SENTIMENT ANALYSIS OF SMALL CAP AND BIG CAP COMPANIES IN UNITED STATES INDEXES

MUHAMMAD KHAIRULRAZI BIN MOHD RIZA

SESSION 2022/2023

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
FEBRUARY 2023

SENTIMENT ANALYSIS OF SMALL CAP AND BIG CAP COMPANIES IN UNITED STATES INDEXES

BY

MUHAMMAD KHAIRULRAZI BIN MOHD RIZA

SESSION
2022/2023

THIS PROJECT REPORT IS
PREPARED FOR

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
IN PARTIAL FULFILLMENT
FOR

BACHELOR OF COMPUTER SCIENCE
B.CS (HONS) DATA SCIENCE

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
FEBRUARY 2023

DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this thesis has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

_____

Name of candidate: Muhammad Khairulrazi Bin Mohd Riza

Faculty of Computing & Informatics

Multimedia University

Date: 1/2/2023

ACKNOWLEDGEMENT

I extend my sincerest appreciation and gratitude to all individuals who have contributed to the completion of this Final Year Project 1 and Final Year Project 2. I would like to specially acknowledge my supervisor, Mr. Nathar Shah Packier Mohammad, and my group member, Imran Kamil Bin Saiful Rijal, for their unwavering support and motivation throughout the writing process, especially during the literature review, data collection, and data labelling phase. I am deeply grateful for their valuable time and efforts in proofreading and rectifying any errors in the report. Their guidance and encouragement were instrumental in making this project a success.

ABSTRACT

Sentiment analysis of social media has gained significant attention in recent years due to the vast amount of information available on these platforms. In this systematic literature review (SLR), we examine the various approaches and techniques used for sentiment analysis of social media data. We focus specifically on the use of Support Vector Machine (SVM) and lexicon-based methods. We present a comprehensive review of the existing literature on these approaches and discuss their advantages and limitations. Our review also includes a discussion of the challenges and future directions in the field of sentiment analysis of social media. Overall, our review highlights the potential of using SVM and lexicon-based approaches for sentiment analysis on social media and provides valuable insights for researchers working in this area.

Keywords: *sentiment analysis, social media, support vector machine, lexicon, literature review, natural language processing*

TABLE OF CONTENTS

ACKNOWLEDGEMENT **5**
ABSTRACT **6**
TABLE OF CONTENTS **7**
LIST OF TABLES **9**
Chapter 6: Evaluation of Findings 9
LIST OF FIGURES **10**
Chapter 2: Literature Review 10
Chapter 3: Theoretical Framework 10
Chapter 4: Research Methodology 10
Chapter 5: Implementation 10
Chapter 6: Evaluation of Findings 10
1 Introduction **11**
1.1 Problem Statement 12
1.2 Project Objectives 13
1.3 Expected Findings 13
1.4 Project Scope 14
1.5 Chapter Organisation 14
2 Literature Review **15**
2.1. Background 15
2.2. Overall survey of sentiment analysis 16
2.2.1 General overview of the sentiment analysis field. 16
2.2.2 Techniques used in sentiment analysis 18
2.2.3 Challenges faced by authors when doing sentiment analysis 19
2.3 Survey of sentiment analysis using Support Vector Machine (SVM) in social media 21
2.3.1 Support Vector Machine (SVM) in sentiment analysis 21
2.3.2 Challenges of using Support Vector Machine in Sentiment Analysis 22
2.4 Survey of sentiment analysis using Lexicon-based approaches in social media 23
2.4.1 Lexicon-based approaches in sentiment analysis 23
2.4.2 Challenges of using Lexicon-based approaches in Sentiment Analysis 24
3 Theoretical Framework **25**
3.1 Evaluating performances of techniques chosen to perform sentiment analysis . 25
3.1.1 Performance of sentiment analysis using Support Vector Machine (SVM) on social media data 26
3.1.2 Performance of sentiment analysis using Lexicon-based approaches on social media data. 29
3.2 Transitioning from FYP1 to FYP2 30
4. Research Methodology **32**
4.1 Survey Methodology 33
4.2 Implementation Methodology 38

4.2.1 Birds eye view of Implementation process 38
5. Implementation **41**
5.1 Implementation plan 41
5.2 Implementation of Sentiment Analysis 42
5.2.1 Data Collection 43
5.2.2 Data Labelling 44
5.2.3 Data Preprocessing 45
5.2.3 TextBlob Implementation 48
5.2.4 ChatGPT Implementation 49
5.2.5 Support Vector Machine (SVM) Implementation 51
6 Evaluation of Findings **52**
6.1 TextBlob Results 52
6.2 ChatGPT Results 55
6.3 Support Vector Machine Results 56
6.4 Performance Comparison 59
6.5 Discussion of Results 59
7 Conclusion **61**
References **63**
8. Appendices **68**
8.1 Appendix A 68
8.1.1 FYP Meeting Log 1 68
8.1.2 FYP Meeting Log 2 74
8.1.3 FYP Meeting Log 3 77
8.1.4 FYP Meeting Log 4 82
8.1.5 FYP Meeting Log 5 87
8.1.6 FYP Meeting Log 6 92
8.1.7 FYP Meeting Log 1 97
8.1.8 FYP Meeting Log 2 103
8.1.9 FYP Meeting Log 3 109
8.1.10 FYP Meeting Log 4 115
8.1.11 FYP Meeting Log 5 120
8.1.12 FYP Meeting Log 6 126
8.2 Appendix B 132
8.3 Appendix C 143

LIST OF TABLES

Chapter 6: Evaluation of Findings

Table 1. Accuracy of TextBlob……..53
Table 2. Accuracy of ChatGPT…….…….55
Table 3. Hyperparameter settings for SVM ……..……57
Table 4. Performance of SVM model (Default vs Tuned).....57
Table 5. Performance comparison……..59

LIST OF FIGURES

Chapter 2: Literature Review

Figure 1. Number of review papers from each database used……..……...…16
Figure 2. Annual Twitter users 2012 to 2022……..……..…..18

Chapter 3: Theoretical Framework

Figure 3. Accuracy of ML algorithms obtained by Ashwin et al……..……..26
Figure 4. Comparison of the systems obtained by S. Nax et al……..…..…...27
Figure 5. Comparison of each SVM kernel obtained by N. Hasanati et al…….…….,…28
Figure 6. Overall flow of FYP1 and FYP2…….……..31

Chapter 4: Research Methodology

Figure 7. Survey methodology process…….……36
Figure 8. FYP2 implementation flowchart……...…….……38

Chapter 5: Implementation

Figure 9. Gantt chart for final year project 2……..…….……..41
Figure 10. Flow of data collection……...……43
Figure 11. Sentiment distribution of big cap and small cap tweets……...……....45
Figure 12. Flow of TextBlob……48
Figure 13. Flow of ChatGPT……49
Figure 14. ChatGPT sentiment…….……50
Figure 15. Flow of SVM……51

Chapter 6: Evaluation of Findings

Figure 16. Manual vs TextBlob sentiment distribution…….……53
Figure 17. Polarity vs Subjectivity (smallcap).....54
Figure 18. Polarity vs Subjectivity (bigcap).....54
Figure 19. Small cap word cloud……55
Figure 20. Big cap word cloud……55
Figure 21. Manual vs ChatGPT sentiment distribution (small cap).....56
Figure 22. Manual vs ChatGPT sentiment distribution (big cap).....56
Figure 23. Manual vs Tuned SVM sentiment distribution (small cap).....58
Figure 24. Manual vs Tuned SVM sentiment distribution (big cap).....58

1 Introduction

The process of extracting subjective information from text data is known as sentiment analysis, also known as opinion mining. It has become an important research area in natural language processing because of the increasing availability of text data in the form of user reviews, comments, and posts. Social media platforms such as Twitter, Facebook, and Instagram provide a rich source of data for studying public opinion on various topics.

In recent years, various approaches and techniques have been proposed for sentiment analysis of social media data. Among these, support vector machine (SVM) and lexicon-based methods have received significant attention due to their effectiveness and ease of use. SVM is a type of supervised machine learning algorithm that can be used to classify text data into different categories. Lexicon-based methods, on the other hand, rely on the use of dictionaries or pre-defined lists of words and their associated sentiments to classify text data.

The goal of the first semester of this final year project is to provide a comprehensive overview of existing research on sentiment analysis using SVM and lexicon-based approaches, as well as other techniques. We will discuss the advantages and limitations of these approaches, as well as the difficulties and potential future development in the field. Our review will be useful for researchers working in the subject of text analysis and for those interested in understanding the current state of the art in this field.

For the second part of this final year project, the goal is to implement the two chosen techniques, SVM and lexicon-based approaches, alongside an additional technique called ChatGPT. This implementation involves performing sentiment analysis on Twitter data, specifically targeting big cap and small cap companies in the US indexes. A detailed comparative analysis is then conducted to evaluate the performance of the chosen techniques. We aim to provide comprehensive understanding of sentiment analysis in the context of the stock market and contribute to valuable insights to the field.

1.1 Problem Statement

The problem that we are addressing is the lack of a comprehensive overview of the existing research on sentiment analysis of social media data using SVM and lexicon-based approaches. The use of social media platforms has increased significantly in recent years for various purposes, including marketing, customer service, and political campaigns. As a result, there is a need to understand the sentiment of social media users on various topics in order to effectively target and engage them. However, there is a lack of a systematic review that compares and contrasts the different approaches and techniques used for sentiment analysis of social media data using SVM and lexicon-based methods. This makes it difficult for researchers and practitioners to understand the current state of this field and to identify the strengths and limitations of different approaches.

1.2 Project Objectives

- Develop comprehensive understanding of sentiment analysis techniques of Support Vector Machine (SVM), Lexicon-based approach (TextBlob) and ChatGPT.
- Investigate the performance of sentiment analysis techniques in predicting sentiment for small cap and big cap companies in the US indexes.
- Compare accuracy and effectiveness of SVM, TextBlob and ChatGPT in sentiment analysis of small cap and big cap companies in the US indexes.
- Assess the impact of hyperparameter tuning on the performance of SVM in sentiment analysis.

1.3 Expected Findings

The expected findings from this project could include the following:

- Understanding of the effectiveness of SVM and lexicon-based methods in sentiment analysis on social media data.
- Characteristics of sentiment analysis.
- Understanding the strengths and limitations of SVM and lexicon-based approaches on sentiment analysis from social media data.
- Analysis of the potential use of sentiment analysis to predict stock market trends.
- Identify the most performing technique for sentiment analysis of big cap and small cap companies in US indexes.

1.4 Project Scope

Sentiment analysis of small- and large-cap companies listed on United States indices is part of this initiative. The project's scope includes, but is not limited to:

- Extract meaningful findings from articles related to sentiment analysis.
- Create a literature review paper about sentiment analysis techniques.
- Select 2 techniques to implement in sentiment analysis of small cap and big cap companies in United States indexes for FYP2.
- Implement and evaluate the performance of the chosen techniques.

1.5 Chapter Organisation

The rest of the paper is organised as follows: Chapter 2 provides a literature review on sentiment analysis using Support Vector Machine (SVM) and lexicon-based approaches. Chapter 3 is the theoretical framework. This section will explain about the performance of sentiment analysis from the chosen technique obtained from conducting literature review and discusses the flow of the final year project. Chapter 4 outlines the survey methodology for the first phase of this final year project and the implementation methodology for the second phase of this final year project which contains the overall birds eye view of the implementation. Chapter 5 presents the implementation plan which includes detailed explanation of each implementation used for this project. Chapter 6 will discuss the performance of the techniques chosen in order to perform sentiment analysis for this project. Chapter 7 will provide the conclusion and recommendations.

2 Literature Review

2.1. Background

The process of extracting subjective information from textual data is known as sentiment analysis, also known as opinion mining. Because of the increasing availability of textual data in the form of user reviews, comments, and posts, it has become an important research area in natural language processing. Twitter, Facebook, and Instagram, for example, provide a rich source of data for studying public opinion on a variety of topics.

In recent years, various approaches and techniques have been proposed for sentiment analysis of social media data. Among these, support vector machine (SVM) and lexicon-based methods have received significant attention due to their effectiveness and ease of use. SVM is a type of supervised machine learning algorithm that can be used to classify text data into different categories. Lexicon-based methods, on the other hand, rely on the use of dictionaries or pre-defined lists of words and their associated sentiments to classify text data.

In this literature review, we aim to provide a comprehensive overview of the existing research on sentiment analysis using SVM and lexicon-based approaches and other techniques as well. We will discuss the advantages and limitations of these approaches, as well as the difficulties and future directions in the field. Our review will be useful for researchers working in the area of sentiment analysis and for those interested in understanding the current state of the art in this field.

2.2. Overall survey of sentiment analysis

2.2.1 General overview of the sentiment analysis field.

Sentiment analysis is a branch of natural language processing (NLP) that aims to identify and extract subjective information from text data automatically. This includes determining the sentiment or emotional tone of a piece of text, as well as identifying the opinions or attitudes expressed by the writer towards a particular subject.

According to a recent survey by Accenture (2020), sentiment analysis is one of the fastest-growing applications of Natural Language Processing (NLP), and its applications are increasing in a wide range of industries, including marketing, customer service, social media, finance, and healthcare.

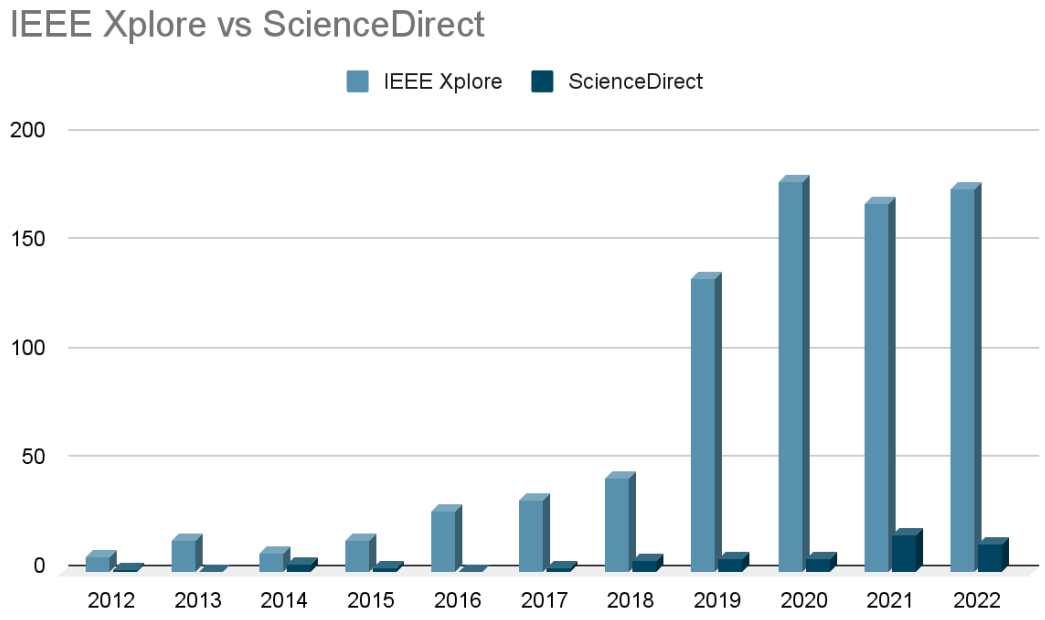


Figure 1. Number of review papers from each database used

Based on **Figure 1**, analysis of review papers on sentiment analysis published from 2012 to 2022 shows a steady increase in the number of papers, indicating the growing interest and significance of the field. Therefore, supporting the claim made by Accenture (2020) that sentiment analysis is a rapidly growing application of Natural Language Processing (NLP) and its use is increasing in various industries.

Sentiment analysis, which has been studied extensively in recent times, has been put to use in many different fields. Some examples include analysing social media posts, evaluating customer opinions, and reviewing products. The technique has also been utilised in various industries like politics, finance, and healthcare.

One industry that we will be focusing on in this literature review is social media. Sentiment analysis in social media has gained significant attention in recent years, largely due to the rapid growth in the number of social media users. According to Statista (2021), the number of social media users worldwide has steadily increased from 2.8 billion in 2016 to 4.9 billion in 2021. For example, **Figure 2** demonstrates the rising number of Twitter users from the year 2012 to 2022.

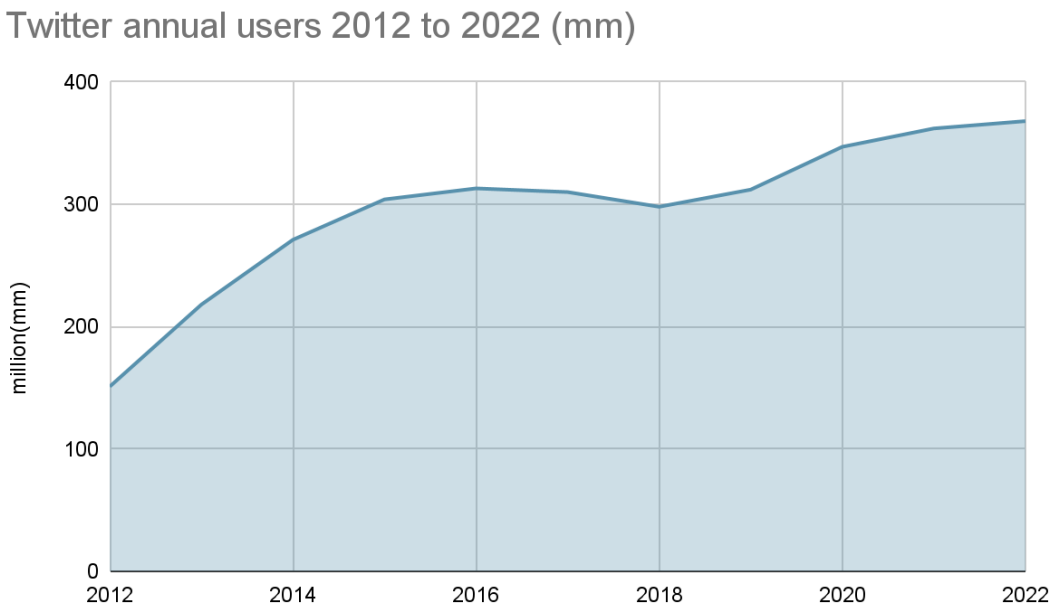


Figure 2 Annual Twitter users 2012 to 2022

As a result of this growth, the volume of user-generated data on social media platforms has increased dramatically, making social media an attractive source of data for sentiment analysis. Overall, sentiment analysis is a significant and expanding area of study that has the capability to change the way we comprehend and make choices utilising a large amount of textual data.

2.2.2 Techniques used in sentiment analysis

Sentiment analysis is a field that has seen numerous approaches and techniques proposed. Some of these include lexicon-based methods, machine learning approaches, hybrid methods that combine lexicon-based and machine learning techniques, comparative methods using both lexicon and machine learning.

One popular technique used in sentiment analysis is lexicon-based approach. This approach is often used as a simple and fast baseline, but it can be limited in its accuracy because sentiment can be expressed in many different ways and can depend on context. Other than that, another popular technique used in sentiment analysis is the use of machine learning algorithms. According to Xu et al. (2022), sentiment analysis that uses machine learning, rather than relying on lexicon-based models, trains the model to recognise sentiments through the use of sentimental cues in the text, allowing for automatic detection. Additionally, hybrid methods, which use both lexicon-based and machine learning techniques, have also been proposed to leverage the strength of both approaches. Comparative methods, which use both lexicon-based and machine learning techniques, have been used to compare the performance of different methods and identify the most effective approach for a particular task.

Overall, these different approaches have been widely studied and have demonstrated varying levels of success in sentiment analysis tasks, depending on the specific use case and the dataset used for evaluation.

2.2.3 Challenges faced by authors when doing sentiment analysis

Despite the promising results of various techniques for sentiment analysis, there are also limitations and challenges that need to be addressed. One of the challenges of sentiment analysis is the presence of negation. For example, M. Bouazizi and T. Ohtsuki (2019) explained that it can be difficult to determine if the presence of negation switches the polarity and, even in multi-class classification, a polarity switch doesn’t necessarily mean the sentiment is the opposite of the negated

word. As an example, the authors provide a tweet that includes the word “happy” and negation, which typically expresses happiness, but the negation in the tweet actually shows anger rather than sadness.

Other than that, according to M. Bouazizi and T. Ohtsuki (2019), context dependency is another challenge in sentiment analysis. They mention that tweets are often in response to other tweets, making their sentiment highly dependent on context. As an example, the authors cite a tweet that was considered neutral but was labeled as showing anger by some annotators due to assumptions about dissatisfaction with an event. The authors point out that machine struggle with imagining these scenarios and accurately determining the sentiment.

Additionally, M. Bouazizi and T. Ohtsuki (2019) also note that polysemy presents a challenge in sentiment analysis. They explain that words can have multiple meanings depending on their context and that even similar meanings can indicate different emotions. The authors give the example of the word “mad” which can mean angry or crazy, but can also be used as an adverb meaning “very”. They note that despite the presence of two sentimental words, the tweet “It was mad fun man!” was clussified as showing anger. The authors suggest that better PoS-tagging could help identify “mad” as an adverb.

In addition, M. F. R. Abu Bakar (2020) identified 3 more challenges in sentiment analysis which include the difficulty in handling informal patterns and evolving language on social media sites such as Facebook and Twitter. Additionally,

the authors found that there is a lack of relevant resources and training data for Malay sentiment analysis, this is due to the language barrier, which makes research in this area more time-consuming compared to English language sentiment analysis. Another challenge is the detection of sarcasm in text, which is complex as the true meaning of sarcasm in text depends on the individual's desire.

2.3 Survey of sentiment analysis using Support Vector Machine (SVM) in social media

2.3.1 Support Vector Machine (SVM) in sentiment analysis

Support vector machine (SVM) is a popular machine learning technique used in sentiment analysis. The basic idea behind SVM is to find the optimal hyperplane that separates the data points into different classes, in this case, positive, negative, and neutral sentiments. The hyperplane is selected in a way that maximises the margin, which represents the interval between the hyperplane and the nearest data points from each class. SVM can also be used for multi-class sentiment analysis by training multiple binary classifiers.

SVM has been widely used in sentiment analysis and has been shown to achieve good performance in various domains such as social media, customer feedback, and product reviews.

2.3.2 Challenges of using Support Vector Machine in Sentiment Analysis

One of the challenges of using SVM in sentiment analysis is its computational complexity when dealing with large amounts of labelled data. SVM models require a large amount of labelled data to achieve good performance, according to a study by S. Naz, A. Sharan, and N. Malik (2018), making them difficult to apply in situations where labelled data is scarce

Another challenge of using SVM in sentiment analysis is selecting the appropriate kernel. The choice of kernel has a significant impact on the performance of the SVM model, but finding the best kernel for a specific dataset can be challenging. Experimentation and trial-and-error are often necessary to determine the best kernel for a specific dataset. For example, in the paper “Implementation of Support Vector Machine with Lexicon Based for Sentiment Analysis on Twitter” by N. Hasanati, Q. Aini, and A. Nuri (2022), the authors address the differences between using different kernels on Twitter dataset and found that the “rbf” kernel is the best kernel to use. The accuracy for each kernel is displayed in Figure 6.

Additionally, SVM are not good at handling multi-class classification tasks. Sentiment analysis often involves classifying text into multiple sentiment classes (positive, negative, and neutral), and SVM are typically designed to handle binary classification problems. This makes it difficult to apply SVM models to multi-class sentiment analysis problems.

In conclusion, while SVM has been shown to achieve good performance in sentiment analysis, it also has some limitations that need to be considered. These limitations include reliance on a large amount of labelled data, difficulties selecting the appropriate kernel, and poor performance on multi-class classification tasks. Despite these challenges, SVM remains a popular and widely used technique in sentiment analysis, and ongoing research continues to address and improve these challenges.

2.4 Survey of sentiment analysis using Lexicon-based approaches in social media

2.4.1 Lexicon-based approaches in sentiment analysis

Lexicon-based approach in sentiment analysis involves utilising a dictionary or lexicon of words and their sentiment polarities (positive, negative, or neutral) to categorise the sentiment expressed in text. This approach is rooted in natural language processing and computational linguistics, and it has been widely used in many different domains such as social media, customer reviews, and opinion mining. The lexicon-based approach has been shown to be simple and effective in determining the sentiment of text, but it also has limitations such as its inability to capture the context and meaning of words, and its dependence on the quality and coverage of the lexicon used. This approach is widely used in research and industry, and it is a popular choice for sentiment analysis due to its ease of implementation. However, it is important to note that recent research has shown that more advanced

methods such as neural networks and machine learning algorithms can achieve better performance in sentiment analysis.

2.4.2 Challenges of using Lexicon-based approaches in Sentiment Analysis

The challenges of using the lexicon-based approach in sentiment analysis are many, and recent studies have highlighted several limitations of this approach. One of the challenges of using a Lexicon-based approach in sentiment analysis is its dependency on the presence of predefined words in dictionaries. According to Khatoon S. et al. (2020), the accuracy of lexicon-based sentiment analysis is impacted by the complexity of the dictionaries, including the presence of usage patterns, rules, and linguistic structures. Additionally, the sentiment orientation of a word in one dictionary can be different from its orientation in another. This makes it difficult to fully understand the diversity of lexicon-based approaches due to the incompleteness of vocabulary.

Another challenge of using a Lexicon-based approach is explained by Khan et al. (2016). The author mentioned that the techniques rely on lexical resources, and the overall effectiveness is highly dependent on the quality of the lexical resources. The polarity of a piece of text can be obtained based on the polarity of the words that compose it. As a result, it is not designed to cover all aspects of language especially when it comes to slang, sarcasm, and negation. This is because of the complexity of natural languages.

Moreover, Akter et al. (2016) pointed out that relying solely on sentiment words is inadequate. This is due to several issues, including words having varying meanings depending on the context, sentences containing sentiment words sometimes lacking sentiment expression, and numerous sentences without sentiment words still implying a sentiment.

All in all, lexicon-based approach is a widely used method in sentiment analysis but it also faces several challenges such as limitations in handling complex sentiments, difficulties in dealing with negations and idiomatic expressions, and its reliance on predefined sentiment dictionaries which may not always accurately reflect the sentiment expressed in a given text. However, with the advancement of NLP techniques and the increasing availability of high-quality sentiment lexicons, the performance of lexicon-based approaches in sentiment analysis continues to grow.

3 Theoretical Framework

3.1 Evaluating performances of techniques chosen to perform sentiment analysis .

This section provides an evaluation of the performance of the chosen sentiment analysis techniques, namely Lexicon-based approaches and Support Vector Machine (SVM) on social media data, as identified in the Literature Review.

3.1.1 Performance of sentiment analysis using Support Vector Machine (SVM) on social media data

A study published in 2021 by Ashwin et al. titled "Sentiment analysis and classification of Indian farmers' protest using Twitter data" gathered around 20,000 tweets to analyse the sentiments expressed by the public regarding a protest. They used Bag of Words and TF-IDF techniques and discovered that Bag of Words produced better results compared to TF-IDF. Additionally, they applied machine learning algorithms like Support Vector Machines, Naive Bayes, Random Forest, and Decision Trees and found that Random Forest achieved the highest level of classification accuracy.

	Accuracy	
	Bag of Words	TF-IDF
Naïve Bayes	72.9	71.33
Decision Tree	79.78	77.62
Random Forest	96.62	95.51
SVC	83.45	83.04

Figure 3. Accuracy of ML algorithms obtained by Ashwin et al. (2021)

Figure 3 exhibits the accuracy of the ML algorithms used by Ashwin et al. (2021) when considering the analysis of twitter data about a protest. Among the machine learning techniques, Naive Bayes had the lowest accuracy of 72%, while Random Forest demonstrated the highest accuracy of 96.62%. SVM and Decision Tree had intermediate accuracy values of 83.45% and 79.78%, respectively.

Additionally, S. Naz et al. (2018) proposed a method for Twitter sentiment classification using Support Vector Machine (SVM). Experiments showed that unigram features performed best among four different n-gram feature sets, and combining their method with n-grams improved SVM classifier accuracy. The proposed approach achieved an accuracy of 81.0%, outperforming the OPAL and VCU-TSA systems with 79.2% and 77.5% accuracy, respectively. **Figure 4** shows a brief comparison of the systems.

	Avg Recall	Accuracy
Proposed approach	0.617	81.0%
OPAL	0.616	79.2%
VCU-TSA	0.448	77.5%

Figure 4. Comparison of the systems obtained by S. Nax et al. (2018)

Another study by N. Hasanati et al. (2022) applied SVM algorithm for text analysis of Covid-19 vaccine tweets, utilizing lexicon method to label data into positive, neutral, and negative classes. The best parameters for SVM were determined using grid search and 10-fold cross-validation, and the rbf kernel resulted in highest accuracy (85.28%), precision (75.96%), and recall (76.36%) among the linear, rbf, and poly kernel tested. **Figure 5** shows a brief comparison of the performance of each kernel from SVM obtained from the authors.

Kernels	Best Parameter Value	Accuracy	Precision	Recall
linear	C=1	84.55%	71.60%	72.19%
RBF	C=100, gamma=0.01	84.88%	74%	73.88%
Poly	C=10, degree=1	84.69%	70.67%	72.64%

Figure 5. Comparison of SVM kernel obtained by N. Hasanati et al. (2022)

Another study by Dangi et al. (2022) compared the effectiveness of five machine learning algorithms (Random Forest Classifier, Multinomial NB Classifier, Logistic Regression, Support Vector Machine, and Decision Tree) in analyzing two Twitter datasets to identify sentiments in COVID-19. The evaluation metrics used were recall, f1, support, and precision. The study found that logistic regression performed better in classifying positive class in dataset 1, while SVM performed better in classifying negative tweets in dataset 2. The accuracy for the classifiers in dataset 1 ranged from 0.97 to 0.98 and in dataset 2 ranged from 0.94 to 0.99.

Another study by Katarya et al. (2022) investigated people's sentiments on the unconventional coronavirus via Twitter messages. They tested 5 machine learning classifiers namely, Random Forest, Logistic Regression, Multinomial Naive Bayes, KNN, and SVM using precision and balanced accuracy. The results obtained by the authors showed that the KNN classifier achieved the best precision of 81%, while Logistic Regression had the best recall, f1, accuracy, and balanced accuracy of approximately 73%. Support Vector Machine classifier obtained a consistent score of approximately 67% across all metrics.

In general, all of the other sentiment analysis techniques performed fairly well with Support Vector Machine (SVM). The results obtained from researchers were consistent and showed that these techniques have the potential to be used for sentiment analysis in various applications, especially in the domain of social media. The use of SVM in particular showed promising results, with its accuracy and robustness in handling large datasets making it a valuable tool for sentiment analysis. In conclusion, all of the various techniques tested by the researchers demonstrated their effectiveness, reliability, and accuracy.

3.1.2 Performance of sentiment analysis using Lexicon-based approaches on social media data.

The study “Analysing sentiment system to specify polarity by lexicon-based” conducted by Abd, D., Abbas, A., & Sadiq, A. (2021) employed a corpus-based approach for sentiment analysis at the phrase level and experimented their model using the IMDB dataset. In their study, Abd, D., Abbas, A., & Sadiq, A. (2021) generated a sentiment lexicon by categorising the words into positive and negative categories based on the training set. To optimise their model's performance, they performed four experiments using sentiment dictionaries consisting of 25000, 30000, 34000, and 40000 review terms and applied them to classify the test set. While the second experiment achieved the highest accuracy of 76.585% on the IMDB dataset, the results of the four trials were relatively consistent, hovering around 76%.

In another study by V. Ramanathan and T. Meyyappan conducted a study in 2019 where they investigated the impact of four factors - domain-specific ontology, entity-specific opinion extraction, a lexicon-based approach combined with conceptual semantic sentiment analysis - on the sentiment analysis of tweets related to tourism in Oman. The study results showed that the combination of lexicon-based approach and conceptual semantic sentiment analysis significantly enhanced the sentiment analysis accuracy to 85.54%. The authors' conclusion was that a majority of people hold positive opinions about tourism in Oman.

Additionally, H. Karamollaoğlu, İ. A. Doğru, et al. (2018) proposed a method to determine the sentiment density of Turkish tweets drawn from Twitter, using the Lexicon-based approach. The method leveraged the SentiWordNet sentiment dictionary by using the English equivalent of Turkish words in the tweets. The sentiment analysis process was applied to comments obtained from social media platforms with an average success rate of 80%.

3.2 Transitioning from FYP1 to FYP2

This final year project is divided into two phases. The first phase involves a comprehensive literature review of the techniques used in sentiment analysis, specifically focusing on *Support Vector Machine (SVM) and lexicon-based approach.* The aim of this phase is to identify the strengths and weaknesses of these techniques, as well as how they have performed in different domains, the challenges they have

faced, and any potential improvements. The overall flow of this FYP is presented in **Figure 6.**

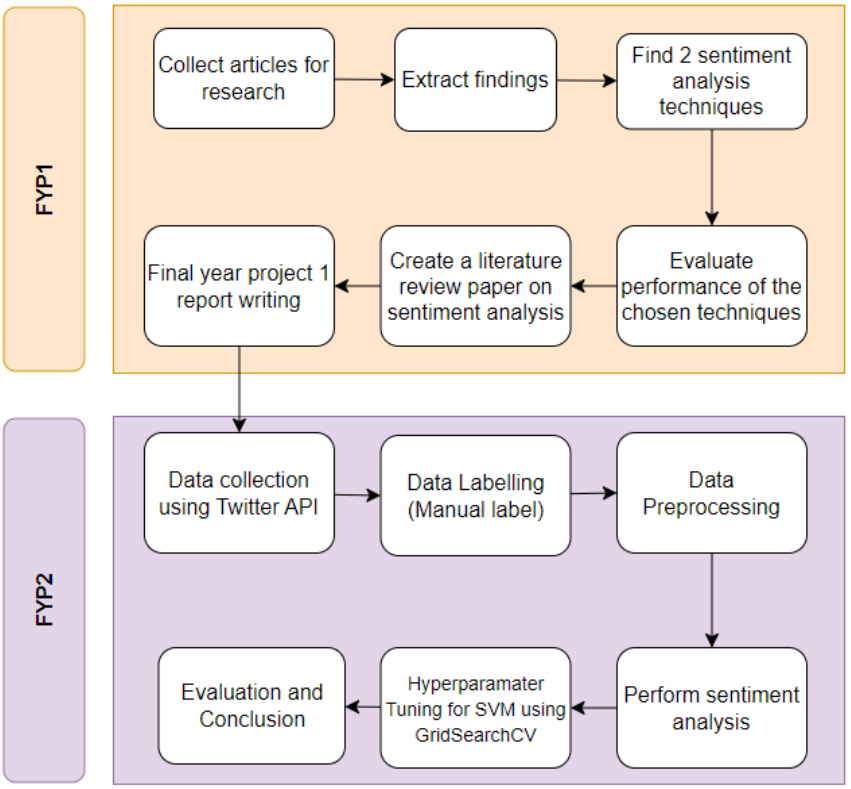


Figure 6. Overall flow of FYP1 and FYP2

A survey on the sentiment analysis techniques of Support Vector Machine (SVM) and lexicon-based approach is presented in ***Chapter 2*** and ***Chapter 3.1***. This survey provides an overview of the key concepts, techniques, and applications of these approaches, as well as an assessment of their performance and limitations in the field of sentiment analysis.

The second phase of the project is the implementation phase, where the SVM and lexicon-based approaches will be utilised to perform sentiment analysis on

Twitter data with the goal of predicting the sentiment of small cap and big cap companies in the United States Indexes.

Additionally, during the implementation phase of FYP2 and extensive discussions with the team members and supervisors, another technique, namely, ChatGPT has been incorporated to perform sentiment analysis. This is due to the booming and surprising rise of LLM (large-language models). Making the total technique to perform sentiment analysis to be 3, which is, Lexicon, Support Vector Machine, and ChatGPT.

4. Research Methodology

The first phase of the final year project involves conducting a literature review on opinion mining on social media data using Support Vector Machine (SVM) and lexicon-based approach. The goal of this phase is to understand the current state of 34 the art in sentiment analysis on social media and to identify the strengths and limitations of different approaches.

After doing the first phase of the final year project, the second phase of this final year project involves implementing sentiment analysis using the selected techniques, namely, TextBlob (Lexicon), Support Vector Machine (SVM), and ChatGPT. This phase includes the data collection, labelling, preprocessing, and the implementation of the three techniques in order to provide comprehensive evaluation.

4.1 Survey Methodology

There are six steps in conducting the systematic literature review (SLR) which are listed below. Before conducting the literature review, it is crucial to establish its scope and purposes, as well as identify the relevant databases and sources to search. The first step is to identify and collect all of the relevant studies and articles. The studies are then screened and selected using established inclusion and exclusion criteria. The quality and relevancy of the selected studies are critically evaluated. The fourth step involves extracting data from the selected studies and organising it into a structured format. The extracted data is then synthesised and conclusions are drawn based on the findings. Finally, the results are presented and discussed in a clear and concise manner. The SLR process is guided by numerous research questions which are:

Research Questions

1. How does SVM and Lexicon perform when studying sentiments from social media?
2. What difficulties are encountered when conducting sentiment analysis on social media datasets?
3. Is SVM and Lexicon a reliable technique for predicting and analysing social media sentiment?

Answering the questions is important because it can provide insights into the effectiveness of SVM and lexicon-based approaches for text analysis on social media

data. Comparing the performance of these approaches on different datasets can help researchers and practitioners to understand their strengths and weaknesses, and to choose the most suitable approach for a particular task. In addition, understanding the performance of these approaches can help to identify areas for improvement and to guide the development of more advanced techniques and methods for sentiment analysis of social media data.

Other than that, Identifying the challenges faced in this area can help to guide the development of solutions and strategies to overcome these challenges, and to improve the performance of sentiment analysis on social media data.

Lastly, it is important to understand how SVM and lexicon-based approaches perform in social media to provide insights into the potential use of sentiment analysis of social media data. If SVM and lexicon-based approaches are found to be suitable for this purpose, it can have significant implications for researchers and practitioners. Understanding the suitability and limitations of these techniques can help inform the design and implementation of systems for predicting and analyzing social media sentiment.

Search Strategy

After formulating the research questions, we then searched two major databases namely, ScienceDirect and IEEE Xplore - for relevant research papers and literature reviews using keywords such as “sentiment analysis”, “literature review”, “systematic literature review”, “SLR” and “review”.

Study Selection (Exclusion and Inclusion Criteria

As part of the paper screening process, we established criteria for inclusion and exclusion. Inclusion criteria are used to determine which papers will be included in the review process, while exclusion criteria identify papers that will not be included. The inclusion criteria used in this study are as follows:

- The collected research papers and literature reviews are published between the year 2012 and 2022.
- The papers are written in English.
- The research must be related to sentiment analysis.

While the exclusion criteria are as follows:

- Articles and reviews that do not meet the inclusion criteria.
- The method used was not clearly described in the study.
- Studies that were unable to address the research question.

The search process and number of identified articles at each stage are depicted in Figure 7. To begin, we searched the ScienceDirect database using the keyword "sentiment analysis" and filtered the results to show literature reviews only, resulting in a total of 61 articles. In the IEEE Xplore database, we used the keyword "sentiment analysis" and filtered the results to show journal papers only and retrieved a total of 809 articles.

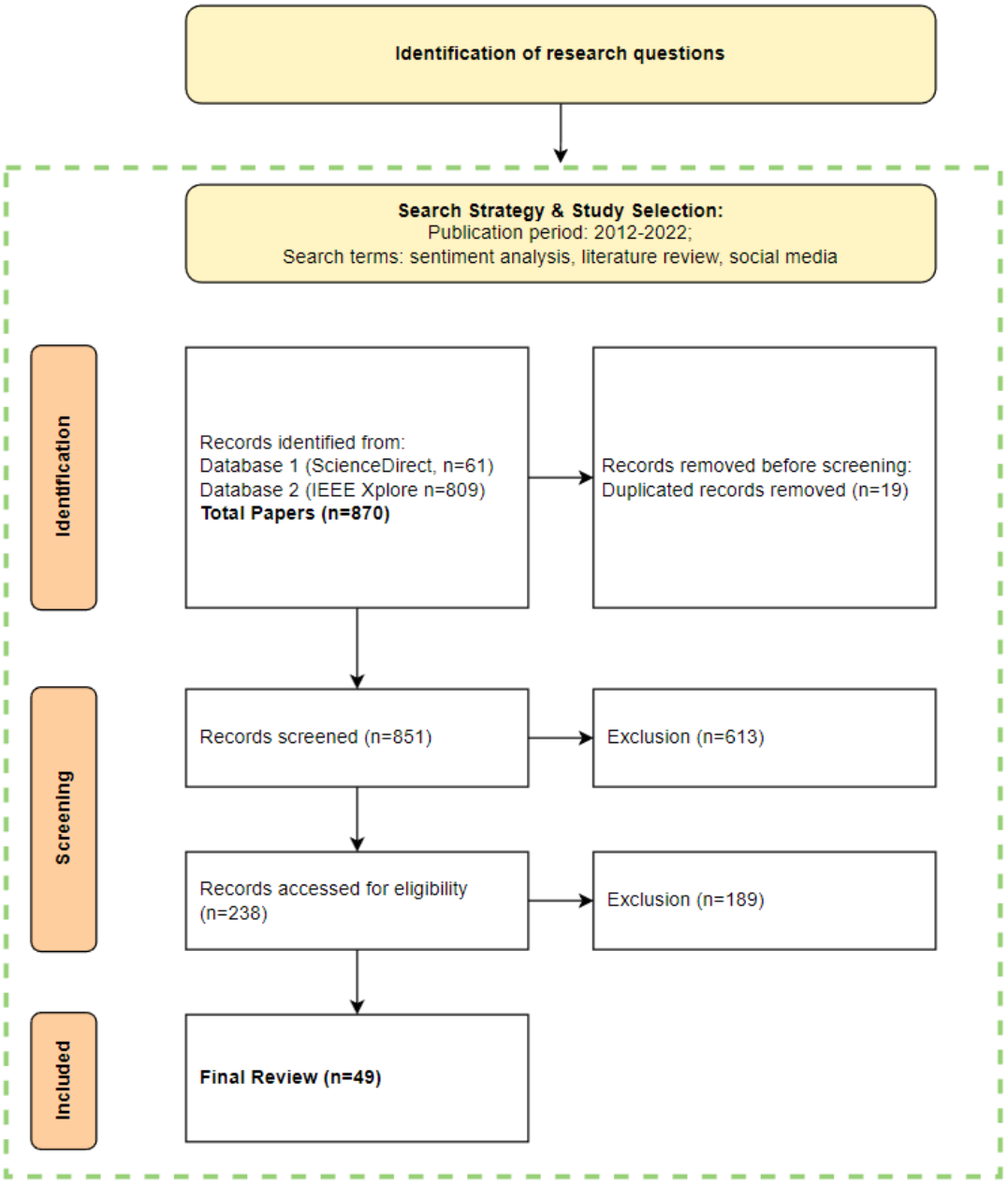


Figure 7. Survey methodology process

Quality Assessment

Maintaining the quality of the papers being reviewed is essential, and one way to do this is by carefully selecting the high-quality articles. High-quality articles can be identified by the number of citations of the paper or the reputation of the author.

Quality assessment ensures that only the most relevant and reliable articles are included in the review.

Data Extraction

During the data extraction phase, we gathered information from each reviewed paper, including year of publication, data set and collection techniques used, data preprocessing methods, sentiment analysis methods, and results. This information was collected and stored in a spreadsheet or document.

Data Synthesis

At this stage, all the papers that have passed the previous stages are ready to be analysed. Initially, 238 papers were identified, however, after implementing the inclusion and exclusion criteria, the list was narrowed to 49 papers that underwent review and analysis.

4.2 Implementation Methodology

4.2.1 Birds eye view of Implementation process

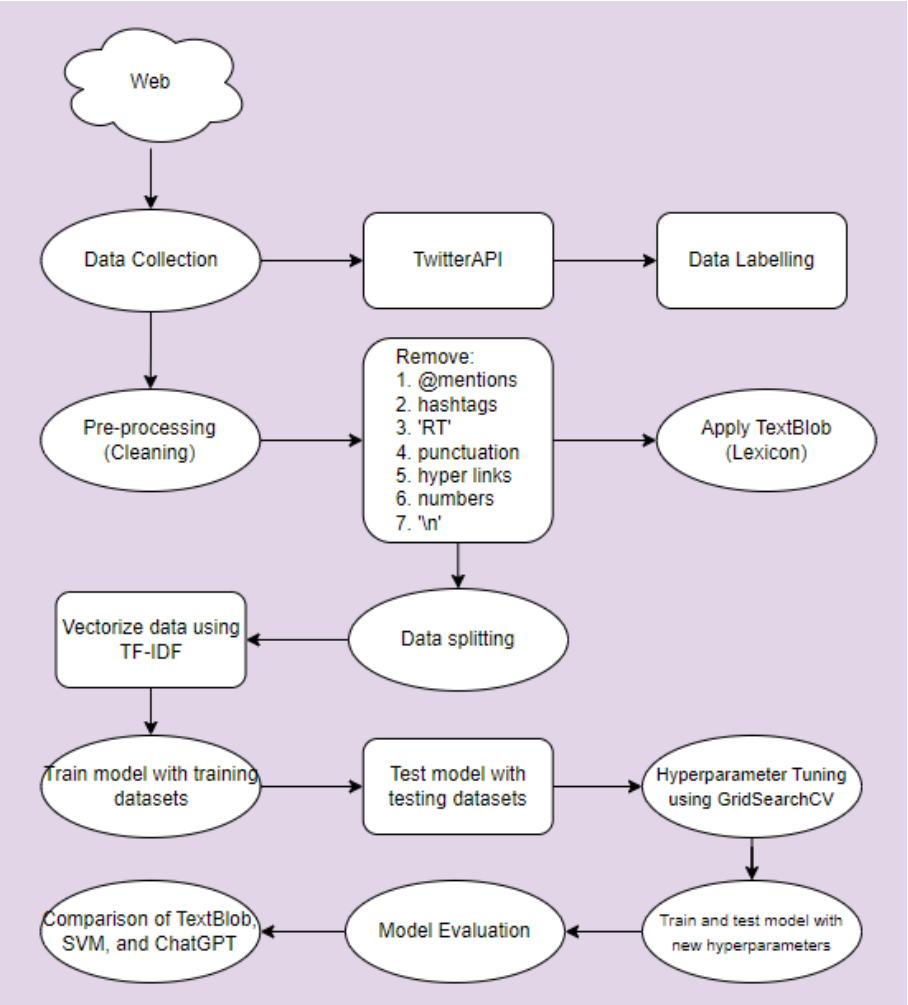


Figure 8. FYP2 implementation flowchart

The methodology employed for FYP2 is presented in **Figure 8,** outlining the flowchart of the research process.

First and foremost, the dataset for small cap and big cap companies are obtained from the API developed by Twitter called TwitterAPI. From there, the

necessary tweets are scraped by querying the company's ticker symbol and ETF symbols for small cap and big cap companies. Around 50 different companies are queried based on their categories (small cap or big cap) and the ETF symbol that tracks the overall performance of small cap and big cap companies. For example, 'IWM' is an index that measures the performance of approximately 2000 small cap companies. 'NDX', on the other hand, tracks the top 100 of the largest companies listed in the NASDAQ stock market. This data collection phase resulted of 5000 tweets for each category (small cap and big cap)

After collecting the datasets, manual labelling of the datasets was performed to assign the sentiments of the tweets to ensure the best accuracy. However, it is a time consuming process. The process of manual labelling the datasets took around 3 weeks to complete.

Next, data preprocessing such as data cleaning was performed. This process includes the removal of @mentions, hashtags, and 'RT' symbol, among other data preprocessing steps. The detailed explanation of data cleaning is explained in ***Chapter 5.2.3.*** This process is crucial for the implementation phase because it ensures that the techniques chosen perform at its best.

After cleaning the datasets, sentiment analysis using TextBlob (Lexicon) was applied to the cleaned tweets. Applying TextBlob (Lexicon) is relatively easy because it doesn't require any data splitting, data labelling or extract any features. At the end of this process, the polarity and subjectivity of the tweet is obtained.

Additionally, the sentiment (positive, negative, neutral) of each tweet is generated from the polarity scores.

Subsequently, ChatGPT is utilised to perform sentiment analysis. ChatGPT has its own API, however, in our case, we used the free version. In this process, each cleaned tweet was copied from the datasets and prompted manually to obtain the sentiments. However, due to the limitation of the free version, this process can be time consuming as ChatGPT only allows approximately 10-20 tweets each prompt.

Next, the cleaned tweets were then divided into training and testing sets, and TF-IDF feature extraction was applied to prepare the data for Support Vector Machine modelling. At first, the SVM model was tested and evaluated using the default parameters. After obtaining the performance of SVM using default parameters, hyperparameter tuning using GridSearchCV was performed to get the best parameter and performance of the SVM model.

Lastly, a final comparison of the results obtained from the three techniques (TextBlob, SVM, and ChatGPT) was conducted using appropriate metrics such as accuracy, F1-score, precision and recall.

5. Implementation

5.1 Implementation plan

		Week													
No.	Tasks	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1.	Data collection														
2.	Data Labelling														
3.	Data Preprocessing														
4.	Perform sentiment analysis														
5.	Performance and Evaluation														
6.	Research paper and report writing														

Figure 9. Gantt chart for final year project 2

FYP2 consists of several phases with their estimated timeframe as outlined in **Figure 9**. The first phase is the data collection phase (3 weeks) from Twitter using TwitterAPI. Next is data labelling which takes approximately 3 weeks to assign sentiment labels to the collected data of small cap and big cap companies. Data preprocessing will take approximately 3 weeks. During this period, the raw data collected from Twitter will undergo cleaning and preprocessing procedures such as removing noise, stop words, tokenization, and address other necessary data cleaning tasks. Next, perform sentiment analysis. In this phase, the implementation of TextBlob (Lexicon), Support Vector Machine, and ChatGPT techniques was carried

out on the cleaned tweets to predict sentiment associated with small cap and big cap companies in the United States Index. The performance and evaluation phase will be conducted simultaneously with the implementation phase. During this phase, the results of the three techniques will be compared using appropriate evaluation metrics. Lastly, from week 9 until the end of semester the focus will be shifted to research paper and FYP2 report writing. This phase involves documenting the entire project.

5.2 Implementation of Sentiment Analysis

This research is conducted using Jupyter Notebook with the Microsoft Visual Studio Code IDE with the help of TwitterAPI for data collection. A total of 10,000 tweets were collected, with 5,000 tweets for small cap companies and another 5,000 tweets for big cap companies.

5.2.1 Data Collection

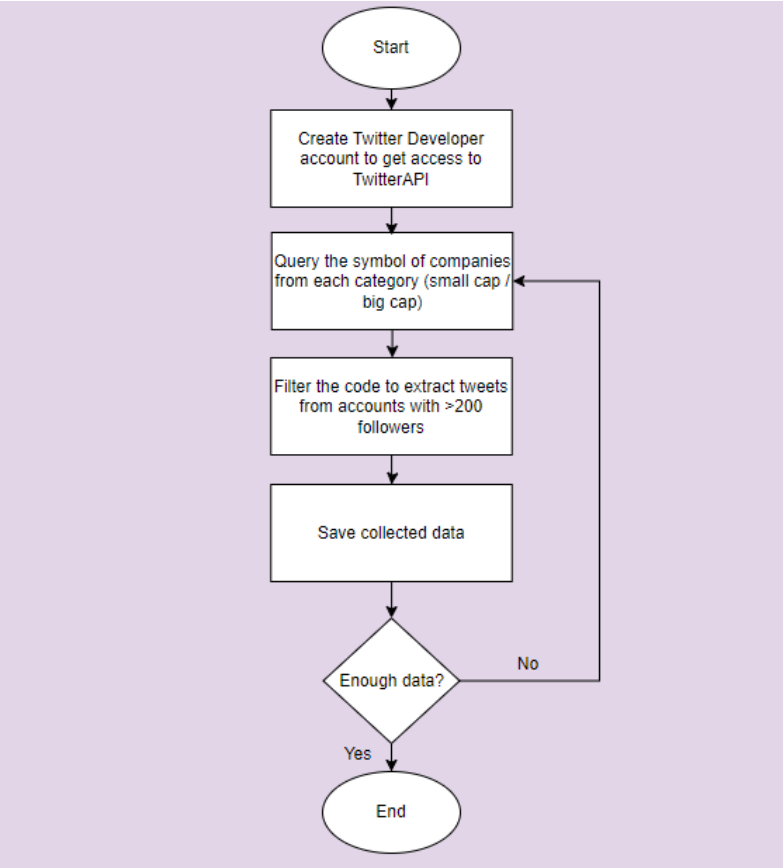


Figure 10. Flow of data collection

In this phase, TwitterAPI is used to find tweets related to companies from each category, namely, small cap and big cap companies. The necessary tweets are collected by querying the company's ticker symbol and ETF symbols. For example, '$TSLA', 'NDX', 'IWM', to name a few. In total, around 50 companies are queried based on their categories (small cap or big cap) and also some ETF symbol that tracks the overall performance of small cap and big cap companies. The code snippet for extracting tweets for small cap and big cap companies is depicted below**.** From the code snippet, is the process of extracting one ticker out of the many tickers/companies in the small cap category.

```
symbol = 'KIDS'
searchResults = api.search_tweets('$' + symbol, count=200, lang='en',
tweet_mode="extended")

# Create a list to store tweet information
tweets = []
for tweet in searchResults:
    user = tweet.user
    #measures taken to avoid bot
    if user.followers_count > 200:
        tweets.append({'Tweets': tweet.full_text, 'Symbol': symbol, 'Follower
Count': user.followers_count})

#create a dataframe from the new tweet
new_tweets = pd.DataFrame(tweets)

# Concatenate new tweet data with the existing smallcap dataframe
smallcap = pd.concat([smallcap,new_tweets])
```

Additionally, the data extraction process includes a filter to specifically get tweets from accounts with more than 200 followers. This measure is taken to ensure that we only get tweets that are less likely to be generated by bots. Despite Twitter's ongoing efforts to eliminate bots under the new leadership of the new CEO, Elon Musk, this measure is still taken to ensure the reliability of the collected tweets. This process is then repeated until we get the amount of tweets that we need in our datasets, in this case, 5,000 tweets for small cap and an additional 5,000 tweets for big cap with a total of 10,000 tweets.

5.2.2 Data Labelling

After collecting the datasets, manual labelling of the datasets was performed to assign the sentiments of the tweets to ensure the best accuracy. While this approach ensured the highest accuracy in sentiment classification, it is also proven to be time-consuming. The manual labelling process required significant attention to

detail, making the process approximately two to three weeks to complete. This manually labelled data can act as ground truth in order to evaluate the techniques applied in this research.

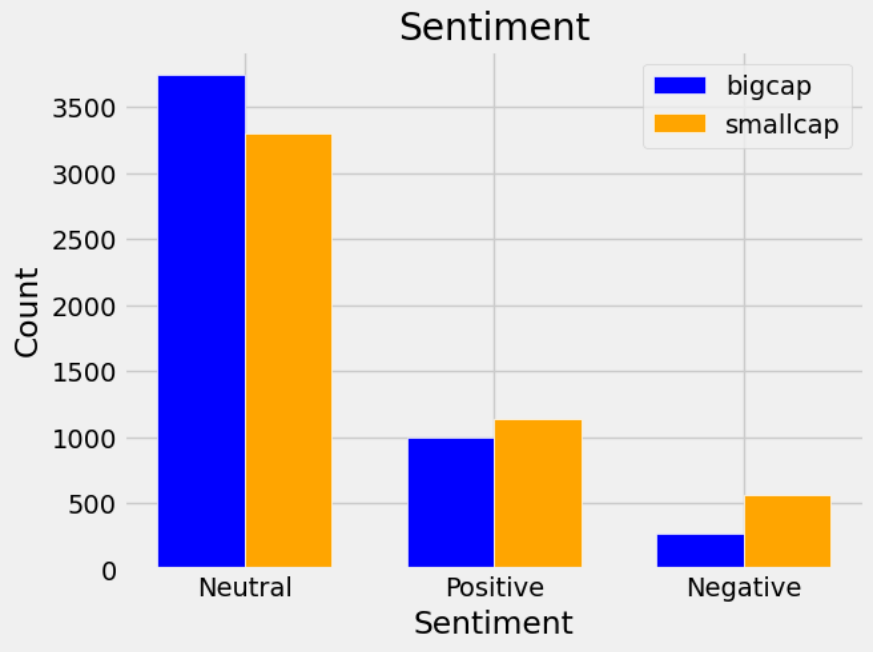


Figure 11. Sentiment distribution of big cap and small cap tweets

From the manual labelling process, it was found that for the small cap dataset, there were 1,139 positive tweets, 561 negative tweets, and 3,299 neutral tweets. Other than that, for big cap companies, there were 993 positive tweets, 265 negative tweets, and 3,742 neutral tweets. These results are depicted in **Figure X** which illustrates the distribution of sentiments of tweets for both big cap and small cap companies.

5.2.3 Data Preprocessing

To prepare the raw data collected from TwitterAPI, a series of data preprocessing steps are taken. The objective of these preprocessing steps was to

ensure that the data is cleaned and in the best format for optimal performance of the Support Vector Machine (SVM) model and TextBlob (Lexicon).

```
def cleanTxt(text):
  text = re.sub(r'@[A-Za-z0-9]+', '', text) # Remove @mentions
  text = re.sub(r'#', '', text) # Remove the '#' symbol
  text = re.sub(r'RT[\s]+', '', text) # Remove RT
  text = re.sub(r'[^\w\s]','',text) # Remove punctuation
  text = re.sub(r'https?:\/\/\S+', '', text) # Remove the hyperlink
  text = re.sub(r'\d+', '', text) # Remove numbers
  text = re.sub(r'\n', '', text) # remove \n
  text = text.lower()
  tokens = word_tokenize(text)
  tokens = [token.lower() for token in tokens if token.lower() not in
stop_words]
  lemmatizer = WordNetLemmatizer()
  tokens = [lemmatizer.lemmatize(token) for token in tokens]
  cleaned_text = " ".join(tokens)

  return cleaned_text
```

The code snippet above demonstrates the implementation of the data preprocessing steps applied on each tweet in the datasets. By applying this *'cleanTxt'* function, it ensures that all the tweets undergo consistent and thorough cleaning, optimising the quality and effectiveness of the sentiment analysis techniques.

Following are the steps involved in the cleaning phase:

1. ***Noise Removal:*** Various types of noise, including @ signs, hashtags, 'RT' symbols, hyperlinks, numbers, white spaces, and punctuations, were removed from the tweets.
2. ***Stop Words Removal:*** Stop words, such as articles, prepositions, and conjunctions (e.g., am, are, an, the, is), were eliminated. These words do not contribute significantly to sentiment classification and are commonly found in any text, including tweets.

3. ***Lowercasing:*** All sentences were converted to lowercase letters to ensure consistency in the text.
4. ***Tokenization:*** The process of tokenization was applied, which involves breaking down the text into smaller units, typically words. This step is crucial for subsequent natural language processing (NLP) tasks as it enables computer analysis, processing, and comprehension of the text.
5. ***Lemmatization:*** Lemmatization, a linguistic technique, was utilized to transform words into their base or root form. This process helps in standardising the words and improves the analysis and understanding of the text. For example, it converts "ate" to its base form, "eat," and reduces "looked" to "look."

5.2.3 TextBlob Implementation

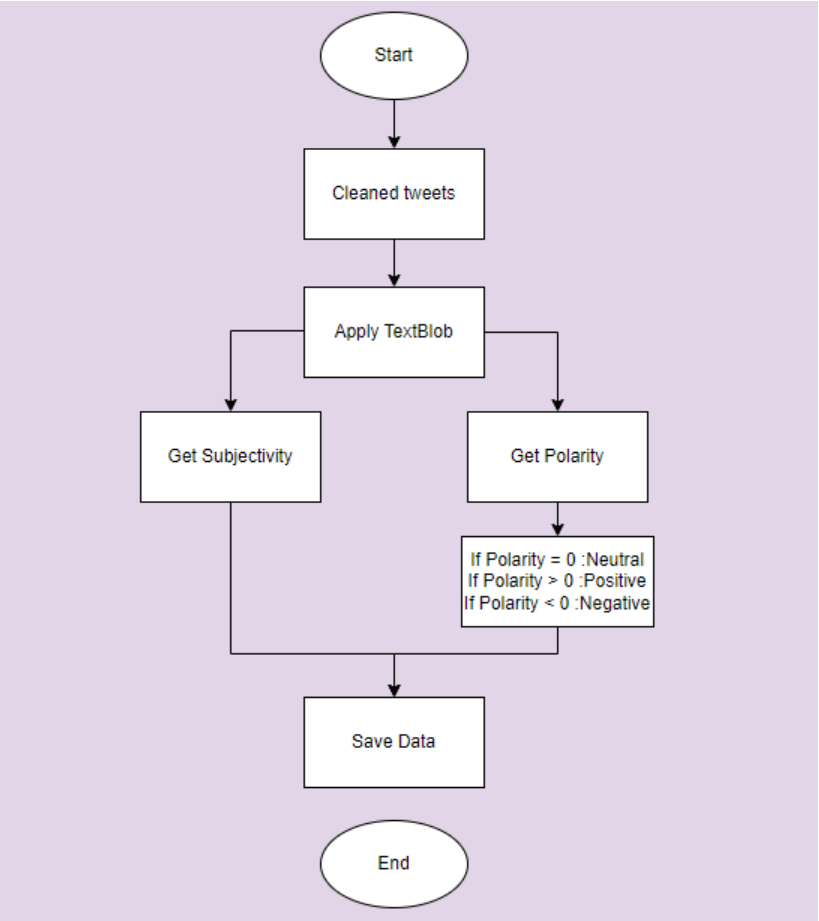


Figure 12. Flow of TextBlob

Figure 12 shows the flow of the application of TextBlob in performing sentiment analysis. After cleaning the tweets, TextBlob is applied for sentiment analysis to get the subjectivity and polarity of each tweet.

```
def getAnalysis(score):
  if score <0:
    return 'Negative'
  elif score == 0:
    return 'Neutral'
  else:
    return 'Positive'
```

Additionally, a function is then created to classify each tweet to their respective sentiments based on the polarity of the tweets. This function takes the polarity of the tweets obtained from TextBlob as input and assigns them to the corresponding sentiment label. If the polarity is less than 0, indicating a negative sentiment, if the polarity is equal to 0, denoting a neutral sentiment, and lastly, if the polarity is greater than 0, representing a positive sentiment. This classification function helps in categorising each tweet into its respective sentiment category.

5.2.4 ChatGPT Implementation

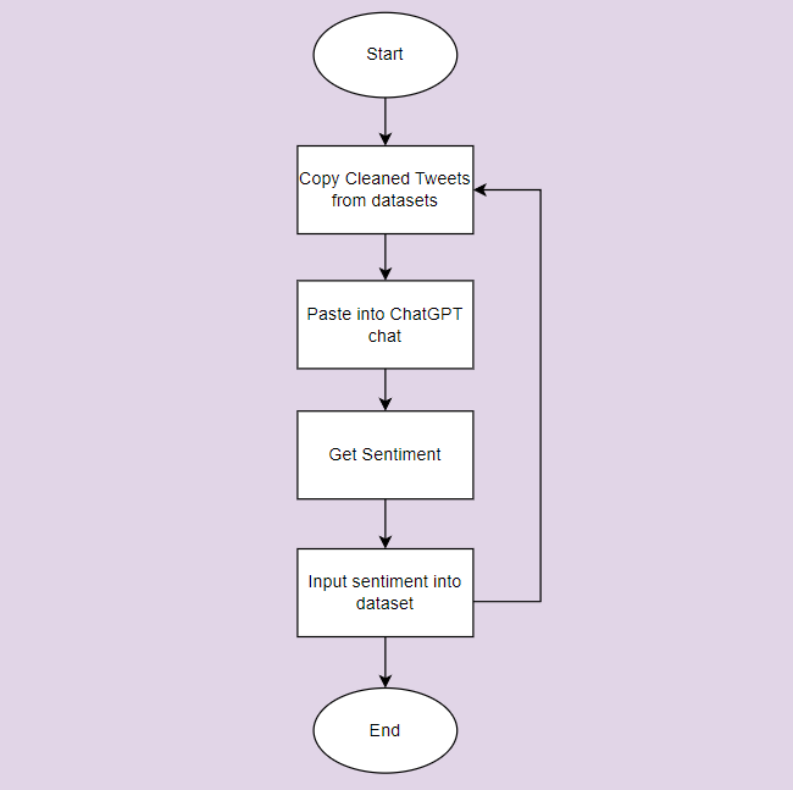


Figure 13. Flow of ChatGPT

In addition to TextBlob and Support Vector Machine (SVM), another technique that is utilised to perform sentiment analysis in this project is ChatGPT. As

we all know, recently, ChatGPT has gained significant attention for its advanced chatbot and language capabilities. **Figure 13** demonstrated the flow of using ChatGPT in performing sentiment analysis. In order to perform sentiment analysis using ChatGPT, each cleaned tweet was manually copied from the datasets and prompted into the chat interface of ChatGPT to obtain the sentiments. However, due to the limitations of the free version, this process can be time consuming as ChatGPT only allows approximately 10-20 tweets each prompt.

Figure 14. ChatGPT Sentiment

Figure 14 provides the screenshot that demonstrates how the sentiment was obtained from the prompt. After getting the results, the sentiments were manually inputted back into the datasets for further analysis and performance comparison with other sentiment analysis techniques used in this project.

5.2.5 Support Vector Machine (SVM) Implementation

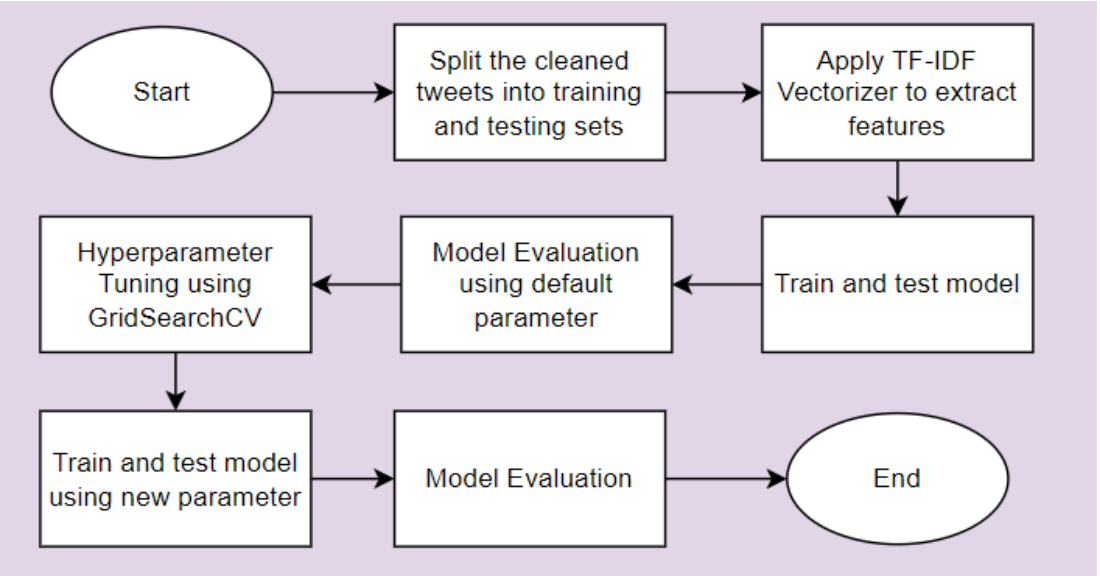


Figure 15. Flow of SVM

Another technique that is used in this project is Support Vector Machine (SVM). The implementation of SVM involves several steps as depicted in the flowchart shown above. Firstly, the cleaned and labelled tweet data is split into training and testing sets.

To extract features from the data, TF-IDF vectorizer is applied to calculate the importance of each word based on its frequency and rarity across the entire dataset. The resulting TF-IDF vectors serve as input to train the SVM model.

After that, the performance of the SVM model is evaluated using its default parameters. This acts as a baseline performance for sentiment analysis using SVM. Additionally, hyperparameter tuning is performed using GridSearchCV to further optimise the model. GridSearchCV systematically explores and tests different

combinations of hyperparameters of the model and finds the best set for the SVM model.

Once the best parameters are obtained through GridSearchCV, the SVM model was retrained and tested using these optimised parameters. Finally, the performance of sentiment analysis using the best SVM parameters is evaluated to assess the improvement of the model after performing hyperparameter tuning.

6 Evaluation of Findings

6.1 TextBlob Results

TextBlob was performed for sentiment analysis of small cap and big cap companies. The results were compared with the manually labelled sentiments to calculate the accuracy. The accuracy achieved for small cap companies using TextBlob was 44%, while for big cap companies, it was 47%. The lower accuracy is attributed to the complexity of stock market tweets, which often contain words that are not in the lexicon dictionary. **Figure 16** was created to compare the performance of TextBlob (Lexicon-based approach) against the manually labelled sentiments for both small cap and big cap companies.

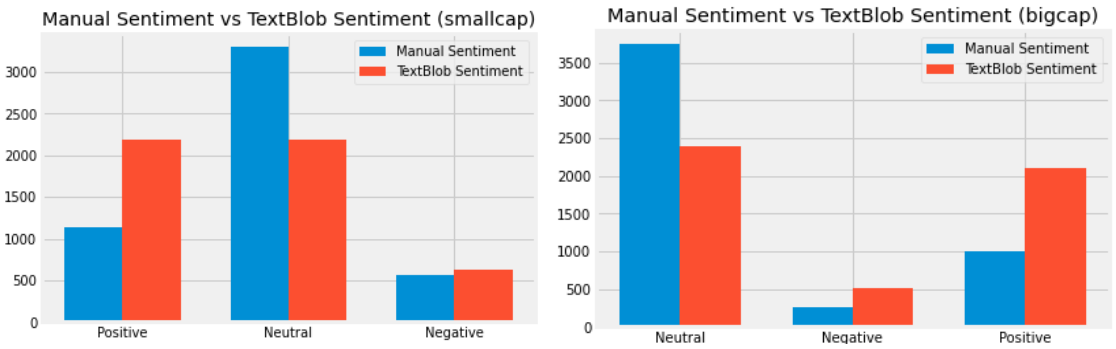


Figure 16. Manual vs. TextBlob sentiment distribution

Additionally, **Figure 17** and **Figure 18** represents the scatterplot of polarity vs. subjectivity that was generated to visualise the relationship between these two sentiment analysis metrics. Subjectivity represents the degree of subjectiveness in the text, while polarity indicates the sentiment orientation (positive, negative, or neutral). The scatter plot provides insights into how subjectivity and polarity are interconnected.

	precision	**recall**	**f1-score**	**Accuracy**
Small Cap	0.53	0.44	0.46	0.44
Big Cap	0.62	0.47	0.51	0.47

Table 1. Accuracy of TextBlob

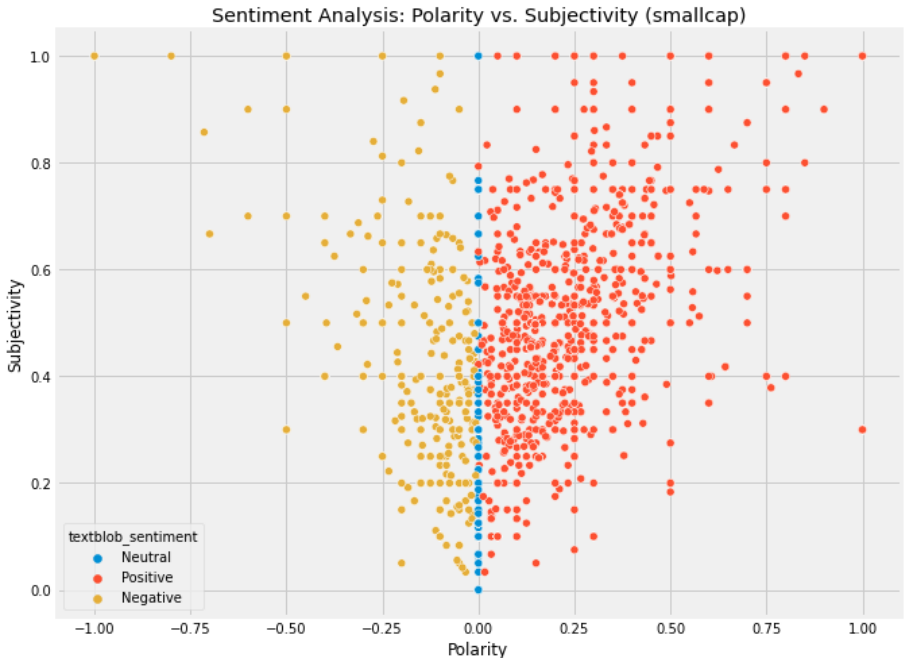


Figure 17. Polarity vs. Subjectivity (smallcap)

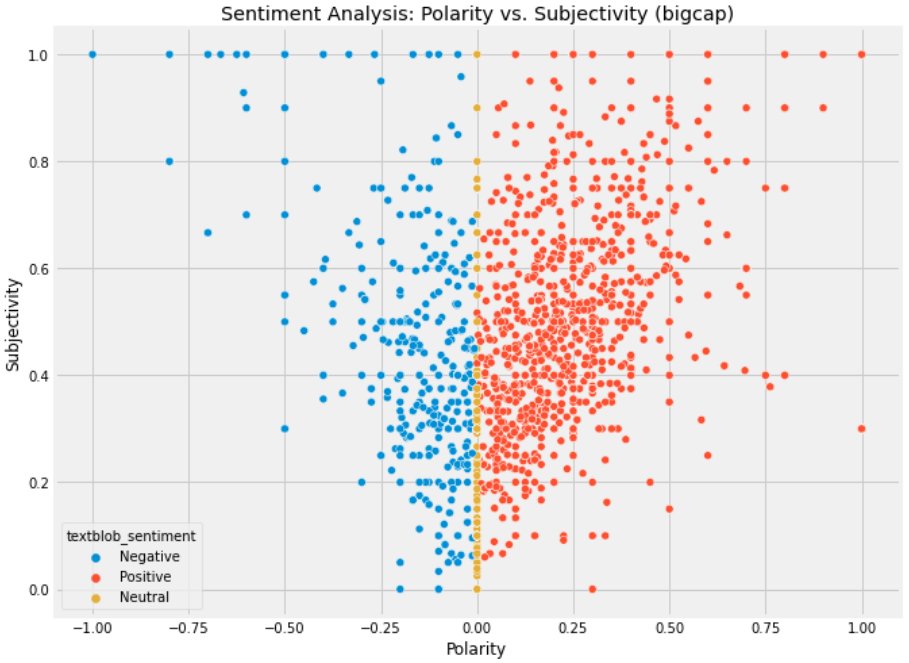


Figure 18. Polarity vs. Subjectivity (bigcap)

Word Clouds displaying the top words for each sentiment class (positive, neutral, negative) is depicted in **Figure 19** and **Figure 20**. These word clouds offer a visual

representation of the most frequently occurring words associated with each sentiment category.

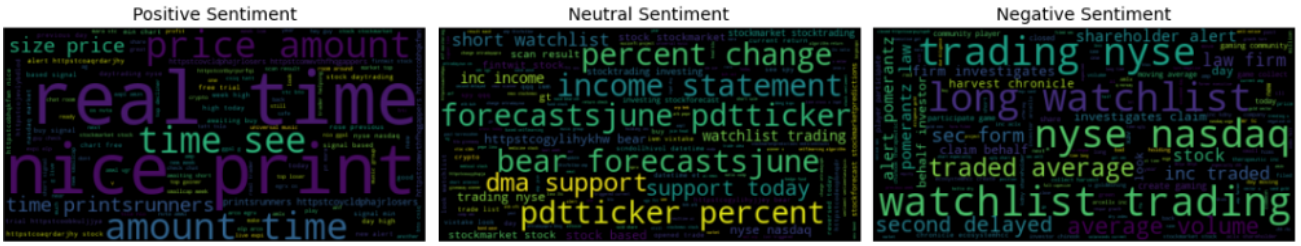


Figure 19. Small cap word cloud

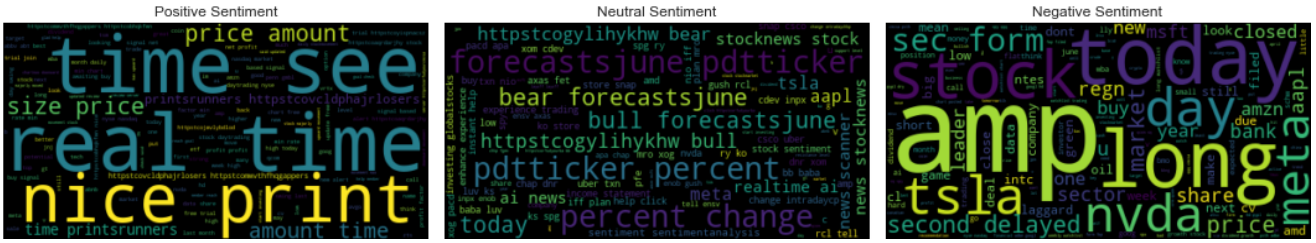


Figure 20. Big cap word cloud

6.2 ChatGPT Results

Additionally, the sentiment analysis results obtained from using ChatGPT, showcased promising accuracy, with 72% accuracy for small cap and 77% accuracy for big cap companies when compared to the manually labelled sentiment.

	precision	**recall**	**f1-score**	**Accuracy**
Small Cap	0.74	0.72	0.64	0.72
Big Cap	0.77	0.77	0.76	0.77

Table 2. Accuracy of ChatGPT

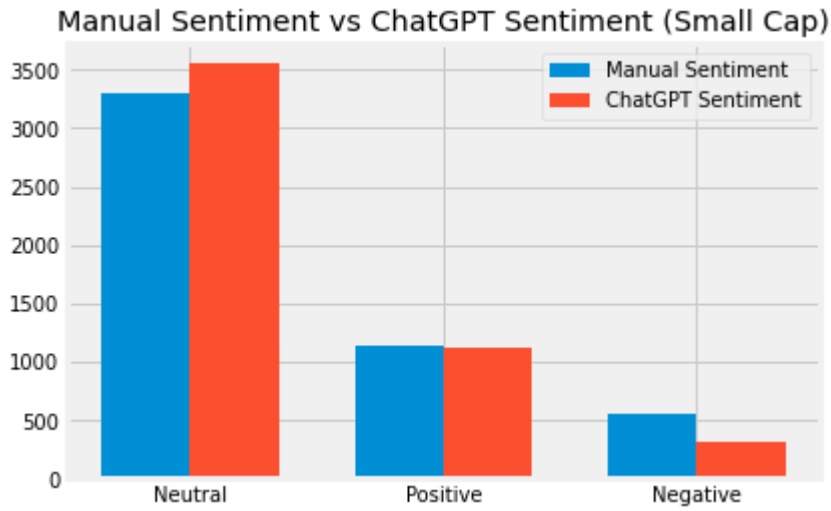


Figure 21. Manual vs ChatGPT sentiment distribution (small cap)

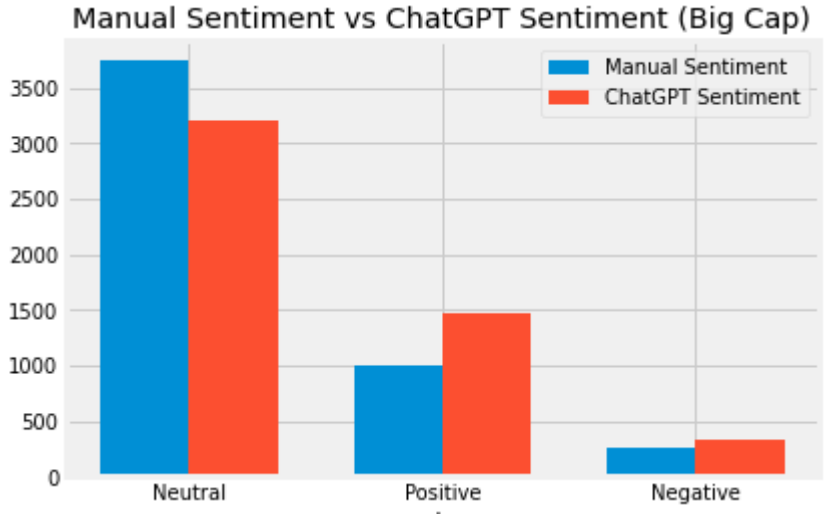


Figure 22. Manual vs ChatGPT sentiment distribution (big cap)

6.3 Support Vector Machine Results

The sentiment analysis results obtained from SVM were also evaluated. Using default parameters, the SVM model achieved an accuracy of 78.3%, precision of 0.81, recall of 0.78 and F1-score of 0.76 for small cap companies and the model achieved an accuracy of 83%, precision of 0.84, recall of 0.83, and F1-score of 0.80 for big cap companies. The performance can be considered relatively good,

indicating the effectiveness of the SVM model in sentiment analysis using its default parameters.

However, after conducting hyperparameter tuning using GridSearchCV, the performance of the SVM model is significantly improved. For small cap companies, the tuned model achieved an accuracy of 79.3% precision of 0.79, recall of 0.79, and F1-score of 0.78. The improved performance of the tuned SVM model performed similarly with big cap companies, the accuracy, precision, recall and F1-score is 85%, 0.85, 0.85, and 0.83 respectively.

	C	kernel	gamma	Accuracy(%) Default	Accuracy(%) Tuned
Small Cap Default	1.0	rbf	scale	78.3	79.3
Big Cap Default	10	rbf	scale	83	85

Table 3. Hyperparameter settings for SVM

	precision	recall	f1-score	Accuracy
Small Cap Default	0.79	0.64	0.69	0.78
Big Cap Default	0.84	0.83	0.80	0.83
Small Cap Tuned	0.79	0.68	0.72	0.79
Big Cap Tuned	0.85	0.85	0.83	0.85

Table 4. Performance of SVM model (Default vs Tuned)

The best hyperparameters identified using GridSearchCV for the SVM model were C=10, gamma='scale', and kernel='rbf', which further enhanced the accuracy of the model. Bar charts comparing the tuned SVM results with the manually labelled sentiments were created for both small cap and big cap companies, providing visual insights into the model classification performance.

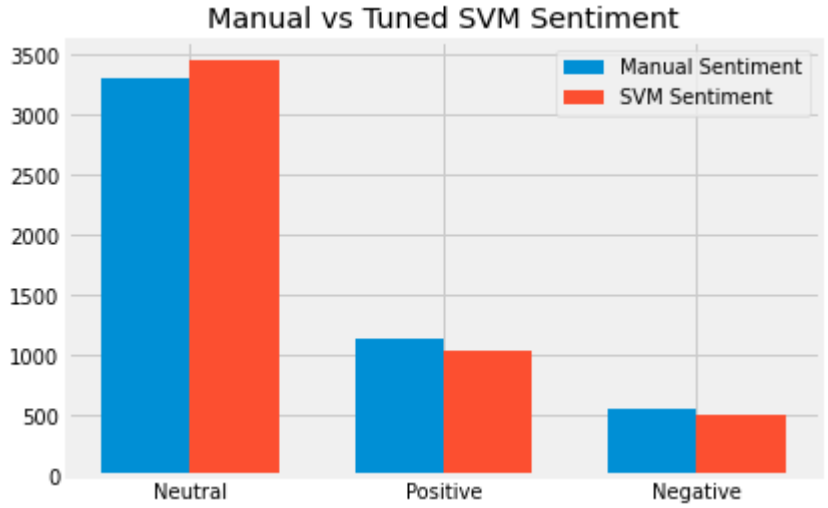


Figure 23. Manual vs. Tuned SVM sentiments distribution(small cap)

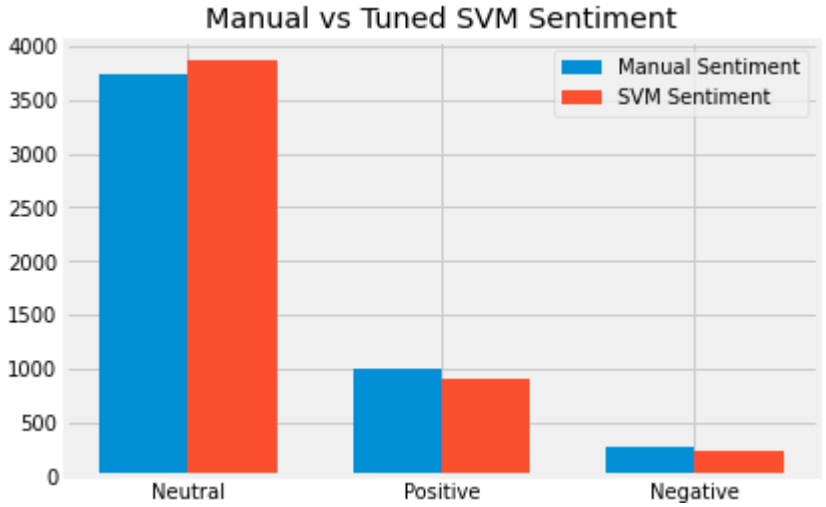


Figure 24. Manual vs. Tuned SVM sentiments distribution (big cap)

6.4 Performance Comparison

Table 5 provides the comparison of performance of the different sentiment analysis approaches used in this project when compared with the manually labelled sentiment. TextBlob, a lexicon-based approach, achieved the lowest accuracy at 44% for small cap companies and 47% for big cap companies. Subsequently, SVM achieved an accuracy of 78.3% for small cap and 83% accuracy for big cap companies. However, after hyperparameter tuning, the tuned SVM model improved to 79.3% for small cap companies and 85% accuracy for big cap companies. Lastly, ChatGPT demonstrated an accuracy of 72% for small cap and 77% accuracy for big cap companies when compared to the manually labelled sentiment.

Methods	Cap Size	Accuracy (%)
TextBlob	Small cap	44
	Big cap	47
SVM	Small cap	78.3
	Big cap	83
SVM (Tuned)	Small cap	79.3
	Big cap	85
ChatGPT	Small cap	72
	Big cap	77

Table 5. Performance comparison

6.5 Discussion of Results

By evaluating the 3 techniques used for sentiment analysis in this project, we can get valuable insights based on their performances. From this analysis, we can identify which method is the best for performing sentiment analysis. TextBlob, a

lexicon-based approach, achieved the lowest accuracy amongst the other 2 techniques used for evaluation in this project, highlighting its simplicity but limited effectiveness in capturing the complexity of stock market tweets. However, TextBlob can serve as a baseline technique and provide a simple interface for sentiment analysis due to its ease of use of datasets that does not require any sentiment label.

The Support Vector Machine (SVM) model demonstrated the best performance out of the all techniques used, even with default parameters. Furthermore, by performing hyperparameter tuning through GridSearchCV, the SVM model achieved further improvements in terms of its performance.

Additionally, the sentiment analysis results obtained using ChatGPT, showcased promising accuracy due to its advanced language capabilities and the ability to understand and generate human-like texts. Its performance, although slightly lower than SVM, indicates its effectiveness in classifying sentiments of stock market tweets without any labelled data. However, it is important to note that the free version of ChatGPT has its limitations in performing sentiment analysis on a large dataset. This is because it only allows a limited number of tweets per prompt. This limitation of the free version makes the process of sentiment analysis to be time-consuming.

Comparing the results of SVM, TextBlob, and ChatGPT with the manually labelled sentiments, it is evident that these approaches have their own strengths and weaknesses. SVM, being a machine learning algorithm, shows the ability to capture

more complex patterns and achieve higher accuracy compared to TextBlob and ChatGPT. SVM also has the flexibility to tune and find the best parameter in order to further enhance its performance, allowing for potential optimization and improved sentiment prediction.

7 Conclusion

In conclusion, this final year project has successfully accomplished its research objectives. Firstly, through extensive literature review and analysis, a comprehensive understanding of the different techniques used in sentiment analysis, mainly Support Vector Machine and Lexicon-based approach was obtained.

The implementation phase of this project demonstrated the successful application of these techniques in performing sentiment analysis using the collected Twitter data, with SVM performing the best out of the 3 techniques tested in this project. Insights such as the sentiment trends in the small cap and big cap companies in the US indexes were gained from this analysis, aiding decision-making processes in the financial sector.

Throughout this project, several major lessons have been learned. One of them is the importance of preprocessing and data cleaning was highlighted for obtaining accuracy sentiment analysis results. Other than that, hyperparameter tuning was found to be crucial for optimising the performance of machine learning models such as SVM. Additionally, the potential of large language models like ChatGPT

was recognised. Although there are some limitations of the free version, the performance of ChatGPT in performing sentiment analysis provides promising results.

As the field of sentiment analysis and the expansion of social media data keeps on increasing every single day, future works should focus on exploring hybrid methods that combine the strength of different techniques to further enhance sentiment analysis accuracy. Additionally, investigating a more advanced or paid version of ChatGPT could provide more capabilities for sentiment analysis tasks. Other than that, expanding the project to include sentiment analysis for additional indexes and larger datasets would provide a broader perspective.

Throughout the final year project, the utilisation of data collection, preprocessing, machine learning, natural language processing, and comparative evaluation contribute to the success of this project, highlighting the program specific skills and knowledge.

In summary, this final year project has provided valuable insights into sentiment analysis techniques, their performance in the financial domain, specifically stock market Twitter data, and future improvement possibilities. Applying program specific skills gave a deeper understanding of sentiment analysis, contributing to both academic knowledge and practical application in the field.

References

[1] Dangi, D., Dixit, D. K., & Bhagat, A. (2022). Sentiment analysis of COVID-19 social media data through machine learning. Multimedia Tools and Applications, 81(29), 42261-42283. doi:10.1007/s11042-022-13492-w

[2] Katarya, R., Nath, G. A., Singhal, D., & Shukla, A. (2022). Analysing the twitter sentiments in COVID-19 using machine learning algorithms. Paper presented at the Proceedings - IEEE International Conference on Advances in Computing, Communication and Applied Informatics, ACCAI 2022, doi:10.1109/ACCAI53970.2022.9752511

[3] S. Naz, A. Sharan and N. Malik, "Sentiment Classification on Twitter Data Using Support Vector Machine," 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, Chile, 2018, pp. 676-679, doi: 10.1109/WI.2018.00-13.

[4] N. Hasanati, Q. Aini and A. Nuri, "Implementation of Support Vector Machine with Lexicon Based for Sentiment Analysis on Twitter," *2022 10th International Conference on Cyber and IT Service Management (CITSM)*, Yogyakarta, Indonesia, 2022, pp. 1-4, doi: 10.1109/CITSM56380.2022.9935887.

[5] Abd, D., Abbas, A., & Sadiq, A. (2021). Analyzing sentiment system to specify polarity by lexicon-based. *Bulletin of Electrical Engineering and Informatics, 10*(1), 283-289. doi:https://doi.org/10.11591/eei.v10i1.2471

[6] Khatoon S., Abu Romman L. and Hasan M.M, A domain-independent automatic labeling system for large-scale social data annotation using lexicon and web-based augmentation. Inf. Technol. Control, 49 (1) (2020), pp. 36-54, 10.5755/j01.itc.49.1.23769

[7] Khan, Muhammad Taimoor, Mehr Durrani, Armughan Ali, Irum Inayat, Shehzad Khalid, and Kamran Habib Khan. (2016) “Sentiment
Analysis and The Complex Natural Language.” Complex Adaptive Systems Modeling 4 (1): 2

[8] Akter, Sanjida, and Muhammad Tareq Aziz. (2016) “Sentiment Analysis on Facebook Group Using Lexicon Based Approach”, in the 2016
3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)

[9] H. Karamollaoğlu, İ. A. Doğru, M. Dörterler, A. Utku and O. Yıldız, "Sentiment Analysis on Turkish Social Media Shares through Lexicon Based Approach," *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, Sarajevo, Bosnia and Herzegovina, 2018, pp. 45-49, doi: 10.1109/UBMK.2018.8566481.

[10] Accenture (2020). Artificial Intelligence: The Future of Business. Accenture. https://www.accenture.com/us-en/insights/artificial-intelligence/ai-future-of-business

[11] Qianwen Ariel Xu, Victor Chang, Chrisina Jayne (2022). A systematic review of social media-based sentiment analysis: Emerging trends and challenges. Decision Analytics Journal, Volume 3, 100073, ISSN 2772-6622.

[12] M. F. R. Abu Bakar, N. Idris, L. Shuib and N. Khamis, "Sentiment Analysis of Noisy Malay Text: State of Art, Challenges and Future Work," in *IEEE Access*, vol. 8, pp. 24687-24696, 2020, doi: 10.1109/ACCESS.2020.2968955.

[13] M. Bouazizi and T. Ohtsuki, "Multi-class sentiment analysis on twitter: Classification performance and challenges," in *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 181-194, September 2019, doi: 10.26599/BDMA.2019.9020002.

[14] Dang, C. N., Moreno-García, M. N., & De la Prieta, F. (2021). Using Hybrid Deep Learning Models of Sentiment Analysis and Item Genres in Recommender Systems for Streaming Services. *Electronics*, *10*(20), 2459. https://doi.org/10.3390/electronics10202459

[15] Statista (2021), Number of social media users worldwide. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[16] Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of Information Management Data Insights*, *1*(2). https://doi.org/10.1016/j.jjimei.2021.100019

[17] V. Ramanathan and T. Meyyappan, "Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism," 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), Muscat, Oman, 2019, pp. 1-5, doi: 10.1109/ICBDSC.2019.8645596.

[18] Christiane Fellbaum, & George Miller. (1998). *WordNet : An Electronic Lexical Database*. MIT Press.

8. Appendices

8.1 Appendix A

8.1.1 FYP Meeting Log 1

TPT3101 Final Year Project (FYP1) Meeting Log
Trimester 1, 2022/23 (Trimester ID:2210)

Meeting Date: 27/10/2022	**Meeting No.: 3**
Meeting Mode: Online	
Project ID: 2189	**Project Type: Research & application based**
Project Title : Sentiment Analysis of Small Cap and Big Cap companies in United States Indexes	
Student ID : 1181103271	**Student Name: Muhammad Khairulrazi Bin Mohd Riza**
Student Programme and Specialisation: **Bachelor's of Computer Science (Hons) Data Science**	
Supervisor Name: Nathar Shah Packier Mohammad	**Co-Supervisor Name:** **(if applicable)**
Collaborating Company: **(if applicable)**	**Company Supervisor Name:** **(if applicable)**

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept / Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- **Read approximately 20-30 literature review articles.**
- **Summarize articles in a tabular format.**
- **Did survey methodology.**
- **Picked a technique for the sentiment analysis project.**
- **Explain how did all the articles do their survey methodology so that we can put that into our context.**
- **Figured out the research questions.**

2. WORK TO BE DONE
[Please write the details of the work to be done, before the next meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept/ Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- **Enhance research questions. (find deeper and more meaningful questions)**
- **Citations.**
- **Explanation of why we choose our techniques.**
- **Improve our survey methodology.**
- **Make a comparison of the general idea and the specific techniques that we chose for this project.**
- **Find more articles to help with development.**

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

Problem encountered - Not finding enough literature review articles.
Solution - Insert a normal research article into the mix.

Problem encountered - Poor survey methodology.
Solution - Explain in more detail and include charts.

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

.....	.....
Supervisor's Signature	Student's Signature
.....	.....
Co-Supervisor's Signature **(if applicable)**	Company Supervisor's Signature **(if applicable)**

IMPORTANT NOTES TO STUDENTS:

1. Items 1 – 3 are to be completed by the students prior to the meeting. Item 4 is to be completed by the supervisor / co-supervisor / company supervisor.

2. Student has to upload the soft copies of the meeting logs in Google Classroom and also attach them along with interim (FYP1) report.
   Minimum requirement is SIX Meeting Logs (Period: Week 4 to Week 14). Students can have fortnightly meetings with the supervisor.

3. Log sheets provide the basis for evaluating the General Effort (Project Management, Attitude, and Technical Competency) of the student, by the supervisor and also for checking the attendance requirement of the student, by the FYP Committee.

   This also provide the student with feedback from the supervisor / co-supervisor / company supervisor on the tasks done and provide the plan for the upcoming tasks. This can provide the motivation for the student to give consistent and efficient effort throughout the period of FYP.

4. Student who fails to meet the minimum requirement (six nos.) of log sheets will not be allowed to submit FYP report.

8.1.2 FYP Meeting Log 2

TPT3101 Final Year Project (FYP1) Meeting Log
Trimester 1, 2022/23 (Trimester ID:2210)

Meeting Date: 24th November 2022	**Meeting No.:** 5
Meeting Mode: Physical Meeting	
Project ID: 2189	**Project Type:** Research-based/Application-based
Project Title : Sentiment Analysis of Small Cap and Big Cap companies in United States	
Student ID : 1181103271	**Student Name:** MUHAMMAD KHAIRULRAZI BIN MOHD RIZA
Student Programme and Specialisation: Bachelor's of Computer Science (Hons) Data Science	
Supervisor Name: Nathar Shah Packier Mohammad	**Co-Supervisor Name:** (if applicable)
Collaborating Company: (if applicable)	**Company Supervisor Name:** (if applicable)

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept / Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Read articles for my techniques used for this project
- Done my survey methodology
- Explained why did I choose my specific techniques
- Explained why did I question my research questions
- Enhanced my research questions
- Found several articles that can guide for my development

2. WORK TO BE DONE
[Please write the details of the work to be done, before the next meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept/ Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Put everything into context/ research template
- Start doing the final literature review

3. PROBLEMS ENCOUNTERED AND SOLUTIONS
[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- Time limitation as busy with other subject's work and projects
- Finding articles for specific techniques
- Time limitation for readings

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

.....	.....
Supervisor's Signature	Student's Signature
.....	.....
Co-Supervisor's Signature **(if applicable)**	Company Supervisor's Signature **(if applicable)**

IMPORTANT NOTES TO STUDENTS:

1. Items 1 – 3 are to be completed by the students prior to the meeting. Item 4 is to be completed by the supervisor / co-supervisor / company supervisor.
2. Student has to upload the soft copies of the meeting logs in Google Classroom and also attach them along with interim (FYP1) report.
   Minimum requirement is SIX Meeting Logs (Period: Week 4 to Week 14). Students can have fortnightly meetings with the supervisor.
3. Log sheets provide the basis for evaluating the General Effort (Project Management, Attitude, and Technical Competency) of the student, by the supervisor and also for checking the attendance requirement of the student, by the FYP Committee.

   This also provide the student with feedback from the supervisor / co-supervisor / company supervisor on the tasks done and provide the plan for the upcoming tasks. This can provide the motivation for the student to give consistent and efficient effort throughout the period of FYP.
4. Student who fails to meet the minimum requirement (six nos.) of log sheets will not be allowed to submit FYP report.

8.1.3 FYP Meeting Log 3

FACULTY OF
COMPUTING
& INFORMATICS

TPT3101 Final Year Project (FYP1) Meeting Log
Trimester 1, 2022/23 (Trimester ID:2210)

Meeting Date: 15th December 2022	Meeting No.: 8
Meeting Mode: Physical	
Project ID: 2189	Project Type: Research-based/Application-based
Project Title : Sentiment analysis of Small Cap and Big Cap companies in United States	
Student ID : 1181103271	Student Name: Muhammad Khairulrazi Bin Mohd Riza
Student Programme and Specialisation: Bachelor's of Computer Science (Hons) Data Science	
Supervisor Name: Nathar Shah Packier Mohammad	Co-Supervisor Name: (if applicable)
Collaborating Company: (if applicable)	Company Supervisor Name: (if applicable)

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept / Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Extract from template draft into articles
- Enhance research questions
- Explained why did I choose my specific techniques
- Find similarities from other articles but for our context

2. WORK TO BE DONE
[Please write the details of the work to be done, before the next meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept/ Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Complete full research articles
- Find the domains of the chosen techniques
- How does the techniques evolve over time

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- Time limitations
- Limited research articles for sentiment analysis from social media

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

No Comment

.....
Supervisor's Signature

.....
Student's Signature

.....
Co-Supervisor's Signature
(if applicable)

.....
Company Supervisor's Signature
(if applicable)

8.1.4 FYP Meeting Log 4

TPT3101 Final Year Project (FYP1) Meeting Log
Trimester 1, 2022/23 (Trimester ID:2210)

Meeting Date: 9th January 2023	**Meeting No.: 10**
Meeting Mode: Physical	
Project ID: 2189	**Project Type:** Research-based/Application-based
Project Title : Sentiment analysis of Small Cap and Big Cap companies in United States	
Student ID : 1181103271	**Student Name: Muhammad Khairulrazi Bin Mohd Riza**
Student Programme and Specialisation: Bachelor's of Computer Science (Hons) Data Science	
Supervisor Name: Nathar Shah Packier Mohammad	**Co-Supervisor Name:** (if applicable)
Collaborating Company: (if applicable)	**Company Supervisor Name:** (if applicable)

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept / Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Full literature review is almost done
- Enhanced Literature review
- Enhanced research questions
- Explained the domains of sentiment analysis techniques

2. WORK TO BE DONE
[Please write the details of the work to be done, before the next meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept/ Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Some short forms change to long form (for user to understand)
- Improve exclusion criteria (in survey methodology)
- Improve "process" graph in Survey Methodology
- Embed "Quality assessment" in survey methodology section into graph (figure 1)
- Show result of data extraction from articles in a table (show their domains, model performance, techniques used and challenges)
- Relate data synthesis to research questions

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- Time management - Busy with other projects and assignments, haven't got the time to complete literature review paper

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

- no comment

.....
Supervisor's Signature

.....
Student's Signature

8.1.5 FYP Meeting Log 5

TPT3101 Final Year Project (FYP1) Meeting Log
Trimester 1, 2022/23 (Trimester ID:2210)

Meeting Date: 12th January 2023	**Meeting No.:** 12
Meeting Mode: Online	
Project ID: 1181103271	**Project Type:** Research-based/Application-based
Project Title : Sentiment analysis of Small Cap and Big Cap companies in United States	
Student ID : 1181103271	**Student Name: Muhammad Khairulrazi Bin Mohd Riza**
Student Programme and Specialisation: Bachelor's of Computer Science (Hons) Data Science	
Supervisor Name: Nathar Shah Packier Mohammad	**Co-Supervisor Name:** (if applicable)
Collaborating Company: (if applicable)	**Company Supervisor Name:** (if applicable)

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept / Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Changed short form to long form (for readers to understand)
- Improved exclusion criteria
- Finished Overall survey of sentiment analysis (Chapter 2)

2. WORK TO BE DONE
[Please write the details of the work to be done, before the next meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept/ Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Survey of sentiment analysis technique for SVM (Chapter 3)
- Survey of sentiment analysis technique for Lexicon (Chapter 4)
- Improve "process" graph in Survey methodology
- Add more figures in the paper
- Relate data synthesis in survey methodology to research questions
- Show results in a table (in progress)

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- Time constraint/ time management - Busy with other projects and assignments, haven't got the time to complete literature review paper on time.

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

- no comment

.....
Supervisor's Signature

.....
Student's Signature

8.1.6 FYP Meeting Log 6

FACULTY OF COMPUTING & INFORMATICS

TPT3101 Final Year Project (FYP1) Meeting Log
Trimester 1, 2022/23 (Trimester ID:2210)

Meeting Date: 26th January 2023	**Meeting No.: 13**
Meeting Mode: Online	
Project ID: 2189	**Project Type:** **Research-based/Application-based**
Project Title : Sentiment analysis of Small Cap and Big Cap companies in United States	
Student ID : **1181103271**	**Student Name:** **Muhammad Khairulrazi Bin Mohd Riza**
Student Programme and Specialisation: **Bachelor's of Computer Science (Hons) Data Science**	
Supervisor Name: **Nathar Shah Packier Mohammad**	**Co-Supervisor Name:** **(if applicable)**
Collaborating Company: **(if applicable)**	**Company Supervisor Name:** **(if applicable)**

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept / Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Completed FYP1 interim report

2. WORK TO BE DONE
[Please write the details of the work to be done, before the next meeting]

Tasks: Problem Formulation and Project Planning / Background Study or Literature Review / Requirement Analysis or Theoretical Framework / Design or Research Methodology / Prototype Development or Proof of Concept/ Draft Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Fine tuning some formatting in FYP1 interim report
- Complete list of tables and figures
- Attach meeting logs in report

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- Time limitations
- Personal problems

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

No Comment

Supervisor's Signature

Student's Signature

8.1.7 FYP Meeting Log 1

MULTIMEDIA UNIVERSITY | FACULTY OF COMPUTING & INFORMATICS

TPT3101 Final Year Project (FYP2) Meeting Log
Trimester 2, 2022/23 (Trimester ID:2220)

Meeting Date: 12/4/2023	**Meeting No.:** 4
Meeting Mode: Physical meeting	
Project ID: 2189	**Project Type:** Research-based
Project Title : Sentiment analysis of Small Cap and Big Cap companies in United States Indexes	
Student ID : 1181103271	**Student Name:** MUHAMMAD KHAIRULRAZI BIN MOHD RIZA
Student Programme and Specialisation: Bachelor of Computer Science (Data Science)	
Supervisor Name: Nathar Shah Packier Mohammad	**Co-Supervisor Name:** (if applicable)
Collaborating Company: (if applicable)	**Company Supervisor Name:** (if applicable)

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Research objective for fyp2 research paper
- Created flowchart of how to implement the project
- Identify the tools needed to complete the project
- Picked 2 research paper as a guidance for this project
- Data collection
- Identified some of the challenges of this project

2. WORK TO BE DONE
[Please write the details of the work to be done, before the next meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Research write up/formal technical write up
- Choose how to do the techniques (parallel or sequential)
- Data Preprocessing (store data in readable format/cleaning/formatting)

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- Time limitation
- Personal issues
- Data Collection (trouble using TwitterAPI for data scraping)

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

.....
Supervisor's Signature

.....
Student's Signature

IMPORTANT NOTES TO STUDENTS:

1. Items 1 – 3 are to be completed by the students prior to the meeting. Item 4 is to be completed by the supervisor / co-supervisor / company supervisor.

2. Student must upload the soft copies of the meeting logs in Google Classroom and attach them along with final (FYP2) report.
   Minimum requirement is SIX Meeting Logs (Period: Week 4 to Week 14). Students can have fortnightly meetings with the supervisor.

3. Log sheets provide the basis for evaluating the General Effort (Project Management, Attitude, and Technical Competency) of the student, by the supervisor and for checking the attendance requirement of the student, by the FYP Committee.

   This also provides the student with feedback from the supervisor / co-supervisor / company supervisor on the tasks done and provides the plan for the upcoming tasks. This can provide the motivation for the student to give consistent and efficient effort throughout the period of FYP.

4. Student who fails to meet the minimum requirement (six nos.) of log sheets will not be allowed to submit FYP report.

8.1.8 FYP Meeting Log 2

MULTIMEDIA UNIVERSITY | FACULTY OF COMPUTING & INFORMATICS

TPT3101 Final Year Project (FYP2) Meeting Log
Trimester 2, 2022/23 (Trimester ID:2220)

Meeting Date: 19/4/2023	**Meeting No.:** 5
Meeting Mode: Online (Google Meet)	
Project ID: 2189	**Project Type:** Research-based
Project Title : Sentiment Analysis of Small Cap and Big Cap companies in United States Indexes	
Student ID : 1181103271	**Student Name:** Muhammad Khairulrazi Bin Mohd Riza
Student Programme and Specialisation: Bachelor of Computer Science Hons. (Data Science)	
Supervisor Name: Nathar Shah bin Packier Mohammad	**Co-Supervisor Name:** (if applicable)
Collaborating Company: (if applicable)	**Company Supervisor Name:** (if applicable)

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Broaden research objective
- experimenting certain aspect when comparing the techniques
- Started some technical writeup for research paper.
- data collection
- data cleaning

2. WORK TO BE DONE
[Please write the details of the work to be done, before the next meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- finish the technical writeup
- explore other twitter data for comparison
- improvise data cleaning steps

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- time limitation
- personal issues
- data collection / data cleaning / data preprocessing

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

.....
Supervisor's Signature

.....
Student's Signature

.....
Co-Supervisor's Signature
(if applicable)

.....
Company Supervisor's Signature
(if applicable)

IMPORTANT NOTES TO STUDENTS:

1. Items 1 – 3 are to be completed by the students prior to the meeting. Item 4 is to be completed by the supervisor / co-supervisor / company supervisor.

2. Student must upload the soft copies of the meeting logs in Google Classroom and attach them along with final (FYP2) report.
   Minimum requirement is SIX Meeting Logs (Period: Week 4 to Week 14). Students can have fortnightly meetings with the supervisor.

3. Log sheets provide the basis for evaluating the General Effort (Project Management, Attitude, and Technical Competency) of the student, by the supervisor and for checking the attendance requirement of the student, by the FYP Committee.

   This also provides the student with feedback from the supervisor / co-supervisor / company supervisor on the tasks done and provides the plan for the upcoming tasks. This can provide the motivation for the student to give consistent and efficient effort throughout the period of FYP.

4. Student who fails to meet the minimum requirement (six nos.) of log sheets will not be allowed to submit FYP report.

8.1.9 FYP Meeting Log 3

TPT3101 Final Year Project (FYP2) Meeting Log
Trimester 2, 2022/23 (Trimester ID:2220)

Meeting Date: 2/6/2023	**Meeting No.:** 10
Meeting Mode: Online (Google Meet)	
Project ID: 2189	**Project Type:** Research-based
Project Title : Sentiment Analysis of Small Cap and Big Cap companies in United States Indexes	
Student ID : 1181103271	**Student Name:** Muhammad Khairulrazi Bin Mohd Riza
Student Programme and Specialisation: Bachelor of Computer Science Hons. (Data Science)	
Supervisor Name: Nathar Shah bin Packier Mohammad	**Co-Supervisor Name:** (if applicable)
Collaborating Company: (if applicable)	**Company Supervisor Name:** (if applicable)

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- finish source code for both techniques for sentiment analysis
- find some external data for comparison

2. WORK TO BE DONE
[Please write the details of the work to be done, before the next meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Documentation of the project (technical writeup) is yet to be done.
- Increase data size
- Optimization (hyperparameter tuning) SVM algorithm
- Integration of ChatGPT into the project
    - How can ChatGPT be integrated in this project?
    - Comparative analysis using ChatGPT?
    - What API do I need to use to integrate ChatGPT?
    - Can ChatGPT help analyze words that are not in the lexicon?
- Comparative analysis against base paper
- Find ChatGPT/Bard API

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- time limitation
- personal issues

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

.....
Supervisor's Signature

.....
Student's Signature

.....
Co-Supervisor's Signature
(if applicable)

.....
Company Supervisor's Signature
(if applicable)

IMPORTANT NOTES TO STUDENTS:

1. Items 1 – 3 are to be completed by the students prior to the meeting. Item 4 is to be completed by the supervisor / co-supervisor / company supervisor.

2. Student must upload the soft copies of the meeting logs in Google Classroom and attach them along with final (FYP2) report.
   Minimum requirement is SIX Meeting Logs (Period: Week 4 to Week 14). Students can have fortnightly meetings with the supervisor.

3. Log sheets provide the basis for evaluating the General Effort (Project Management, Attitude, and Technical Competency) of the student, by the supervisor and for checking the attendance requirement of the student, by the FYP Committee.

   This also provides the student with feedback from the supervisor / co-supervisor / company supervisor on the tasks done and provides the plan for the upcoming tasks. This can provide the motivation for the student to give consistent and efficient effort throughout the period of FYP.

4. Student who fails to meet the minimum requirement (six nos.) of log sheets will not be allowed to submit FYP report.

8.1.10 FYP Meeting Log 4

TPT3101 Final Year Project (FYP2) Meeting Log
Trimester 2, 2022/23 (Trimester ID:2220)

Meeting Date: 14/6/2023	**Meeting No.:** 11
Meeting Mode: Physical Meeting	
Project ID: 2189	**Project Type:** Research-based
Project Title : Sentiment Analysis of Small Cap and Big Cap companies in United States Indexes	
Student ID : 1181103271	**Student Name:** Muhammad Khairulrazi Bin Mohd Riza
Student Programme and Specialisation: Bachelor of Computer Science Hons. (Data Science)	
Supervisor Name: Nathar Shah bin Packier Mohammad	**Co-Supervisor Name:** (if applicable)
Collaborating Company: (if applicable)	**Company Supervisor Name:** (if applicable)

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Recollect data for both big cap and small cap companies
- This is due to the last minute changes in the datasets

2. WORK TO BE DONE
[Please write the details of the work to be done, before the next meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Data labelling (Manual Label)
- Redo source code
- research paper writing

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- time limitation
- personal issues

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

.....
Supervisor's Signature

.....
Student's Signature

.....
Co-Supervisor's Signature
(if applicable)

.....
Company Supervisor's Signature
(if applicable)

IMPORTANT NOTES TO STUDENTS:

8.1.11 FYP Meeting Log 5

FACULTY OF
COMPUTING
& INFORMATICS

TPT3101 Final Year Project (FYP2) Meeting Log
Trimester 2, 2022/23 (Trimester ID:2220)

Meeting Date: 14/6/2023	**Meeting No.:** 12
Meeting Mode: Physical Meeting	
Project ID: 2189	**Project Type:** Research-based
Project Title : Sentiment Analysis of Small Cap and Big Cap companies in United States Indexes	
Student ID : 1181103271	**Student Name:** Muhammad Khairulrazi Bin Mohd Riza
Student Programme and Specialisation: Bachelor of Computer Science Hons. (Data Science)	
Supervisor Name: Nathar Shah bin Packier Mohammad	**Co-Supervisor Name:** (if applicable)
Collaborating Company: (if applicable)	**Company Supervisor Name:** (if applicable)

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Research paper writing and final report writing in progress
- finished labelling the new dataset
- Finish implementation of sentiment analysis on the newly obtained dataset

2. WORK TO BE DONE

[Please write the details of the work to be done, before the next meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Hyperparameter tuning for SVM model
- Continue on research paper and final report writing

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- time limitation
- personal issues

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

.....
Supervisor's Signature

.....
Student's Signature

.....
Co-Supervisor's Signature
(if applicable)

.....
Company Supervisor's Signature
(if applicable)

IMPORTANT NOTES TO STUDENTS:

1. Items 1 – 3 are to be completed by the students prior to the meeting. Item 4 is to be completed by the supervisor / co-supervisor / company supervisor.

2. Student must upload the soft copies of the meeting logs in Google Classroom and attach them along with final (FYP2) report.
   Minimum requirement is SIX Meeting Logs (Period: Week 4 to Week 14). Students can have fortnightly meetings with the supervisor.

3. Log sheets provide the basis for evaluating the General Effort (Project Management, Attitude, and Technical Competency) of the student, by the supervisor and for checking the attendance requirement of the student, by the FYP Committee.

   This also provides the student with feedback from the supervisor / co-supervisor / company supervisor on the tasks done and provides the plan for the upcoming tasks. This can provide the motivation for the student to give consistent and efficient effort throughout the period of FYP.

4. Student who fails to meet the minimum requirement (six nos.) of log sheets will not be allowed to submit FYP report.

8.1.12 FYP Meeting Log 6

TPT3101 Final Year Project (FYP2) Meeting Log
Trimester 2, 2022/23 (Trimester ID:2220)

Meeting Date: 14/6/2023	**Meeting No.:** 12
Meeting Mode: Physical Meeting	
Project ID: 2189	**Project Type:** Research-based
Project Title : Sentiment Analysis of Small Cap and Big Cap companies in United States Indexes	
Student ID : 1181103271	**Student Name:** Muhammad Khairulrazi Bin Mohd Riza
Student Programme and Specialisation: Bachelor of Computer Science Hons. (Data Science)	
Supervisor Name: Nathar Shah bin Packier Mohammad	**Co-Supervisor Name:** (if applicable)
Collaborating Company: (if applicable)	**Company Supervisor Name:** (if applicable)

1. WORK DONE
[Please write the details of the work done, after the last meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Research paper is fully done
- Documentation of implementations

2. WORK TO BE DONE

[Please write the details of the work to be done, before the next meeting]

Tasks: Implementation / Testing (Application-based projects) or Evaluation of Findings and Research Contribution (Research-based projects) / Commercialisation Proposal (Application-based projects) or Research Paper (Research-based Projects) / Draft Final Report Completion

(Please strike out the tasks, which are not applicable)

Details (in point form):

- Finishing up final report writing

3. PROBLEMS ENCOUNTERED AND SOLUTIONS

[Please write the details of the problems encountered, after the last meeting and provide the solutions / plan for the solutions]

- time limitation
- personal issues

4. COMMENTS (Supervisor / Co-Supervisor / Company Supervisor)

.....
Supervisor's Signature

.....
Student's Signature

.....
Co-Supervisor's Signature
(if applicable)

.....
Company Supervisor's Signature
(if applicable)

IMPORTANT NOTES TO STUDENTS:

1. Items 1 – 3 are to be completed by the students prior to the meeting. Item 4 is to be completed by the supervisor / co-supervisor / company supervisor.

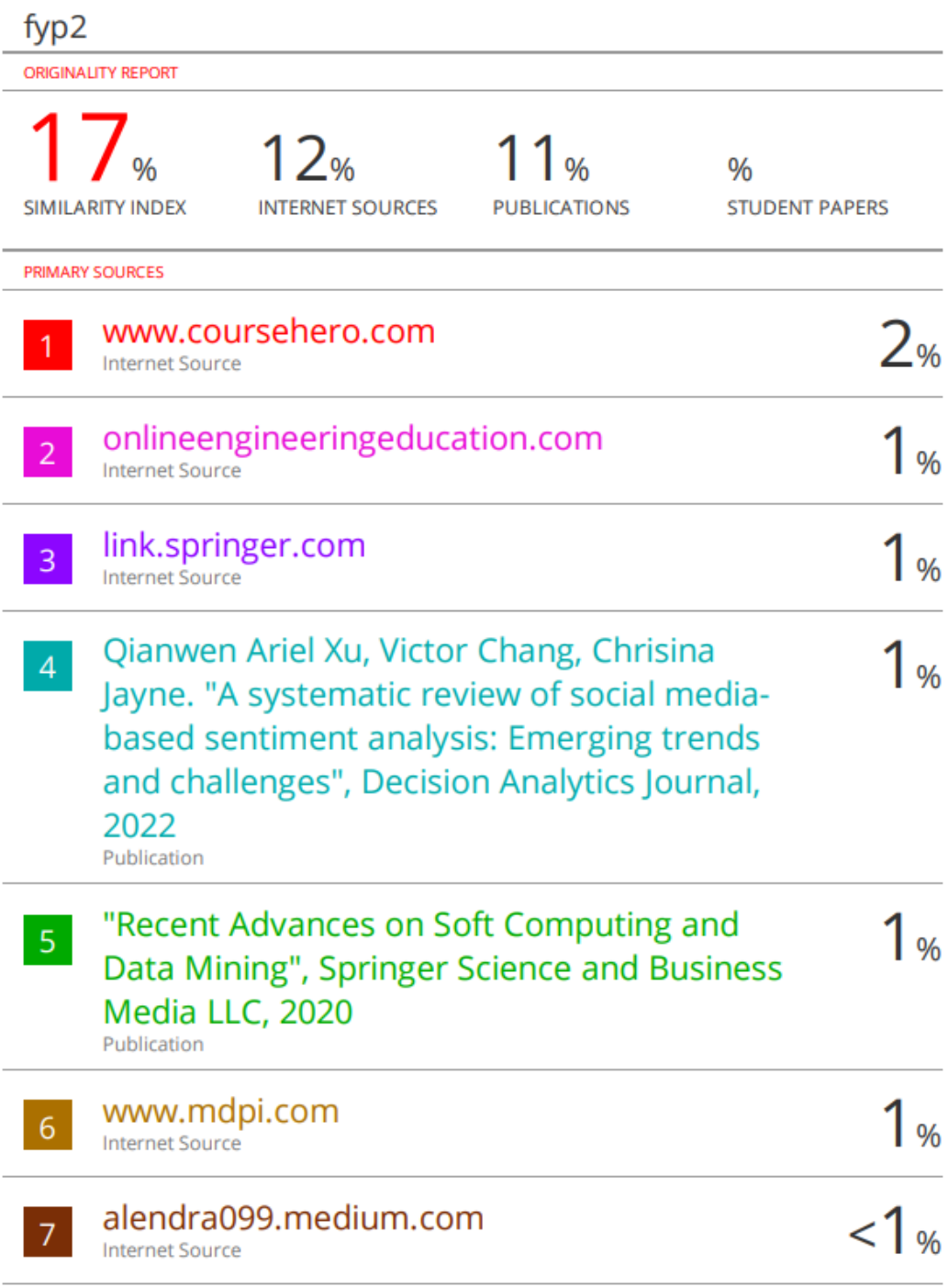
2. Student must upload the soft copies of the meeting logs in Google Classroom and attach them along with final (FYP2) report.
   Minimum requirement is SIX Meeting Logs (Period: Week 4 to Week 14). Students can have fortnightly meetings with the supervisor.

3. Log sheets provide the basis for evaluating the General Effort (Project Management, Attitude, and Technical Competency) of the student, by the supervisor and for checking the attendance requirement of the student, by the FYP Committee.

   This also provides the student with feedback from the supervisor / co-supervisor / company supervisor on the tasks done and provides the plan for the upcoming tasks. This can provide the motivation for the student to give consistent and efficient effort throughout the period of FYP.

4. Student who fails to meet the minimum requirement (six nos.) of log sheets will not be allowed to submit FYP report.

8.2 Appendix B



fyp2

ORIGINALITY REPORT

17%	12%	11%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	2%
2	onlineengineeringeducation.com Internet Source	1%
3	link.springer.com Internet Source	1%
4	Qianwen Ariel Xu, Victor Chang, Chrisina Jayne. "A systematic review of social media-based sentiment analysis: Emerging trends and challenges", Decision Analytics Journal, 2022 Publication	1%
5	"Recent Advances on Soft Computing and Data Mining", Springer Science and Business Media LLC, 2020 Publication	1%
6	www.mdpi.com Internet Source	1%
7	alendra099.medium.com Internet Source	<1%

8	repository.ihu.edu.gr Internet Source	<1 %
9	utpedia.utp.edu.my Internet Source	<1 %
10	MSVPJ SATHVIK. "Enhancing Machine Learning Algorithms using GPT Embeddings for Binary Classification", Institute of Electrical and Electronics Engineers (IEEE), 2023 Publication	<1 %
11	Ashwin Sanjay Neogi, Kirti Anilkumar Garg, Ram Krishn Mishra, Yogesh K Dwivedi. "Sentiment analysis and classification of Indian farmers' protest using twitter data", International Journal of Information Management Data Insights, 2021 Publication	<1 %
12	Minhajul Abedin Shafin, Md. Mehedi Hasan, Md. Rejaul Alam, Mosaddek Ali Mithu, Arafat Ulllah Nur, Md. Omar Faruk. "Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language", 2020 23rd International Conference on Computer and Information Technology (ICCIT), 2020 Publication	<1 %
13	nciet.srpec.org.in Internet Source	<1 %

14	Aniket K. Shahade, K.H. Walse, V.M. Thakare, Mohammad Atique. "Multi-lingual opinion mining for social media discourses: an approach using deep learning based hybrid fine-tuned smith algorithm with adam optimizer", International Journal of Information Management Data Insights, 2023 Publication	<1 %
15	Zulfadzli Drus, Haliyana Khalid. "Sentiment Analysis in Social Media and Its Application: Systematic Literature Review", Procedia Computer Science, 2019 Publication	<1 %
16	ir.uitm.edu.my Internet Source	<1 %
17	digitalcommons.odu.edu Internet Source	<1 %
18	Akshay Kulkarni, Adarsha Shivananda, Anoosh Kulkarni. "Natural Language Processing Projects", Springer Science and Business Media LLC, 2022 Publication	<1 %
19	Gianluca Anese, Marco Corazza, Michele Costola, Loriana Pelizzon. "Impact of public news sentiment on stock market index return and volatility", Computational Management Science, 2023 Publication	<1 %

20	docplayer.net Internet Source	<1 %
21	ebin.pub Internet Source	<1 %
22	max-success.eu Internet Source	<1 %
23	Monali Bordoloi, Saroj Kumar Biswas. "Sentiment analysis: A survey on design framework, applications and future scopes", Artificial Intelligence Review, 2023 Publication	<1 %
24	ijsrset.com Internet Source	<1 %
25	dokumen.pub Internet Source	<1 %
26	www.researchgate.net Internet Source	<1 %
27	Kanika Jindal, Rajni Aron. "A systematic study of sentiment analysis for social media data", Materials Today: Proceedings, 2021 Publication	<1 %
28	spectrum.library.concordia.ca Internet Source	<1 %
29	cris.brighton.ac.uk Internet Source	<1 %

30	etd.lsu.edu Internet Source	<1 %
31	revistaie.ase.ro Internet Source	<1 %
32	"Proceedings of International Conference on Emerging Technologies and Intelligent Systems", Springer Science and Business Media LLC, 2022 Publication	<1 %
33	dspace.khazar.org Internet Source	<1 %
34	"Proceedings of Integrated Intelligence Enable Networks and Computing", Springer Science and Business Media LLC, 2021 Publication	<1 %
35	hdl.handle.net Internet Source	<1 %
36	"Mobile Radio Communications and 5G Networks", Springer Science and Business Media LLC, 2021 Publication	<1 %
37	Mondher Bouazizi, Tomoaki Ohtsuki. "Multi-class sentiment analysis on twitter: Classification performance and challenges", Big Data Mining and Analytics, 2019 Publication	<1 %

38	my.ai.se Internet Source	<1 %
39	ui.adsabs.harvard.edu Internet Source	<1 %
40	Akshay Kulkarni, Adarsha Shivananda. "Natural Language Processing Recipes", Springer Science and Business Media LLC, 2021 Publication	<1 %
41	mcgill-fammedstudies-recherchemedfam.pbworks.com Internet Source	<1 %
42	peerj.com Internet Source	<1 %
43	9pdf.net Internet Source	<1 %
44	Marjan Van de Kauter, Bart Desmet, Véronique Hoste. "The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment", Language Resources and Evaluation, 2015 Publication	<1 %
45	github.com Internet Source	<1 %
46	pure.hud.ac.uk Internet Source	<1 %

47	revista.profesionaldelainformacion.com Internet Source	<1 %
48	search.bvsalud.org Internet Source	<1 %
49	Ankita Sharma, Udayan Ghose. "Toward Machine Learning Based Binary Sentiment Classification of Movie Reviews for Resource Restraint Language (RRL) – Hindi", IEEE Access, 2023 Publication	<1 %
50	Debashis Naskar. "Temporal Emotion Dynamics in Social Networks", Universitat Politecnica de Valencia, 2022 Publication	<1 %
51	Lívia Kelebercová, Michal Munk, František Forgáč. "Could You Understand Me? The Relationship among Method Complexity, Preprocessing Complexity, Interpretability, and Accuracy", Mathematics, 2023 Publication	<1 %
52	core.ac.uk Internet Source	<1 %
53	eprints.utar.edu.my Internet Source	<1 %
54	pnrjournal.com Internet Source	<1 %

55	repository.tudelft.nl Internet Source	<1 %
56	usir.salford.ac.uk Internet Source	<1 %
57	www.qc.dfo-mpo.gc.ca Internet Source	<1 %
58	www.readkong.com Internet Source	<1 %
59	"Advances in Data Science, Cyber Security and IT Applications", Springer Science and Business Media LLC, 2019 Publication	<1 %
60	"Intelligent Data Communication Technologies and Internet of Things", Springer Science and Business Media LLC, 2020 Publication	<1 %
61	Lecture Notes in Computer Science, 2014. Publication	<1 %
62	Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. "Sentiment analysis in Facebook and its application to e-learning", Computers in Human Behavior, 2013. Publication	<1 %
63	Rahul Kumar Singh, Manoj Kumar Sachan, R. B. Patel. "360 degree view of cross-domain	<1 %

	opinion classification: a survey", Artificial Intelligence Review, 2020 Publication	
64	Vibhuti Gupta, Rattikorn Hewett. "Real-Time Tweet Analytics Using Hybrid Hashtags on Twitter Big Data Streams", Information, 2020 Publication	<1 %
65	addictionresearchchair.com Internet Source	<1 %
66	curve.carleton.ca Internet Source	<1 %
67	e-journal.unair.ac.id Internet Source	<1 %
68	ir.jkuat.ac.ke Internet Source	<1 %
69	livecodestream.dev Internet Source	<1 %
70	salford-repository.worktribe.com Internet Source	<1 %
71	techscience.com Internet Source	<1 %
72	umpir.ump.edu.my Internet Source	<1 %
73	www.ijpe-online.com Internet Source	<1 %

74	"Artificial Intelligence and Speech Technology", Springer Science and Business Media LLC, 2022 Publication	<1 %
75	"International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2020 Publication	<1 %
76	Krzysztof Grabczewski. "Handwritten Digit, Recognition Road to Contest victory", 2007 IEEE Symposium on Computational Intelligence and Data Mining, 04/2007 Publication	<1 %
77	Thanveer Shaik, Xiaohui Tao, Christopher Dann, Haoran Xie, Yan Li, Linda Galligan. "Sentiment analysis and opinion mining on educational data: A survey", Natural Language Processing Journal, 2022 Publication	<1 %
78	Communications in Computer and Information Science, 2015. Publication	<1 %
79	Deepali Arora, Kin Fun Li, Stephen W. Neville. "Consumers' Sentiment Analysis of Popular Phone Brands and Operating System Preference Using Twitter Data: A Feasibility Study", 2015 IEEE 29th International	<1 %

Conference on Advanced Information Networking and Applications, 2015
Publication

80	Dimple Tiwari, Bharti Nagpal, Bhoopesh Singh Bhati, Ashutosh Mishra, Manoj Kumar. "A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques", Artificial Intelligence Review, 2023 Publication	<1 %
81	Ramesh Chundi, Vishwanath R. Hulipalled, Jay Bharthish Simha. "Lexicon-based sentiment analysis for Kannada-English code-switch text", IAES International Journal of Artificial Intelligence (IJ-AI), 2023 Publication	<1 %

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		

8.3 Appendix C

Sentiment analysis of Big Cap and Small Cap Companies in the US indexes using TextBlob, Support Vector Machine, and ChatGPT

Muhammad Khairulrazi [1] and Nathar Shah [2]

[1] Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, 63000, Cyberjaya, Selangor

Abstract. Analysing stock market trends through sentiment analysis has become increasingly important for understanding public opinion and sentiment towards companies. This research paper evaluates and compares three approaches for sentiment analysis which is TextBlob, Support Vector Machine (SVM) and ChatGPT using Twitter data against manually labelled datasets. This study focuses on sentiment analysis of small cap and big cap companies in the US indexes. The results show that SVM outperforms TextBlob and ChatGPT, achieving accuracy of 79.3% for small cap and 85% for big cap companies after hyperparameter tuning using GridSearchCV. Among the techniques used in this study, TextBlob performs the lowest accuracy (44% for small cap, 47% for big cap), while ChatGPT demonstrated promising results with 74% accuracy for small cap and 77% accuracy for big cap companies.

Keywords: Stock market, Sentiment analysis, TextBlob, Support Vector Machine, ChatGPT,

1. Introduction

Analysing stock market trends has been regarded as one of the most difficult yet very crucial tasks [1]. One approach to analyse the stock market is through sentiment analysis, which has gained prominence as a vital tool for understanding public opinion and sentiment towards various entities, including companies listed in financial markets [2]. This understanding of sentiment towards companies is valuable for investors, market analysts, and financial decision-makers. With the advent of social media, news articles, and online forums, sentiment analysis techniques have become crucial in extracting meaningful information from vast amounts of unstructured text data [3].

The United States stock market comprises a diverse range of companies, varying in

market capitalization. Big cap companies, characterised by their large market value, and small cap companies, representing relatively smaller firms, often exhibit distinct dynamics and investor sentiments. Assessing sentiment towards both types of companies can help identify potential investment opportunities and gauge market sentiment towards different market segments.

This research paper aims to evaluate and compare three distinct approaches for sentiment analysis: lexicon-based methods, machine learning algorithms, and ChatGPT, a state-of-the-art language model developed by OpenAI. By assessing the performance of these approaches, we seek to determine their effectiveness in analysing sentiment towards big cap and small cap companies in US indexes.

Lexicon-based methods have been widely used in sentiment analysis and rely on predefined lexicons or dictionaries containing sentiment scores for words [4]. These methods assign sentiment scores to individual words or phrases and aggregate them to determine the overall sentiment of a document or text snippet. Previous studies have shown the effectiveness of lexicon-based methods in sentiment analysis tasks. For example, a study by [5] proposed a method to determine the sentiment density of Turkish tweets using SentiWordNet and have an average success rate of 80%.

Machine learning algorithms, on the other hand, employ statistical models to learn patterns and relationships in labelled training data and make predictions on new, unseen data [6][7]. One of the most popular machine learning techniques is Support Vector Machine (SVM). The basic idea behind SVM is to find the optimal hyperplane that separates the data points into different classes[8]. Support Vector Machines (SVM), have been successfully applied to sentiment analysis tasks and demonstrated promising results in various domains, including finance and stock market sentiment analysis such as in [9][10][11][12]

Additionally, in our evaluation, we incorporate ChatGPT, a state-of-the-art language model developed by OpenAI. It is trained using the GPT-3.5 model via RLHF (reinforcement learning from human feedback), which aligns the model to human preference[13]. While there is limited research related to ChatGPT's application in sentiment analysis, one study has demonstrated its effectiveness in sentiment analysis ability, which results in matching and even surpassing fine-tuned models like BERT in various sentiment analysis tasks [14].

By conducting a comparative analysis of these approaches, we aim to provide a comprehensive understanding of their strengths, limitations, and performance in sentiment analysis of big cap and small cap companies in US indexes. The insights derived from this research can assist investors, financial analysts, and decision-makers in making informed decisions based on sentiment analysis of textual data.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work in sentiment analysis, focusing on lexicon-based methods, Support Vector Machine (SVM), and recent advancements in language models. Section 3 presents the methodology employed in our evaluation, including the datasets, feature extraction techniques, and evaluation metrics. Section 4 presents the results and discussion, analysing the performance of each approach in sentiment analysis tasks. Finally, Section 5 concludes the paper, summarising the key findings and outlining avenues for future research.

2. Related Works

This section of the paper examines and highlights the existing studies that are similar to ours. There are many research papers that perform sentiment analysis in various machine learning approaches and techniques. For example, in this study, R Uma et al. [15] focuses on sentiment analysis in the realm of customer reviews using Support Vector Machine (SVM) and Convolutional Neural Network (CNN). After dataset cleaning, text preprocessing, and hyperparameter tuning, the SVM and CNN models achieved accuracy of 96% and 94% respectively on web-scraped customer reviews data.

Another study by Leelawat et al. [16] investigated the impact of COVID-19 pandemic on Thailand's tourism industry using Twitter data. The tweets related to specific cities in Thailand including Bangkok, Chiang Mai, and Phuket were analysed using three machine learning techniques, namely, SVM, Random Forest, and Decision Tree. From this analysis, the SVM algorithm achieved the highest accuracy in sentiment analysis at 77.4%, while Random Forest algorithm achieved 95.4% accuracy for intention analysis.

Additionally, S. Bengesi et al. [17] explores sentiment analysis in the context of the monkeypox outbreak. 500,000 multilingual tweets were collected and analysed using VADER and TextBlob. In addition, they developed and evaluated 56 classification models, employing techniques such as stemming, lemmatization, CountVectorizer, and TF-IDF. The model with TextBlob annotation + lemmatization + CountVectorizer + SVM achieved the highest accuracy of approximately 0.9348.

A study conducted by R. N. Satrya et al. [18] which utilises SVM classification for sentiment analysis on Twitter data related to cryptocurrencies. The study aims to identify positive and negative trends by classifying Twitter comments. The dataset includes tweets about cryptocurrencies from June 2022. The application of the SVM model achieves an accuracy of 93.13% using k-fold cross-validation. However, the highest accuracy of 94.64% is obtained with the undersampling method and an 80:20 train-test split. Other methods such as handling imbalance data and testing of different train-test-split ratios are explored in this study to further explore the potential of SVM model towards Twitter data related to cryptocurrency.
In another study conducted by S. Datta et al. [19] that focus on sentiment analysis of Twitter data related to COVID-10 Omicron variant. In this study, 3 sentiment analyzers are applied, namely, VADER, NLTK, and TextBlob to classify tweets as positive, neutral or negative. The dataset consists of over 78,000 samples and the result of sentiment analysis shows that the TextBlob sentiment analyzer performs the best amongst the other 2 techniques used by the authors in this study.

3. Methodology

The methodology employed in this study is presented in **Figure 1**, outlining the flowchart of the research process. Initially, the necessary datasets were obtained from the TwitterAPI through the stock news feed on Twitter. To collect relevant Twitter data pertaining to small cap and big cap companies, the Twitter API was queried using the ticker symbols of the respective companies, including ETF symbols such as 'IWM' for small cap and 'NASDAQ100' for big cap companies. This data collection phase resulted in the acquisition of 5,000 tweets for each category (small cap and big cap).

Following data collection, manual labelling of the datasets was conducted to assign sentiments to the tweets. While manual labelling ensures accuracy, it is a time-consuming process. Subsequently, data cleaning procedures were performed, including the removal of @mentions, hashtags, and the 'RT' symbol, among other data preprocessing steps. Further details on the data preprocessing process can be found in Section 3.1 of this paper.

After cleaning the data, sentiment analysis using TextBlob was applied to the cleaned tweets. The cleaned tweets were then divided into training and testing sets, and TF-IDF feature extraction was applied to prepare the data for SVM modelling. The SVM model was trained and tested using default parameters, and subsequent hyperparameter tuning was performed using GridSearchCV.

In addition to TextBlob and SVM, ChatGPT was employed for sentiment analysis. The ChatGPT model was utilised to generate sentiment predictions based on the text data. Finally, a comparison of the results obtained from the three techniques (TextBlob, SVM, and ChatGPT) was conducted.

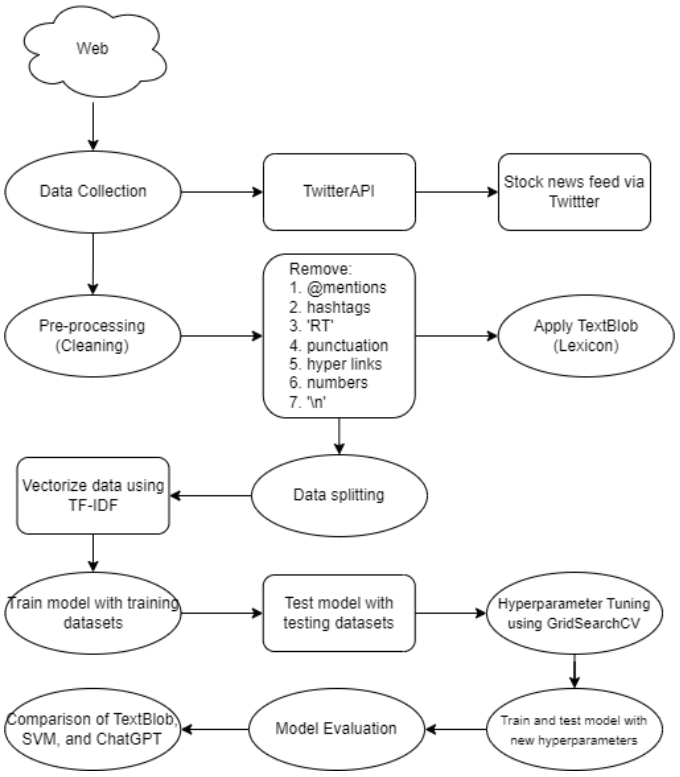


Figure 1: Implementation flowchart

3.1 Data Preprocessing

To prepare the raw data collected from the Twitter API for further processing, a series of data preprocessing steps were implemented. The objective was to ensure that the data is clean and in the appropriate format for optimal performance of the Support Vector Machine (SVM) model. It is important to note that the actual content of the tweets was left

unaltered, without any changes or manipulation. The following preprocessing steps were performed:

1. ***Noise Removal:*** Various types of noise, including @ signs, hashtags, 'RT' symbols, hyperlinks, numbers, white spaces, and punctuations, were removed from the tweets.
2. ***Stop Words Removal:*** Stop words, such as articles, prepositions, and conjunctions (e.g., am, are, an, the, is), were eliminated. These words do not contribute significantly to sentiment classification and are commonly found in any text, including tweets.
3. ***Lowercasing:*** All sentences were converted to lowercase letters to ensure consistency in the text.
4. ***Tokenization:*** The process of tokenization was applied, which involves breaking down the text into smaller units, typically words. This step is crucial for subsequent natural language processing (NLP) tasks as it enables computer analysis, processing, and comprehension of the text.
5. ***Lemmatization:*** Lemmatization, a linguistic technique, was utilized to transform words into their base or root form. This process helps in standardizing the words and improves the analysis and understanding of the text. For example, it converts "ate" to its base form, "eat," and reduces "looked" to "look."

3.2 Data Labelling

In this study, the collected Twitter data underwent manual labelling to assign sentiment labels to the tweets such as positive, negative, or neutral. Manual labelling involves the process of annotating each tweet with its respective sentiment category. The sentiment labels were assigned based on the overall sentiment conveyed in the tweet's content. This process involved considering the contextual information, tone, and expressed emotions within the tweet. While manual labelling ensures accuracy in sentiment classification, it is a time-consuming task that requires human judgement and expertise.

The manually labelled dataset serves as the ground truth for evaluating the performance of sentiment analysis techniques (TextBlob, SVM, ChatGPT) for small cap and big cap companies in US indexes.

3.3 TF-IDF Feature Extraction

To prepare the text data for SVM modelling, TF-IDF (Term Frequency-Inverse Document Frequency) feature extraction was applied. TF-IDF measures the importance of words based on their frequency and rarity across documents. The process involves calculating term frequency (TF), inverse document frequency (IDF), and combining them to obtain TF-IDF scores. These scores represent the significance of words in individual tweets and across the dataset. Transforming the data into numerical feature vectors allows the SVM model to learn patterns for sentiment classification in small cap and big cap companies.

3.4 TextBlob

In this study TextBlob(Lexicon) is utilised for sentiment analysis. It provides a simple interface to analyse the polarity (positive, negative, or neutral) and subjectivity (objective or subjective) of cleaned tweets. TextBlob employs a pre-trained sentiment analysis model and lexicon-based approach, assigning sentiment scores to words and aggregating them for overall sentiment assessment. This baseline technique allows comparison with manual labelling for sentiment classification in small cap and big cap companies.

3.5. Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm was utilised for sentiment analysis in this study. SVM is a supervised learning model commonly used in NLP tasks. It was trained on the cleaned and labelled tweet data, using TF-IDF feature vectors as input. Default parameters were used initially, and hyperparameter tuning was performed with GridSearchCV to optimise performance.

3.6. ChatGPT

ChatGPT, a state-of-the-art language model developed by OpenAI, was employed for sentiment analysis. Despite limited research specifically focusing on sentiment analysis, recent studies have highlighted its effectiveness in language tasks. ChatGPT was used to generate sentiment predictions based on text data, exploring its potential and comparing its performance with other techniques such as TextBlob and SVM. Its powerful language understanding and generation abilities make it a valuable addition to the evaluation

4. Experiment result

This section presents the experimental results of the research conducted using Jupyter Notebook with the Microsoft Visual Studio Code IDE. A total of 10,000 tweets were collected, with 5,000 tweets for small cap companies and an equal number for big cap companies.

From the manual labelling process, it was found that for the small cap dataset, there were 1,139 positive tweets, 561 negative tweets, and 3,299 neutral tweets. Similarly, for big cap companies, there were 993 positive tweets, 265 negative tweets, and 3,742 neutral tweets. These results are visualised in **Figure 2** illustrating the distribution of positive, neutral, and negative tweets for both datasets.

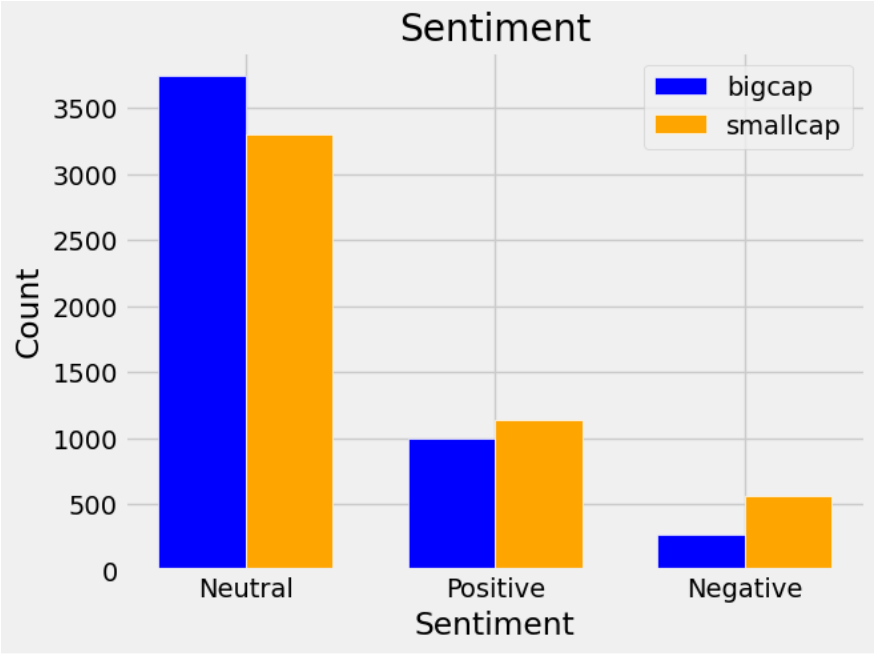


Figure 2: Distribution of Sentiment from Manual Label

4.1 TextBlob

TextBlob was performed for small cap and big cap companies. The results were compared with the manually labelled sentiments to calculate the accuracy. The accuracy achieved for small cap companies using TextBlob was 44%, while for big cap companies, it was 47%. The lower accuracy may be attributed to the complexity of stock market tweets, which often contain nuanced language and diverse sentiment expressions. **Figure 3** was created to compare the performance of the lexicon-based approach (TextBlob) against the manually labelled sentiments for both small cap and big cap companies.

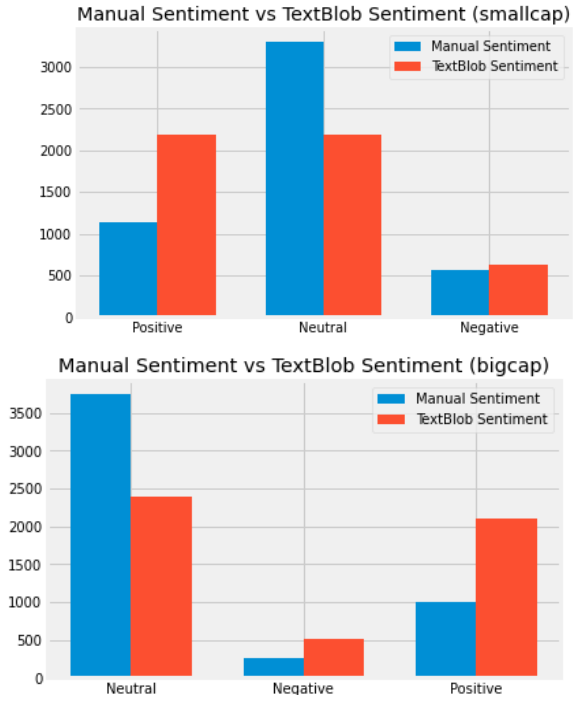


Figure 3. Manual Sentiment vs. TextBlob sentiment

Additionally, **Figure 4** represents the scatterplot of polarity vs. subjectivity that was generated to visualise the relationship between these two sentiment analysis metrics. Subjectivity represents the degree of subjectiveness in the text, while polarity indicates the sentiment orientation (positive, negative, or neutral). The scatterplot provides insights into how subjectivity and polarity are interconnected.

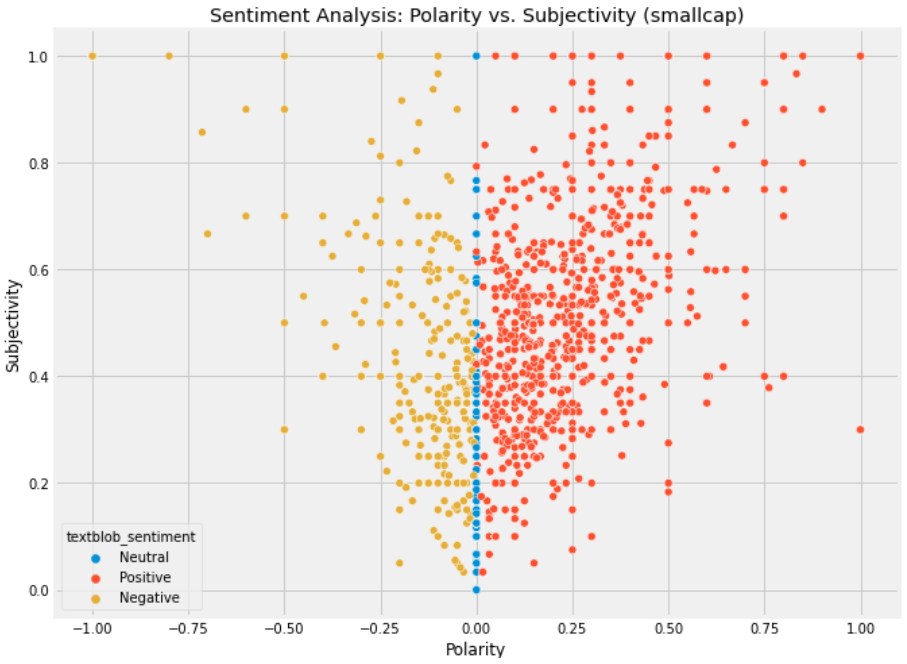


Figure 4. Polarity vs. Subjectivity (smallcap)

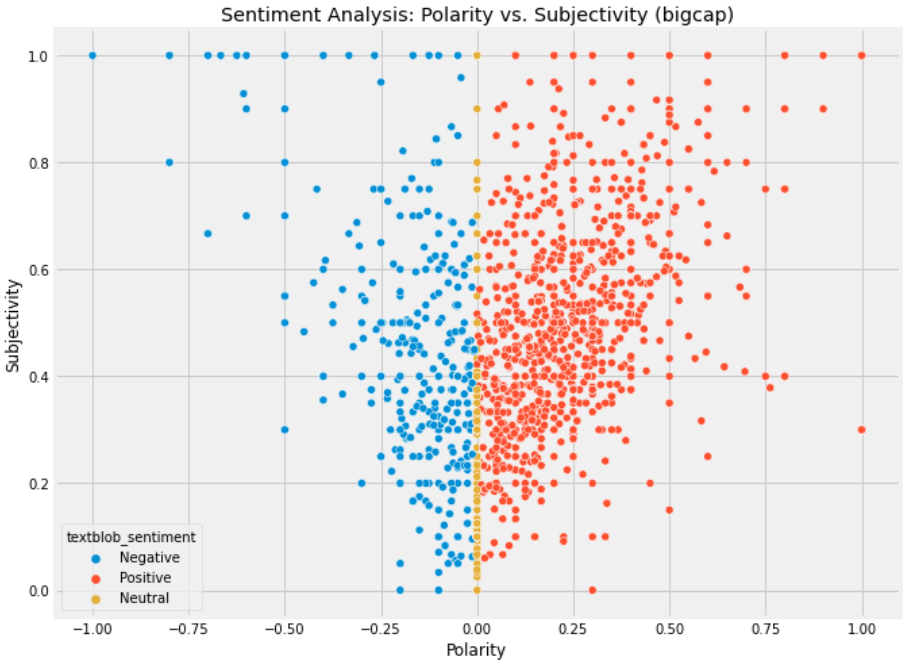


Figure 5. Polarity vs. Subjectivity (bigcap)

Word Clouds displaying the top words for each sentiment class (positive, neutral, negative) is depicted in **Figure 5** and **Figure 6**. These word clouds offer a visual representation of the most frequently occurring words associated with each sentiment category.

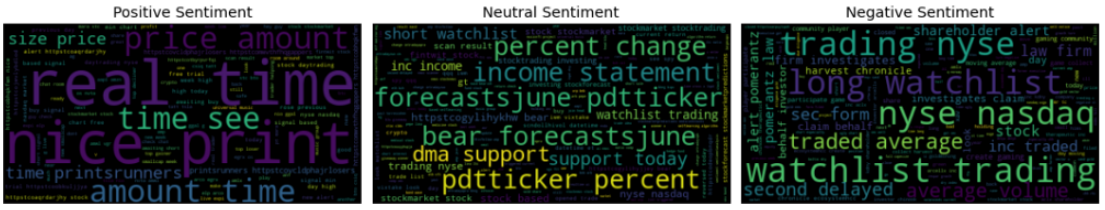


Figure 5. SmallCap Sentiment Word cloud

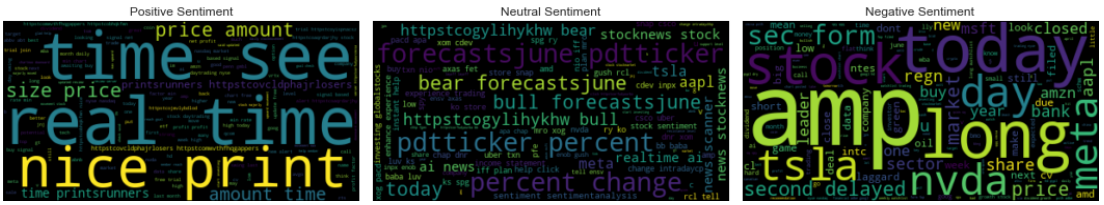


Figure 6. Bigcap sentiment word cloud

4.2 Support Vector Machine

The sentiment analysis results obtained from the Support Vector Machine (SVM) were also evaluated. Using default parameters, the SVM model achieved an accuracy of

78.3%, precision of 0.81, recall of 0.78, and F1-score of 0.76. The performance can be considered relatively good, indicating the effectiveness of the SVM model in sentiment classification.

After performing hyperparameter tuning using GridSearchCV, the SVM model achieved improved performance. For small cap companies, the tuned model achieved an accuracy of 79.3%, precision of 0.79, recall of 0.79, and F1-score of 0.78. Similarly, for big cap companies, the accuracy was 85%, with precision, recall, and F1-score of 0.85 and 0.83 respectively. The best hyperparameters identified were C=10, gamma='scale', and kernel='rbf', which further enhanced the SVM model's performance. Bar charts comparing the SVM results with the manually labelled sentiments were created for both small cap and big cap companies, providing visual insights into the model's sentiment classification performance.

Parameters	**C**	**kernel**	**gamma**	**Accuracy(%) small cap**	**Accuracy(%) big cap**
Default	1.0	rbf	scale	**78.3**	**83**
After Tuned	10	rbf	scale	**79.3**	**85**

Table 1. Hyperparameter settings for SVM

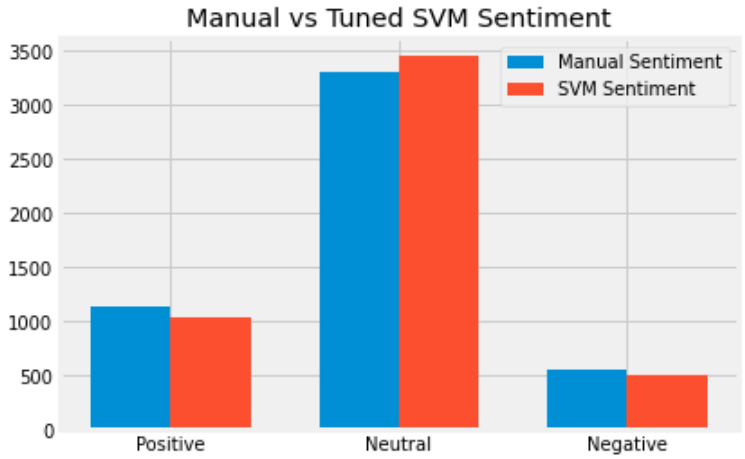


Figure 7. Manual vs. SVM sentiments(smallcap)

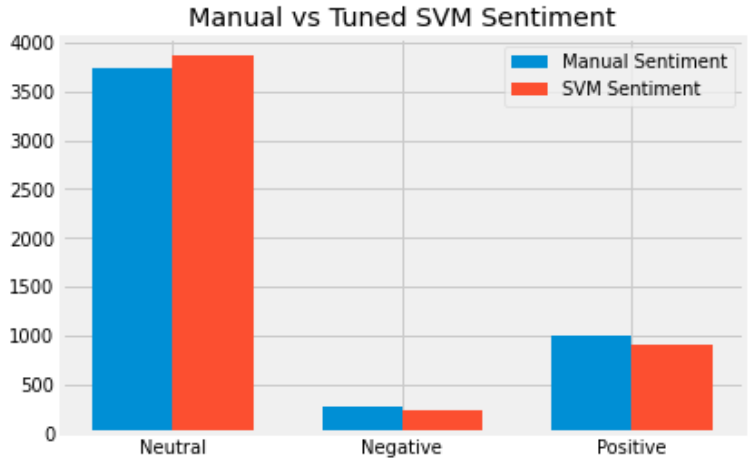


Figure 8. Manual vs. SVM sentiments (bigcap)

4.3. Performance Comparison

Table 2 displays the comparison of performance of the different sentiment analysis approaches. TextBlob, a lexicon-based approach, achieved the lowest accuracy amongst the other 2 techniques at 44% for small cap companies and 47% for big cap companies. SVM initially achieved an accuracy of 79% for small cap companies and 85% for big cap companies. However, after hyperparameter tuning, the tuned SVM model improved to 79.3% accuracy for small cap companies and 85% accuracy for big cap companies. Lastly, ChatGPT demonstrated accuracy of 74% for small cap and 77% for big cap companies.

Methods	Cap Size	Accuracy (%)
TextBlob	Big cap	46.52
	Small cap	43.57
SVM	Big cap	83
	Small cap	78
SVM (Tuned)	Big cap	85
	Small cap	79.3
ChatGPT	Big cap	77.44
	Small cap	72.43

4.4 Discussion

TextBlob, a lexicon-based approach, achieved moderate accuracy, with scores of 44% for small cap and 47% for big cap companies. This can be attributed to the challenges posed by the complex language and nuanced expressions often found in stock market tweets. However, TextBlob can serve as a baseline technique and provides a simple interface for sentiment analysis. The Support Vector Machine (SVM) model demonstrated good performance, even with default parameters. The initial results yielded an accuracy of 78.3% for small cap and big cap companies, with precision, recall, and F1-score values indicating effective sentiment classification. Through hyperparameter tuning with GridSearchCV, the SVM model achieved further improvements, reaching an accuracy of 79.3% for small cap and 85% for big cap companies. These enhancements highlight the significance of parameter optimization in enhancing the SVM model's performance.

Comparing the results of SVM and TextBlob with the manually labelled sentiments, it is evident that both approaches have strengths and limitations. SVM, being a machine learning algorithm, demonstrates the ability to capture more intricate patterns and achieve higher accuracy compared to the lexicon-based approach of TextBlob. However, SVM requires careful parameter selection and hyperparameter tuning to unleash its full potential.

Additionally, the sentiment analysis results obtained using ChatGPT, a state-of-the-art language model, showcased promising accuracy, with 74% for small cap and 77% for big cap companies. Despite the limited research specifically focusing on ChatGPT for sentiment analysis, its language understanding capabilities and contextual comprehension contribute to its effectiveness in sentiment classification tasks.

Overall, the findings from this study emphasise the importance of selecting the appropriate sentiment analysis approach based on the context and requirements. TextBlob serves as a simple baseline technique, while SVM demonstrates robust performance with hyperparameter optimization. ChatGPT, although still an area of ongoing research, shows promise in sentiment analysis tasks.

5. Conclusion and Future Work

In this research paper, we conducted a comparative analysis of sentiment analysis approaches for small cap and big cap companies in US indexes. Three distinct techniques were evaluated: TextBlob, Support Vector Machine (SVM), and ChatGPT.

The experimental results provided insights into the performance of each approach. TextBlob, a lexicon-based method, achieved lowest accuracy, highlighting its simplicity but limited effectiveness in capturing the complexity of stock market tweets. SVM demonstrated good performance, both with default parameters and after hyperparameter tuning using GridSearchCV. The SVM model's accuracy and performance were further improved, emphasising the significance of parameter optimization. ChatGPT, on the other hand, showed promising accuracy in sentiment analysis, leveraging its advanced language understanding capabilities.

The comparative analysis of these techniques shed light on their strengths and limitations. TextBlob provides a baseline approach, while SVM and ChatGPT offer more advanced and sophisticated sentiment analysis capabilities. The findings highlight the importance of selecting the appropriate technique based on the specific requirements and context of sentiment analysis tasks.

In conclusion, this research paper contributes to the field of sentiment analysis by evaluating and comparing different approaches for sentiment classification in the context of small cap and big cap companies. The findings provide guidance for researchers and practitioners in selecting suitable techniques for sentiment analysis tasks and emphasise the importance of parameter optimization for achieving optimal performance. Further research can focus on exploring other advanced language models and refining sentiment analysis techniques to enhance their effectiveness and applicability in the financial domain.

References

[1] R. Ren, D. D. Wu and T. Liu, "Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine," in IEEE Systems Journal, vol. 13, no. 1, pp. 760-770, March 2019, doi: 10.1109/JSYST.2018.2794462.

[2] Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear.” Procedia - Social and Behavioral Sciences, 26, 55–62. https://doi.org/10.1016/j.sbspro.2011.10.562

[3] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[4] Lili Shang, Hao Xi, Jiaojiao Hua, Huayun Tang, Jilei Zhou, A Lexicon Enhanced Collaborative Network for targeted financial sentiment analysis, Information Processing & Management, Volume 60, Issue 2, 2023, 103187, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2022.103187.

[5] H. Karamollaoğlu, İ. A. Doğru, M. Dörterler, A. Utku and O. Yıldız, "Sentiment Analysis on Turkish Social Media Shares through Lexicon Based Approach," 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, 2018, pp. 45-49, doi: 10.1109/UBMK.2018.8566481.

[6] E. Cambria, B. Schuller, Y. Xia and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," in IEEE Intelligent Systems, vol. 28, no. 2, pp. 15-21, March-April 2013, doi: 10.1109/MIS.2013.30.

[7] Natt Leelawat, Sirawit Jariyapongpaiboon, Arnon Promjun, Samit Boonyarak, Kumpol Saengtabtim, Ampan Laosunthara, Alfan Kurnia Yudha, Jing Tang, “Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning,” Heliyon, Volume 8, Issue 10, 2022, ISSN 2405-8440, doi: 10.1016/j.heliyon.2022.e10894.

[8] Y. B. P. Pamukti and M. Rahardi, "Sentiment Analysis of Bandung Tourist Destination Using Support Vector Machine and Naïve Bayes Algorithm," 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering

(ICITISEE), Yogyakarta, Indonesia, 2022, pp. 391-395, doi: 10.1109/ICITISEE57756.2022.10057802.

[9] Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. International Journal of Information Management Data Insights, 1(2). https://doi.org/10.1016/j.jjimei.2021.100019
[10] S. Naz, A. Sharan and N. Malik, "Sentiment Classification on Twitter Data Using Support Vector Machine," 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, Chile, 2018, pp. 676-679, doi: 10.1109/WI.2018.00-13.

[11] N. Hasanati, Q. Aini and A. Nuri, "Implementation of Support Vector Machine with Lexicon Based for Sentiment Analysis on Twitter," 2022 10th International Conference on Cyber and IT Service Management (CITSM), Yogyakarta, Indonesia, 2022, pp. 1-4, doi: 10.1109/CITSM56380.2022.9935887

[12] Katarya, R., Nath, G. A., Singhal, D., & Shukla, A. (2022). Analysing the twitter sentiments in COVID-19 using machine learning algorithms. Paper presented at the Proceedings - IEEE International Conference on Advances in Computing, Communication and Applied Informatics, ACCAI 2022, doi:10.1109/ACCAI53970.2022.9752511

[13] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 4299–4307.

[14] Wang, Z., Xie, Q., Ding, Z., Feng, Y., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv preprint arXiv:2304.04339.

[15] R. Uma, A. S. H, P. Jawahar and B. V. Rishitha, "Support Vector Machine and Convolutional Neural Network Approach to Customer Review Sentiment Analysis," 2022 1st International Conference on Computational Science and Technology (ICCST), CHENNAI, India, 2022, pp. 239-243, doi: 10.1109/ICCST55948.2022.10040381.

[16] Natt Leelawat, Sirawit Jariyapongpaiboon, Arnon Promjun, Samit Boonyarak, Kumpol Saengtabtim, Ampan Laosunthara, Alfan Kurnia Yudha, Jing Tang, Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning, Heliyon, Volume 8, Issue 10, 2022, e10894, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2022.e10894.

[17] S. Bengesi, T. Oladunni, R. Olusegun and H. Audu, "A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets," in IEEE Access, vol. 11, pp. 11811-11826, 2023, doi: 10.1109/ACCESS.2023.3242290.

[18] R. N. Satrya, O. N. Pratiwi, R. Y. Fa'rifah and J. Abawajy, "Cryptocurrency Sentiment Analysis on the Twitter Platform Using Support Vector Machine (SVM) Algorithm," 2022 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS), Bandung, Indonesia, 2022, pp. 01-05, doi: 10.1109/ICADEIS56544.2022.10037413.

[19] S. Datta, P. Mitra and P. Kundu, "Sentiment Analysis on Twitter Data of Omicron (B.l.l.529) using Natural Language Processing," 2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, 2022, pp. 1-6, doi: 10.1109/IATMSI56455.2022.10119296.