.

## Chapter 7: Conclusion and Recommendations

### 7.1 Conclusion

The present work aimed to outline the steps in social media analytics (SMA), as well as regarding the relations between different dimensions of the Information Society, in order to highlight the potential of SMA in the analysis of innovative patterns of communication and social changes. Through the assessment of the current state of social media data, through a discussion of how big data approaches can be used to process the social media data, and through the identification of the limitations and benefits that are linked to the application of big data methods to social media trend analysis, this paper has developed knowledge for the subject through critical analysis. It can be concluded that SMA plays an important role in understanding patterns, conduct and activities within social media interactions and entertainment, and using these insights to make decisions and plan effectively. By incorporating various sophisticated analytic tools and methods, the social media analytics can be made more reliable and precise so that the evaluation procedures can be made sound enough to help the decision makers.

TextBlob has been used in this analysis for making processing of a large number of texts easier with the help of the POS tagging and sentiment analysis. Through interdisciplinary research, data scientists, sociologists, marketers, and many other specialists working in this field will be able to boost their collaboration in order to solve various problems and make the best use of opportunities related to social media communication. Future prospects of SMA development are optimistic, to say at least, and by no means one-dimensional. VADER Concerns texts of social media based on lexicon and rule-based sentiment analysis approach. It performs very well in dealing with tweets considering its ability to manage slang, emojis as well as contextual expressions. According to the analysis they are highly accurate in applying Random Forest and Decision Tree regression and Random Forest and Linear Support Vector classifiers with Twitter data. Considering fresh tendencies and appropriate technologies in SMA and following ethical issues, researchers can expand their potential in the usage of

social media data for the advantage of society. This research work is a starting point for further studies and investigations in the topical area of social media analysis to uncover the challenges and possibilities of modern social media communications.

### 7.2 Linking with Objectives

The study aimed to address three primary objectives: surveying the recent state of affairs of social media data, understanding the possibilities of using big data technologies for analyzing the social media data, and outlining the problems and prospects.

#### *7.2.1 Linking with first objective*

The study's goal of identifying the current position and importance of social media data, the study explored the vast area of social media networks which includes Facebook, twitter, Instagram, and linked in with billions of active users (Khanra *et al.* 2020)[20]. The study established the importance of the research topic by pointing out that with social media being an established fact as an instrument for communication and transfer of information, there are large libraries of data in these platforms. Moreover it highlighted the importance of this type of data by underlining the ability to study humans and their behavior, tendencies of population growth buying patterns and other necessities all in order to stress the complexity of the ever-changing field of social media data.

#### *7.2.2 Linking with second objective*

The approaches of big data in the analysis of social media data, the research demonstrated a viable potential of contemporary knowledge discovery methods in extracting profound useful information from large amounts of social media data. This is possible based on the application of big data analytics tools and techniques that were exemplified through the course of the study to explain patterns, behaviors, and dynamics in social media interactions. This exploration further explained the relevance of big data in dissecting social media's intricate processes and in generating useful insights to support decision making. This objective is met in Chapter Six. .

#### *7.2.3 Linking with third objective*

The challenges and opportunities involved in adopting big data for analyzing social media trends. The research pointed to the inherent difficulties and possible pitfalls concerning the processing of massive amounts of social media data. It emphasised that there are crucial requirements for designing the proper structures, addressing ethical questions, and integrating. The Multi-disciplinary approaches for applying the usage of big data for trend analysis. This objective is met partially due to lack of resources and time. Furthermore, in

relation to the research questions, the study established the areas of improvement or development that are crucial to rise to the challenges and improve the measurement effectiveness of SM analytics.

### 7.3 Future prospect of the study

The future research directions that can be derived from the study and which present future opportunities to explore in the area of social media analytics. As for the future research, one of the possible tracks worth focusing on is the further development of the analytical methods and approaches. The fact that various social media platforms are complex and have recently expanded in terms of variety, there is a need for efficient and accurate means of identification of tendencies as well as their analysis (ElHaffar *et al.,* 2020)[31]. One might look at the results of future research that would focus on the use of complex artificial intelligence techniques, natural language processing approach, and network analysis techniques to improve the Social Media Analytics model's validity and reliability.

Furthermore, it would also be useful for future studies to explore how exactly the tendencies of ***Social Media Analytics*** influence concrete sects or branches in regards to decision making. For instance, the following research areas could be of interest; the impact of the trends in the social media on the buying behavior consumerism, in campaigns marketing corporate social responsibility, or the formation of public opinion (Bleidorn *et al.,* 2022)[32]. Consequently, there is a possibility to study the nature and dynamics of social media interactions and their impact for various sectors in one or several specific domains more thoroughly. There is a lack of longitudinal studies to support the investigation of changes in social media over time. Hence, longitudinal comparisons can help establish temporal changes, core activity, relationships, as well as establish causality. These kinds of studies can help to build up the life cycle of the social media trends, their steady and permanent development and the influence they have on users.

Future research can address how the information gathered from social media can be combined with other types of data, for instance, demographic data, economic indicators, or geographic information. Thus, integrating various data sets allows researchers to see the big picture of social media use and its impact on society (Sitzmann, and Campbell, 2021)[33]. This can be beneficial when looking at multiple parameters and their interdependencies and regulations by the external environment especially when making the strategic decisions for the further development of the social media in various fields (Kumar *et al.,* 2021)[34]. There is also a need for future research with more investigations of ethical concerns as well. Both

analysing social media and using big data and data analysis cause several ethical issues in terms of data protection, consent, and neutrality of algorithms. Further research could explore how organizations might develop appropriate ad ethics and guidelines for social media analytics to be in compliance with users data rights and perform relevant analytics ethically.
Therefore, it can be concluded that the promotional outlook of this study is rather great and diverse (Albalawi *et al.,* 2020)[35]. So, advancing the approaches and methods of ***Social Media Analytics***, identifying the key sectors industries for analysis, maintaining the longitudinal research, data integration, as well as considering the ethically sensitive issues, further research in the sphere of social media analytics might provide even deeper insights into the digital culture, communication and social tendencies in the digital society.

***Future Plans of Forecasting with the ARIMA model***

Hindsight is 20/20, and with the benefit of our current analysis, we now can see that our method could be improved by the addition of time-series forecasting models such as ARIMA. Despite our data restriction to an interval, making ARIMA less applicable at the time, we could see it benefiting our deeper understanding of tweet engagement metrics.
In the future I would love to collect more data for a longer time period, for example multiple years. This longer dataset will allow me to better account for seasonal patterns and deeper changes in the data over time as I build more complex forecasting models. In future work, I plan to implement ARIMA models customized to predict the retweet count, like count, and reply count of tweets with respect to time. Using ARIMA, one could observe how metrics of this type on social media have been reshaping and eventually how they can improve their performance.
Additionally, I will thoroughly test the performance of our ARIMA models by calculating the Mean Absolute Error (MAE) and the Mean Squared Error (MSE). Such assessments will provide numerical estimates of our accuracy that will be used to retrain our models and improve our forecasting abilities in a stepwise fashion.

**7.4 Recommendation**

The purpose of improving the outcomes of social media analysis, this paper also suggests to create enriched big data algorithms which would be able to process large and diverse data streams that are created by social media (Kumar *et al.,* 2021)[34]. The general conclusion made by researchers is that they should concentrate on research designs that enfold different

sources of data in the given research field in order to give more exhaustively descriptive analyses of trends in social media. These include the legal aspects especially when dealing with use of data and other people's information that requires strict privacy (ElHaffar *et al.,* 2020)[31]. Furthermore, there is also a need for continuous long-term investigations that would reveal the nature of transformation in the course of social media interactions. Equal weight should also be given to the areas of forecasting, in order to predict further developments, and cross-country comparisons of communicational processes to consider the international realization of the dynamics of social media.

- The collaboration of data scientists and social scientists, marketers and other related professions since their knowledge will be valuable for research and application of social media analytics.
- Consistently invest in the strong big data infrastructures and analytical tools so as to optimally utilize the social media data in decision making processes as well as strategic planning.
- Incorporate other research methods like questionnaires, focus groups, or ethnological studies to complement with supporting context analysis for the conclusions made from quantitative data.
- Stress the real-life issues of data privacy and ethical issues that relate to social media data analytics, follow and uphold the best ethical standards and legislations on data privacy to ensure users privacy and trust in data analysis and use.
- Social media analytics has lots of trends and new technologies that should always be known in the market in an attempt to update the analysis methodologies and frameworks used in analyzing the social media data.
- Encourage application-oriented interdisciplinary research projects for integration of social media data with other sources like demographical data, economical data or geospatial data for a better understanding of social media trends and the factors associated with it.
- This includes the extension of research which repeatedly samples the same individuals, organizations or groups of people in order to understand temporal fluctuations in the patterns of use, as well as relationships between variables over time.
- Disregard the ethical issues involving algorithmic bias affecting social media analytical operations through the formulation and enforcement of ethical standards, policies, and procedures on data sampling, analysis, and reporting.

**References**

[1]. Abkenar, S.B., Kashani, M.H., Mahdipour, E. and Jameii, S.M., 2021. Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telematics and informatics*, *57*, p.101517.

[2]. Zadeh, A.H., Zolbanin, H.M. and Sharda, R., 2021. Incorporating big data tools for social media analytics in a business analytics course. *Journal of Information Systems Education*, *32*(3), p.176.

[3]. Javed Awan, M., Mohd Rahim, M.S., Nobanee, H., Munawar, A., Yasin, A. and Zain, A.M., 2021. Social media and stock market prediction: a big data approach. *MJ Awan, M. Shafry, H. Nobanee, A. Munawar, A. Yasin et al.," Social media and stock market prediction: a big data approach," Computers, Materials & Continua*, *67*(2), pp.2569-2583.

[4]. Chaudhary, K., Alam, M., Al-Rakhami, M.S. and Gumaei, A., 2021. Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics. *Journal of Big Data*, *8*(1), p.73.

[5]. Zhu, B., Zheng, X., Liu, H., Li, J. and Wang, P., 2020. Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. *Chaos, Solitons & Fractals*, *140*, p.110123.

[6]. Manguri, K.H., Ramadhan, R.N. and Amin, P.R.M., 2020. Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan Journal of Applied Research, pp.54-65.

[7]. Jung, S.H. and Jeong, Y.J., 2020. Twitter data analytical methodology development for prediction of start-up firms' social media marketing level. Technology in Society, 63, p.101409.

[8]. Antonakaki, D., Fragopoulou, P. and Ioannidis, S., 2021. A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. Expert systems with applications, 164, p.114006.
[9] Shang, W.L., Chen, J., Bi, H., Sui, Y., Chen, Y. and Yu, H., 2021. Impacts of COVID-19 pandemic on user behaviors and environmental benefits of bike sharing: A big-data analysis. Applied Energy, 285, p.116429.

[10] Islam, M.R., Liu, S., Wang, X. and Xu, G., 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Network Analysis and Mining, 10(1), p.82.

[11] Aliahmadi, A., Nozari, H. and Ghahremani-Nahr, J., 2022. Big Data IoT-based agile-lean logistic in pharmaceutical industries. International Journal of Innovation in Management, Economics and Social Sciences, 2(3), pp.70-81.

[12] Singh, M., Jakhar, A.K. and Pandey, S., 2021. Sentiment analysis on the impact of coronavirus in social life using the BERT model. Social Network Analysis and Mining, 11(1), p.33.

[13] Grover, P., Kar, A.K. and Dwivedi, Y.K., 2022. Understanding artificial intelligence adoption in operations management: insights from the review of academic literature and social media discussions. Annals of Operations Research, 308(1), pp.177-213.

[14] Coffey, Y., Bhullar, N., Durkin, J., Islam, M.S. and Usher, K., 2021. Understanding eco-anxiety: A systematic scoping review of current literature and identified knowledge gaps. The Journal of Climate Change and Health, 3, p.100047.

[15] Abbate, S., Centobelli, P., Cerchione, R., Nadeem, S.P. and Riccio, E., 2024. Sustainability trends and gaps in the textile, apparel and fashion induustries. Environment, Development and Sustainability, 26(2), pp.2837-2864.

[16] Kushwaha, A.K., Kar, A.K. and Dwivedi, Y.K., 2021. Applications of big data in emerging management disciplines: A literature review using text mining. International Journal of Information Management Data Insights, 1(2), p.100017.

[17] Wang, J., Yang, Y., Wang, T., Sherratt, R.S. and Zhang, J., 2020. Big data service architecture: a survey. Journal of Internet Technology, 21(2), pp.393-405.

[18] Lawn, S., Oster, C., Riley, B., Smith, D., Baigent, M. and Rahamathulla, M., 2020. A literature review and gap analysis of emerging technologies and new trends in gambling. International journal of environmental research and public health, 17(3), p.744.

[19] Sivarajah, U., Irani, Z., Gupta, S. and Mahroof, K., 2020. Role of big data and social media analytics for business to business sustainability: A participatory web context. Industrial Marketing Management, 86, pp.163-179.

[20] Khanra, S., Dhir, A. and Mäntymäki, M., 2020. Big data analytics and enterprises: a bibliometric synthesis of the literature. Enterprise Information Systems, 14(6), pp.737-768.

[21] Tao, D., Yang, P. and Feng, H., 2020. Utilization of text mining as a big data analysis tool for food science and nutrition. Comprehensive reviews in food science and food safety, 19(2), pp.875-894.

[22] Iqbal, R., Doctor, F., More, B., Mahmud, S. and Yousuf, U., 2020. Big data analytics: Computational intelligence techniques and application areas. Technological Forecasting and Social Change, 153, p.119253.

[23] Liu, X., Shin, H. and Burns, A.C., 2021. Examining the impact of luxury brand's social media marketing on customer engagement: Using big data analytics and natural language processing. Journal of Business research, 125, pp.815-826.

[24] Balaji, T.K., Annavarapu, C.S.R. and Bablani, A., 2021. Machine learning algorithms for social media analysis: A survey. Computer Science Review, 40, p.100395.

[25] Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R. and Mora, H., 2020. A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. Industrial Marketing Management, 90, pp.523-537.

[26] Misra, N.N., Dixit, Y., Al-Mallahi, A., Bhullar, M.S., Upadhyay, R. and Martynenko, A., 2020. IoT, big data, and artificial intelligence in agriculture and food industry. IEEE Internet of things Journal, 9(9), pp.6305-6324.

[27] Sheng, J., Amankwah‐Amoah, J., Khan, Z. and Wang, X., 2021. COVID‐19 pandemic in the new era of big data analytics: Methodological innovations and future research directions. British Journal of Management, 32(4), pp.1164-1183.

[28] Nemes, L. and Kiss, A., 2021. Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication, 5(1), pp.1-15.

[29] Journalofbigdata, (2021) Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics. Accessed From: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00466-2

[30] Khan, W., Ghazanfar, M.A., Azam, M.A., Karami, A., Alyoubi, K.H. and Alfakeeh, A.S., 2022. Stock market prediction using machine learning classifiers and social media, news. Journal of Ambient Intelligence and Humanized Computing, pp.1-24.

[31] ElHaffar, G., Durif, F. and Dubé, L., 2020. Towards closing the attitude-intention-behavior gap in green consumption: A narrative review of the literature and an overview of future research directions. Journal of cleaner production, 275, p.122556.

[32] Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C.J., Sosa, S.S., Roberts, B.W. and Briley, D.A., 2022. Personality stability and change: A meta-analysis of longitudinal studies. Psychological bulletin, 148(7-8), p.588.

[33] Sitzmann, T. and Campbell, E.M., 2021. The hidden cost of prayer: Religiosity and the gender wage gap. Academy of Management Journal, 64(4), pp.1016-1048.

[34] Kumar, S., Kar, A.K. and Ilavarasan, P.V., 2021. Applications of text mining in services management: A systematic literature review. International Journal of Information Management Data Insights, 1(1), p.100008.

[35] Albalawi, R., Yeap, T.H. and Benyoucef, M., 2020. Using topic modeling methods for short-text data: A comparative analysis. Frontiers in artificial intelligence, 3, p.42.