SENTIMENT ANALYSIS ON PERSONAL BRAND IN MALAYSIA USING HYBRID APPROACH

MOHD FIKRI BIN MOHD HANIM

UNIVERSITI TEKNOLOGI MALAYSIA

PSZ 19:16 (Pind. 1/13)

UNIVERSITI TEKNOLOGI MALAYSIA

DECLARATION OF THESIS / UNDERGRADUATE PROJECT REPORT AND COPYRIGHT

Author's full name : MOHD FIKRI BIN MOHD HANIM

Date of Birth : 01 JUNE 1997

Title : SENTIMENT ANALYSIS ON PERSONAL BRAND IN MALAYSIA USING HYBRID APPROACH

Academic Session : 2022/2023

I declare that this thesis is classified as:

☐	**CONFIDENTIAL**	(Contains confidential information under the Official Secret Act 1972)*
☐	**RESTRICTED**	(Contains restricted information as specified by the organization where research was done)*
☑	**OPEN ACCESS**	I agree that my thesis to be published as online open access (full text)

1. I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:
2. The thesis is the property of Universiti Teknologi Malaysia
3. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
4. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

SIGNATURE OF STUDENT

MCS211043

MATRIX NUMBER

Date: 21 JULY 2023

SIGNATURE OF SUPERVISOR

Dr. Sharin Hazlin Binti Huspi

NAME OF SUPERVISOR

Date: 21 JULY 2023

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

"I hereby declare that we have read this thesis and in my opinion this thesis is suffcient in term of scope and quality for the award of the degree of Master of"

Data Science

Signature : 

Name of Supervisor : Dr. Sharin Hazlin Binti Huspi

Date : 21 JULY 2023

SENTIMENT ANALYSIS ON PERSONAL BRAND IN MALAYSIA USING HYBRID APPROACH

MOHD FIKRI BIN MOHD HANIM

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Science (Data Science)

Faculty of Computing
Universiti Teknologi Malaysia

JULY 2023

DECLARATION

I declare that this thesis entitled *"Sentiment Analysis on Personal Brand in Malaysia Using Hybrid Approach"* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	.....
Name	:	Mohd Fikri Bin Mohd Hanim
Date	:	21 JULY 2023

DEDICATION

This thesis is dedicated to my parents who encourage and support my Master's degree journey. Thank you to my supervisor for guiding me throughout this thesis. Thank you too to my supportive friends who are also struggling to finish their own thesis and has helped me with this thesis either physically or morally.

ACKNOWLEDGEMENT

I am honoured to be able to complete my project under supervision of Dr. Sharin Hazlin Binti Huspi. Her guidance helps me progress despite having no IT background, despite being stuck in loop of mistake countless time and despite losing vision of the objective, she has helped me clear the path countless times.

ABSTRACT

This thesis analyses tweets and replies regarding a selected group of Twitter celebrities and personal brands to understand their image and messaging efficacy. Twitter is crucial for personal branding and reputation management. Thus, individuals and businesses must understand how their audience views their online presence. To do this, 37 Twitter accounts were selected and their tweets and replies scraped from January to December 2021 along with their follower's count. The accounts were divided into 5 categories. The tweets were collected using the SNScrape tool, resulting in a total of 216,856 tweets and text data was cleaned and pre-processed. A hybrid approach to sentiment analysis was implemented, which used a combination of lexicon and machine learning techniques. The VADER lexicon sentiment analyser was used to classify the scraped tweets. The results showed that 59% of tweets were neutral, 32% were positive, and 9% were negative. This provided a general overview of the overall sentiment of the tweets. To further analyse the data, machine learning techniques were applied. The Support Vector Machine (SVM) and the result are discussed. Mainly due to lack of data, prediction done by SVM obtained low recall value. The result from the VADER sentiment analyser is discussed based on the category of the personal brand, engagement rate, and specific personal brand. Generally, the data shows that positive tweet will result in more positive replies and vice versa which applies to all category. In term of engagement rate, negative tweet in Technology category received a lot more engagement compare to positive tweet while positive tweet in Celebrity and Politic received more engagement than negative tweet. In Entrepreneur and Entertainer category, mean engagement is relatively high for both positive and negative. The content analysis of specific personal brands reveals that tweets that stimulate discussions, debates, controversies, and resonate with the audience's emotions are more likely to receive high engagement.

ABSTRAK

Tesis ini menganalisis tweet dan balasan mengenai sekumpulan selebriti Twitter dan jenama peribadi terpilih untuk memahami imej dan keberkesanan pemesejan mereka. Twitter adalah penting untuk penjenamaan peribadi dan pengurusan reputasi. Oleh itu, individu dan perniagaan mesti memahami cara penonton mereka melihat kehadiran dalam talian mereka. Untuk melakukan ini, 37 akaun Twitter telah dipilih dan tweet serta balasan mereka dikikis dari Januari hingga Disember 2021. Akaun-akaun ini telah dikategorikan kepada 5 kategori. Tweet-tweet dikumpulkan menggunakan alat SNScrape, menghasilkan jumlah tweet keseluruhan sebanyak 216,856. Proses pembersihan dan pra-pemprosesan teks telah dilakukan. Pendekatan hibrid untuk analisis sentimen telah digunakan, dengan menggunakan gabungan teknik leksikon dan pembelajaran mesin. Analisis sentimen menggunakan leksikon VADER telah digunakan untuk mengklasifikasikan tweet yang dikumpulkan. Keputusan menunjukkan bahawa 59% daripada tweet adalah neutral, 32% adalah positif, dan 9% adalah negatif. Ini memberikan gambaran keseluruhan tentang sentimen tweet tersebut. Untuk menganalisis data dengan lebih mendalam, teknik pembelajaran mesin telah digunakan. Mesin Vector Sokongan (SVM) dan keputusannya telah dibincangkan. Terutamanya disebabkan oleh kekurangan data, ramalan yang dibuat oleh SVM mendapat nilai recall yang rendah. Keputusan daripada analisis sentimen VADER telah dibincangkan berdasarkan kategori jenama peribadi, kadar penglibatan, dan jenama peribadi tertentu. Secara amnya, data menunjukkan bahawa tweet positif akan mendapat balasan positif dan sebaliknya, dan ini berlaku untuk semua kategori. Dari segi kadar sambutan, tweet negatif dalam kategori Teknologi menerima lebih banyak sambutan berbanding tweet positif, manakala tweet positif dalam kategori Selebriti dan Politik menerima lebih banyak sambutan berbanding tweet negatif. Dalam kategori Pengusaha dan Penghibur, kadar penglibatan purata adalah tinggi untuk tweet positif dan negatif. Analisis kandungan jenama peribadi tertentu mendedahkan bahawa tweet yang merangsang perbincangan, perdebatan, kontroversi dan lebih menyentuh emosi penonton lebih cenderung untuk menerima sambutan yang tinggi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

API	-	Application Programming Interface
CSV		Comma-Separated Values
IoT	-	Internet of Things
ML	-	Machine Learning
MT	-	Machine Translation
NLTK	-	Natural Language Toolkit
RBF	-	Radial basis function
SMOTE		Synthetic Minority Oversampling Technique
SVC	-	Support Vector Classification
SVM	-	Support Vector Machine
TF-IDF		Term Frequency-Inverse Document Frequency
URL	-	Uniform Resource Locator
UTM	-	Universiti Teknologi Malaysia
VADER	-	Valence Aware Dictionary for Sentiment Reasoning

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The pandemic of COVID-19 in 2020 has indirectly accelerate the digitization of the service industry and many other fields. There is a booming in the number of employees who work remotely, utilization of E-commerce, digital content consumption, digital health solution and many other solutions that utilize usage of IoT to keep businesses running. Dataportal reported that in 2022, user is spending 43% longer on social media and 16% more time creating and uploading videos compare to 2020.

Every day, a wide array of issues is discussed on social media sites. Generally, there is many different topic and method of presentation across different social media. Twitter is a microblogging platform that allows users to share short messages, or "tweets," with their followers. Launched in 2006, Twitter has since become a widely-used and influential social media platform, particularly in the realm of politics, news, and entertainment. In 2020, the COVID-19 pandemic led to a significant increase in social media usage, as people turned to online platforms for communication, information, and entertainment while physical distancing measures were in place. Having more than 300 million users in 2022, Twitter is one of the most used platforms. With more than 200 billion tweets per year, they can have mixture of sentiment in one thread whether it is positive, negative or neutral.

As a microblogging platform, Twitter is particularly well-suited for the analysis of public sentiment, as it allows users to share their thoughts and feelings on a wide range of topics in real-time. This has made it a popular platform for researchers studying public opinion, as well as businesses and organizations looking to gauge the response to their products, services, or messaging. Sentiment analysis usually used by

businesses to identify pain points that might need an improvement. Given the frequent criticisms about the site and how it can lead to such abusive discourse, the cancellation of celebrities, and our society's reliance on it for news, it can be difficult to see the platform's benefits. Despite this, there is also a lot more people using social media to create positive awareness, teaching lesson, promote businesses or recruiting for a job. Twitter might generally know as a casual social media; thus, this type of social media usually gives a more genuine impression of someone's personality and interest.

The purpose of this study is to examine the sentiment of tweets about Twitter celebrities and personal brands in order to gain insight into their perceived image and the effectiveness of their messaging. Analyzing the sentiment of tweets about these individuals enables us to gain insights into how they are perceived by the public and the extent to which their personal brand resonates with their followers. This research will contribute to our understanding of the role that social media plays in forming public opinion as well as the influence that influencers have on the people, they have following them.

1.2 Problem Background

Sentiment analysis is a form of data mining that identifies specific data from a given piece of content and labels it as positive, negative, or neutral. Conducting a sentiment analysis on textual data is a method that is prevalent among firms that wish to monitor how customers feel about a certain product or service and gain a better understanding of what those customers want.

Twitter is a social media platform that is utilised by a lot of people and allows users to share their thoughts and opinions in real time over a number of different issues. This has made it a popular platform for researchers studying public sentiment, as well as businesses and organizations looking to gauge the response to their products, services, or messaging. Twitter celebrities, also known as "influencers," are individuals who have gained a large following on the platform and often use it as a platform to promote their personal brand or business. Personal brands, on the other

hand, are the way in which individuals present themselves and their professional identities online.

Recently, a surge of new influencers has emerged, driven by their expertise or passion for professional skills. These influencers vary in their approach, with some focusing on sharing useful knowledge while others adopt a contrasting stance. However, regardless of the influencer's sentiment, it is common to observe mixed sentiments in the replies they receive. This suggests that Malaysians' reactions to positive tweets may not always be positive, and vice versa. Analysing the sentiment of tweets related to these individuals and their personal brands can offer insights into their perceived image and the effectiveness of their messaging.

1.3 Problem statement

As the word implies, microblogging is the posting of short messages such as "I am eating lunch" and is a passive kind of blogging. Microblogging platforms allow users to broadcast and share information about their daily activities, views, news headlines, current status, and other interests in a simple and straightforward manner. There are also commercial or purposeful microblogs used to promote websites, services, goods, or persons via the use of microblogging on prominent platforms such as Twitter, Facebook, etc. as marketing and public relations services. As of the end of April 2022, statistics indicate that Twitter had an active daily user base of 211 million. This signifies the significant reach and impact of the platform. Furthermore, Twitter has witnessed a growing trend of influencers leveraging the platform to build their businesses and personal brands.

Despite the potential value of analysing Twitter sentiment on celebrities and personal brands, there has been limited research on this topic. This study aims to fill this gap by examining the sentiment of tweets and replies about a selected group of Twitter celebrities and personal brands in order to understand their perceived image and the effectiveness of their messaging. The research questions guiding this study are:

How is the sentiment of tweets and replies about Twitter celebrities and personal brands related to their perceived image?
How does the sentiment of tweets and replies about Twitter celebrities and personal brands relate to the effectiveness of their messaging?
How do different categories of personal brands reflect on the personal brand's tweet polarity and the replies polarity?

By including both tweets and replies in the analysis, this study will provide a more comprehensive understanding of the sentiment surrounding these individuals and their personal brands on Twitter. This study will contribute to our understanding of the role of social media in shaping public opinion and the influence of influencers on their followers. It will also provide valuable insights for individuals and organizations looking to use Twitter as a platform for promoting their personal brand or business.

1.4 Aim and objective

There are several reasons to do influencer sentiment research. To measure public opinion, you may want to look at how positive or negative sentiment is toward a certain influencer. The aim of this study is to examine the sentiment of tweets and replies about a selected group of Twitter celebrities and personal brands in order to understand their perceived image and the effectiveness of their messaging. The main target of this project is to classify the tweet and it's replies into positive, negative and neutral sentiment using the proposed hybrid approach.

Objective of this project:

1. To perform classification of a Twitter post (influencer's tweets and the replies) sentiment using hybrid approach
2. To measure the engagement rate of each Tweet using personal brand's number of followers
3. To investigate the relationship between the sentiment of tweets and replies, and the engagement rate of tweets of personal brands

1.5 Scope of the project

The scope of the project will be focusing on the profile of the selected Twitter user. All tweet from 1st of January 2021 to 31st December 2021 will be collected and analysed. Twitter data by the following characteristic will be retrieved from Twitter by using SNScrape;

(a) This study will focus on tweets and replies about a selected group of 37 selected Twitter account with minimum of 3000 followers based in Malaysia

Twitter account with poor engagement and level of activity despite having good followers will be removed from the list of candidates

The analysis will be limited to tweets and replies posted in 2021

The study will not include private or deleted tweets or replies

1.6 Significant of study

Celebrities are increasingly relying on social media as a medium to communicate with their audience and market their work because of its growing importance. Celebrities often utilise social media platforms in order to keep their followers up to speed on both their personal and professional life, to interact with their devoted following, and to construct and preserve their own personal brands. Analysing the sentiment of celebrities' social media posts allows us to delve into their strategic use of social media platforms in shaping their public image and influencing public perception. This information can be valuable for understanding the role that social media plays in the personal branding strategies of celebrities and the impact it has on their public image. Additionally, studying the sentiment of celebrities' social media posts can provide insights into the ways in which celebrities engage with their audience and use social media to influence public opinion.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The literature review will begin by discussing the concept of personal branding and the role of social media in shaping celebrity image. It will then examine the various factors that influence the behavior of celebrities on social media. The review will also consider the potential impact of social media behavior on celebrity reputation and the effects of negative behavior on their public image. Next, the literature review will explore the approaches and methodologies used in previous research on this topic and identify any trends that need to be considered before initiating the experiment. The comprehensive overview of the current state of the art in sentiment analysis will include a discussion on how the rate of engagement is calculated based on industry practice. Following that, an exploration of content analysis will be presented to provide a rough idea of its principles. Finally, the importance of sentiment analysis in personal branding will be highlighted.

2.2 Human behavior in social media

In this day and age, the use of social media has evolved to the point where it is now an essential component of our everyday lives. It makes it possible for us to connect with other people and to share our opinions and experiences with a wide audience. However, the way people behave on social media can often be vastly different from their offline behavior. Understanding these factors provides insights into how people utilize social media for personal branding and the effects it has on their relationships, both online and offline, as well as their overall well-being.

2.2.1 Twitter

Twitter is a social media platform where user is able to send and read short text, video or voice messages called "tweets.". Tweets can only be 280 characters long, and users have the option of making them public or keeping them private. Twitter is well-known for its ability to link individuals from all over the world and for its real-time, public character. Recently recorded statistic shows that there is 1.3 billion Twitter account and 35% of them are monthly active user (Daniel Ruby, 2022). Individuals, corporations, and organizations frequently use it to share their news, thoughts, and information with their followers and to engage in conversation with those followers.

2.2.2 Social media professionalism

Social media professionalism is important because it can have a significant impact on an individual's reputation and career and consequently, the company they are working with. Almost everyone in today's digital world uses social media platforms to connect with others, share information, and build their own personal brands. However, social media platforms also have the potential to be dangerous tools. A person's reputation can damage when they post inappropriate or unprofessional content on their social media accounts, which can also hurt their image or the company they are working with. Marie E. (2017) stated in her article that social media screening is inefficient, this is due to ability of social media to configure the visibility of certain post or basically hide the inappropriateness for the purpose of screening. Another alternative is to implement real-time processing of the social media for screening purposes.

In 2016, Martina Temmerman studied how journalists behave in social media based on their employment status. One part of the study was to find out if employed journalists act differently or have different priorities in social media compared to freelance journalists. The study found that employed journalists were more likely to use social media for news gathering and dissemination, while freelance journalists were more likely to use it for networking and self-promotion. Additionally, the study

found that employed journalists were more likely to adhere to ethical guidelines and professional standards on social media, while freelance journalists were more likely to engage in opinionated or controversial content. These findings imply that work position can have a major impact on the way in which journalists utilize social media, and they highlight the need for future research on this subject.

These findings have implications for all social media users, as they suggest that employment status can affect the way people use these platforms. For example, employed individuals may be more likely to use social media for professional purposes, such as news gathering and dissemination, networking, or self-promotion. On the other hand, freelance or independent individuals may be more likely to use social media for personal or leisure purposes, such as sharing opinions or engaging with friends and family. Additionally, the study's findings on the impact of employment status on adherence to ethical guidelines and professional standards on social media highlight the need for all users to be aware of and consider these factors when using these platforms. In summary, this research suggests that the way people use social media can be influenced by their employment status, and that it is important for all users to be mindful of these factors when engaging with these platforms.

2.2.3 Personal Branding and professional profile

Personal branding is the process of developing and promoting a distinct public image or identity associated with a person's name or reputation. Creating a distinct personal style or image, communicating consistently and effectively through various channels (such as social media or personal websites), and developing a strong network of contacts and supporters can all contribute to this.

In today's digital world, the landscape of personal branding has significantly evolved. Unlike in the year 2000, where only actors, musicians, athletes, and models could easily garner attention and build networks, now anyone can create and establish their own personal brand. Personal branding is the most effective method for demonstrating your one-of-a-kind qualities, areas of expertise, and distinctive values

or features. It includes all of the different methods that you develop a public persona for your company and advertise yourself to the public.

Personal branding and having a professional profile are two notions that are connected but still separate from one another. Personal branding refers to the process of developing and publicizing an easily recognizable and one-of-a-kind brand identity, whereas a professional profile is a compact overview of an individual's qualifications, experiences, and successes in their field. However, personal brand can still be professionalized by bringing value into the brand. For example, Ali Abdaal is one of the personal brands actually gives value to people. He started by doing video about his studying journey as a Medicine student in Cambridge University and now he's an entrepreneur who help company work on their marketing programs. Figure 2.1 presenting his website where he lay all of his works which include podcast, video course and many other informative materials.

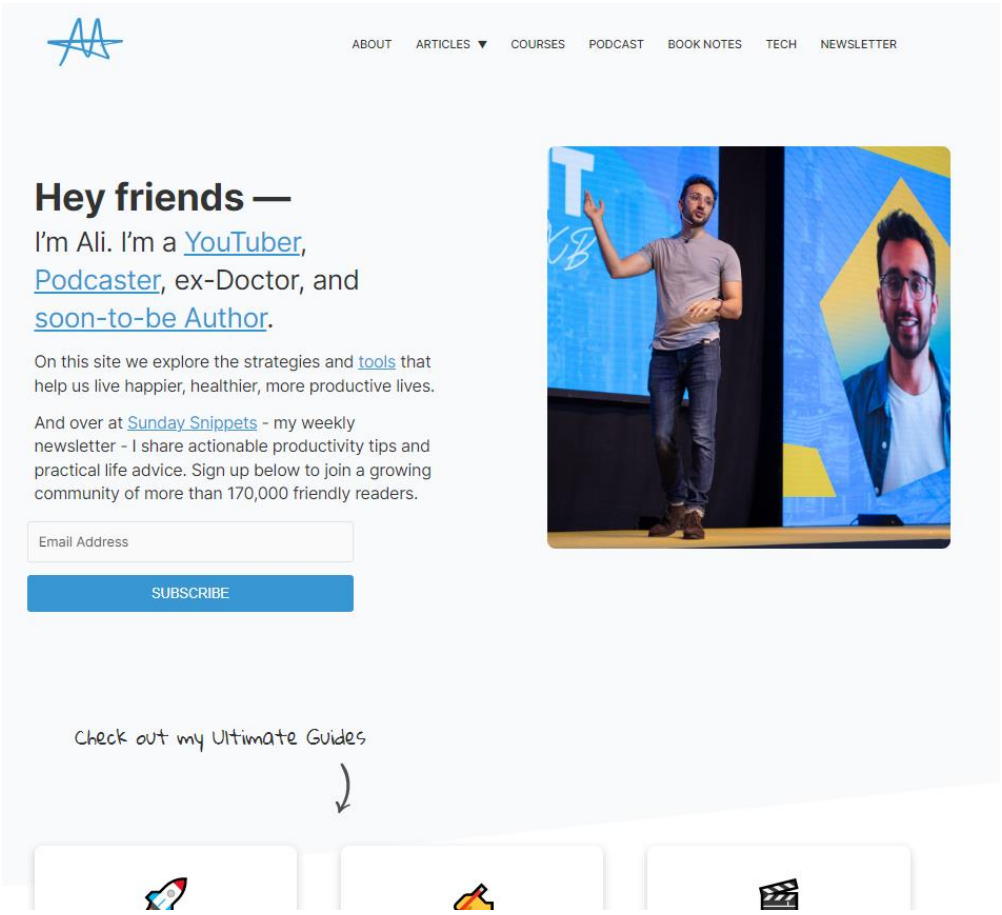


Figure 2.1 Ali Abdaal's websites (Retrieved from: https://aliabdaal.com/)

2.3 Sentiment analysis

Sentiment analysis which also known as opinion mining, is a branch of research that investigates how individuals feel or think about a variety of topics, including but not limited to issues, events, people, problems, programmes, goods, and organisations, as well as the characteristics of these entities (Yue et al., 2019). Company use sentiment analysis to understand their audience or target market not just how much the audience mention the brand, but also what they feel about the brand. The data contained in tweets are extremely diverse and unstructured, and, depending on the context, they can be either good or negative, or even neutral (A. and Sonawane, 2016).

2.3.1 Sentiment Library

A library of sentiment analysis is a set of tools and algorithms that can be used to text data in order to perform sentiment analysis on such data. There is a wide variety of sentiment analysis libraries available, each of which possesses its own exceptional collection of characteristics and capacities. The NLTK library, which is written in Python, the Stanford CoreNLP library, which is written in Java, and the TextBlob library, which is written in Python are some popular sentiment analysis libraries. The majority of the time, these libraries offer a diverse selection of pre-trained models and algorithms that are able to be applied to text data in order to carry out sentiment analysis. In addition to this, they might include tools for preprocessing and cleaning text data, as well as tools for analyzing the effectiveness of sentiment analysis models.

2.3.2 Sentiment analysis techniques

Sentiment analysis may be broken down into three primary categorization levels: the document level, the sentence level, and the aspect level. The results of document-level sentiment analysis are presented at the document level. A paper that expresses an opinion is categorized as conveying either a positive or negative view or neutral. It takes into consideration the entire article as a fundamental information unit.

Analysis of sentiment on a sentence-by-sentence basis seeks to categories the feelings conveyed in each individual sentence. Before moving on to an analysis of the polarity of feelings, it is necessary to determine whether or not the statement in question is subjective or objective. It establishes whether the statement is positive or negative in its expression of thoughts. If the sentiment analysis performed at the document level as well as the sentence level is unable to correctly understand the actual sentiment, a more granular analysis may be achieved by sentiment categorization at the aspect or feature level. (Bing Liu, 2012). Machine learning and lexicon-based sentiment analysis are the main approaches that has been used for scholar to extract sentiments.

A lexicon is a library of words that can be used to interpret the semantic orientation, leading to an increased comprehension of the valence and intensity of emotional states. Lexicon-based approaches usually use this library

There are two different approaches for lexicon-based approach. The dictionary-based approach is typically utilised by first locating the opinion seed words and afterwards employing these variables in order to tie the words to their respective synonyms and antonyms. The list will be updated to include these new words, and then the list will be utilised as the basis for the subsequent iteration. When there are no new words discovered, the iteration will be stop (Medhat, 2014). The dictionary approach performs very well when dealing with terms pertaining to the general domain, but it does not perform well when dealing with terms pertaining to the specific domain or context (Medhat, 2014), and most importantly, the dictionary approach is unable to handle informal words or slang terms.

The corpus-based method is an alternative to the lexicon-based method. The corpus-based method, in contrast to the dictionary-based method, is capable of carrying the domain specificity. In order to get accurate results, the volume of the corpus typically needs to be quite substantial and needs to be produced by an expert in the relevant field (C. Olivia, 2017).

The machine learning method is yet another important approach that is frequently utilised in the process of sentiment analysis. When training an algorithm, the machine learning approach will frequently make use of labelled datasets (also known as supervised learning). An automated classifier will take the training dataset to learn distinguishing the data according to specific set of rules. Machine learning approach has been recorded to perform much more accurate compare to lexicon approach (Pang et al., 2002; Chaovalit and Zhou, 2005). However, machine learning approach is not necessarily accurate across multiple domains, thus researchers (Andreevskaia & Bergler, 2008) has introduced hybrid approach.

Hybrid approach is combination of both lexicon and machine learning approaches. The advantage of a hybrid approach is that it lets you find and measure sentiment at the concept level, and a powerful supervised learning algorithm gives you high accuracy. The text will use a dictionary method of lexicon approach to locate similar word to classify the word's sentiment and overall sentence's sentiment are calculated based on the collective polarity (Srivats Athindran. N, 2018). The same sentences will be fed into a machine learning algorithm to measure its sentiment.

2.3.3 Sentiment analysis polarity

In sentiment analysis, polarity refers to the attitude or emotion expressed in a piece of text. The polarity of a piece of text can be positive, negative, or neutral, and is typically determined by the use of specific keywords or phrases that indicate emotion. Polarity is an important aspect of sentiment analysis because it helps to determine the overall sentiment of a piece of text and can be used to evaluate the emotional tone of a document or conversation (Cambria. E, 2016). The polarity of the text often involving experts in opinion mining as opinion mining focuses more on determining the sentiment expressed in a piece of text and polarity is one of the key factors used.

The use of emojis in Twitter has had a significant impact on how sentiment analysis is applied. The use of emojis in Twitter has made it easier for people to express

their emotions and attitudes in a concise and visual way. However, one of the common challenges in machine learning is detecting a sarcasm. To accurately determine the sentiment of tweets that contain emojis, algorithms must be trained to recognize and interpret the meaning of emojis in the context of the tweets in which they are used. Figure 2.2 shows example of emoji along with its polarity distribution (Kralj Novak et al. ,2015).

Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)	Unicode name
	0x1f602	14622	0.805	0.247	0.285	0.468	0.221		FACE WITH TEARS OF JOY
	0x2764	8050	0.747	0.044	0.166	0.790	0.746		HEAVY BLACK HEART
	0x2665	7144	0.754	0.035	0.272	0.693	0.657		BLACK HEART SUIT
	0x1f60d	6359	0.765	0.052	0.219	0.729	0.678		SMILING FACE WITH HEART-SHAPED EYES
	0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093		LOUDLY CRYING FACE

Figure 2.2 The sentiment distribution table of emoji

In lexicon-based sentiment analysis, the polarity of individual words is often evaluated by professionals with expertise in the fields of natural language processing and sentiment analysis. These experts compile dictionaries or lexicons of words and phrases that have been manually labeled with their corresponding polarity (i.e., positive, negative, or neutral). The polarity of a word is determined based on its common usage and the emotions and attitudes it typically expresses. For example, a word like "love" is likely to have a positive polarity because it is typically used to express positive emotions and attitudes. On the other hand, a word like "hate" is likely to have a negative polarity because it is typically used to express negative emotions and attitudes.

2.3.4 Sentiment analysis techniques previous research benchmark

The topic of previous research benchmarking holds significance in the field of sentiment analysis. In this subtopic, a summarized Figure 2.1 will be presented,

outlining the performance of diverse sentiment analysis methods proposed by different researchers. The table will include details about the machine learning algorithms employed, the corresponding performance metrics, and the authors responsible for introducing the methods.

The purpose of this subtopic is to provide a comprehensive overview of the current state of the art in sentiment analysis, and to compare the performance of different methods. The aim of our work is to present the material in a concise and clear manner, with the intention of offering a valuable reference for researchers and practitioners interested in sentiment analysis. By providing a comprehensive overview and comparative analysis of different methods, the aim is to facilitate understanding and provide insights in a format that is accessible and beneficial to the target audience. The Table 2.1 will be based on a review of the literature in sentiment analysis, and will include information about the most recent and relevant research in the field. Figure 2.3 presents a column chart showcasing recent studies on sentiment analysis. Its purpose is to facilitate easy comparison of the performance of different methods and enable readers to identify the strengths and weaknesses associated with each approach.

Table 2.1 Summary of sentiment analysis benchmarking

Approach	System	Accuracy	Author
Deep learning	Baidu Dictionary	95	Yuguo Tao (2022)
Hybrid	Linear SVM + Sentiment Lexicon	82.3	Mudinas, Zang (2012)
Lexicon	BPANN	95	A. Shar ma, Dey (2012)
Lexicon	Dictionary based lexicon	80	Hu and Liu (2004)
Lexicon	Pattern	69	Heidi Nguyen (2018)
Lexicon	PMI	66	Turney (2002)
Lexicon	SentiWordNet	80	Heidi Nguyen (2018)
Lexicon	VADER	83	Heidi Nguyen (2018)
ML	Neighbour sentiment algorithm	61.35	García-Díaz Pilar (2022)

Table 2.1 Summary of sentiment analysis benchmarking (Continued)

Approach	System	Accuracy	Author
ML	Neighbour sentiment algorithm	68.9	García-Díaz Pilar (2022)
ML	Gradient Boosting	87	Heidi Nguyen (2018)
ML	Linear SVC	87.3	Saura J. (2022)
ML	Logistic Regression	84.8	Saura J. (2022)
ML	Logistic Regression	90	Heidi Nguyen (2018)
ML	Maximum Entropy	81	Pang and Lee (2002)
ML	Multinomial Naïve Bayes	75.8	Saura J. (2022)
ML	Naïve Bayes	81.5	Pang and Lee (2002)
ML	Naïve Bayes	90.25	Boiy E., Hens P. (2007)
ML	Naïve Bayes with Subjectivity identification	87.2	Pang and Lee (2004)
ML	Random Forest	55.1	Saura J. (2022)
ML	SVM	82.9	Pang and Lee (2002)
ML	SVM	89	Heidi Nguyen (2018)
Pre-trained	Flair	96	Jeremy DiBattista (2021)
Pre-trained	TextBlob	66	Jeremy DiBattista (2021)
Pre-trained	VADER	69	Jeremy DiBattista (2021)

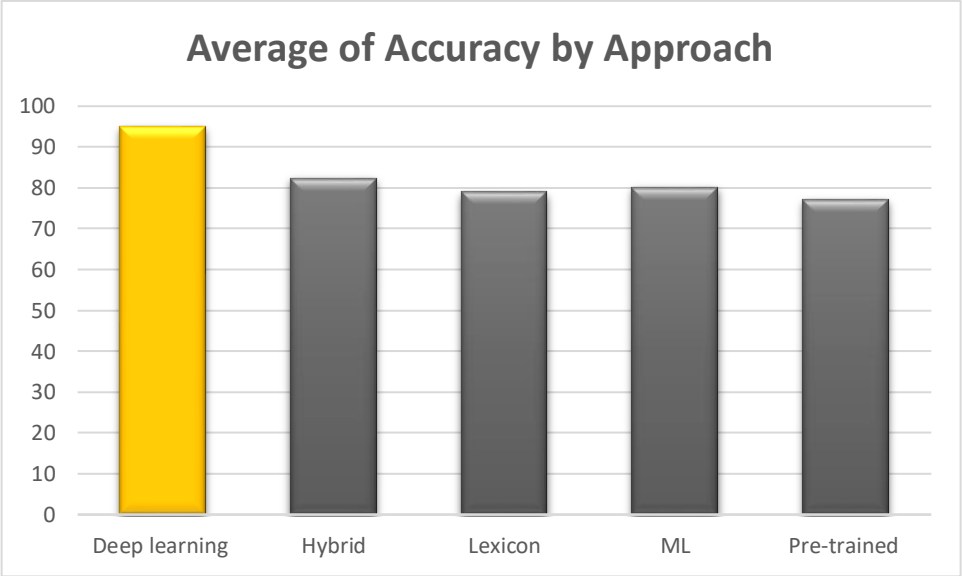


Figure 2.3 Comparison of different sentiment analysis approach

Based on Figure 2.3, our analysis reveals that deep learning techniques produce the highest average accuracy among all the approaches considered. Although there is only have one sample for deep learning, the bar chart of average accuracy obtained from the benchmarking demonstrates its superiority over other techniques. This highlights the potential of deep learning in sentiment analysis and warrants further investigation. For the purpose of doing sentiment analysis, using deep learning techniques might not be the best choice for a number of different reasons. One of the reasons behind this is that in order to reach a high level of accuracy, deep learning necessitates the use of a substantial quantity of labelled training data. The process of acquiring this might be time-consuming and costly, and as a result, it might not be possible for certain projects. Additionally, deep learning models can be computationally intensive, requiring specialized hardware and infrastructure to train and run effectively. This can make them difficult to implement and maintain in some settings.

2.4 Rate of Engagement

According to Smith and Gallicano (2015), engagement pertains to the public's sentiments towards social media content and their subsequent actions in response to it. This encompasses activities such as searching for content, leaving comments, expressing approval through likes, and sharing content with others online. Engagement rate is a crucial metric that measures the level of interaction and involvement between an audience and a particular piece of content, typically in the context of social media platforms. It provides valuable insights into how well a brand, influencer, or individual is connecting with their target audience and the effectiveness of their content strategy. Understanding and analyzing engagement rate is essential for assessing the impact and success of digital marketing efforts and making informed decisions to optimize audience engagement.

Engagement rate serves as a key performance indicator (KPI) for evaluating the effectiveness of online content and social media campaigns. Numerous researchers have extensively explored the measurement of engagement rate, taking into account various factors such as different social media platforms, types of engagement (Mariani et al, 2018), metrics used for evaluation (Bonson et al., 2015), and assigning varying weights to these metrics (Mariani et al, 2018; Vadivu et al., 2015). A wide range of studies has been conducted to develop comprehensive frameworks and methodologies for measuring engagement rate across different contexts.

However, recent articles have presented alternative approach for calculating engagement rate, deviating from the traditional approaches. These methods highlight the importance of considering follower count and total impressions as key factors in determining engagement rate.

In one article published by the CFI Team in March 2023, an approach was proposed that incorporates follower count as a metric for calculating engagement rate. This method considers the size of an individual's or organization's follower base and assesses the level of engagement relative to the total number of followers. The formula used by the CFI Team, as shown in 2.1, allows for flexibility in defining "Total

Engagement," enabling users to select specific metrics they wish to include. For instance, if a user wants to focus on likes as their chosen metric, they would need to consistently apply this metric across different time spans or when comparing different accounts for research purposes.

$$Engagement\ Rate = \frac{Total\ Engagement}{Total\ Followers} \times 100 \quad (2.1)$$

This approach, based on total impressions, is the same approach utilized by the Twitter Analytics Dashboard to calculate the engagement rate. By considering the number of times a tweet or post was loaded and displayed, this approach offers a comprehensive evaluation of content reach and effectiveness, providing valuable insights for social media analysis. Statusbrew (2023) highlights the significance of this approach in capturing the broader impact and visibility of social media content beyond traditional engagement metrics.

In summary, the approach that utilizes total impressions offers a valuable perspective that captures the overall impact and visibility of social media content. It provides a broader evaluation beyond traditional engagement metrics, acknowledging the exposure and reach of the content. However, due to the lack of data available for calculating total impressions, an alternative approach that considers the total number of followers is often employed. Although it may not provide as comprehensive a view as the total impression method, using the total number of followers still offers insights into the level of engagement relative to the size of the follower base. Therefore, in the absence of total impression data, the method based on follower count serves as a suitable alternative for evaluating engagement rate.

2.5 Content Analysis

Content analysis is a widely employed research methodology used to identify and analyze specific words, themes, or concepts within qualitative data, particularly textual data. It enables researchers to quantitatively examine and interpret the presence, significance, and interrelationships of these elements, thereby facilitating inferences about the underlying messages, authors, audiences, as well as the cultural and temporal contexts associated with the analyzed text.

In the field of content analysis, various approaches have been established, including conceptual analysis, content analysis by frequency, comparative analysis, discourse analysis, and narrative analysis (Muhammad Hassan, 2022). The selection and adaptation of these methods depend on the research question or objective and the nature of the content being analyzed.

For instance, L. Humphreys et al. (2014) employed conceptual analysis by coding each tweet based on the presence of personally identifiable information or other personal details, aligning with their research objective of investigating the extent to which Twitter users share identifiable information on the platform. Similarly, De la Torre-Dı´ez (2012) conducted a study examining the purpose and usage of Facebook and Twitter groups related to colorectal cancer, breast cancer, and diabetes. To ensure that the coding framework reflected the study's purpose, the groups were coded according to categories such as support, research, fight, or prevention, thereby capturing the majority purpose of each group.

To summarize, it is crucial to that the selection and application of appropriate content analysis approaches should be guided by the research objectives and the specific nature of the content under investigation.

2.6 Importance of Sentiment Analysis on Personal Brand Building

In today's digital age, individuals and businesses have an unprecedented ability to connect with their audience through social media platforms and online forums. This has led to an explosion of user-generated content, such as tweets, comments, and reviews, which can provide valuable insight into how a personal brand is perceived. However, the sheer volume of this data can make it difficult for individuals to manually analyze and make sense of it (Sanjeev V.,2022)

That's where sentiment analysis comes in. By using natural language processing, machine learning, and text analytics, sentiment analysis algorithms can automatically analyze and classify large volumes of text data in a fraction of the time it would take to do manually. This makes it possible for individuals to quickly and easily understand the overall sentiment of the public towards their personal brand, as well as to identify specific patterns and trends that may be influencing that sentiment.

Personal branding is the process of creating and promoting a certain image or reputation of oneself in order to establish oneself as an authority or expert in a certain field or industry. With sentiment analysis, individuals can identify what people like and dislike about their personal brand and take steps to improve it. For example, in our case, Twitter data are used to examine the pattern on specific categories and sentiment along with the reply's sentiment. In summary, using sentiment analysis to observe social media feedback can help personal brands to identify opportunities for growth and expansion, as well as to identify potential areas of risk.

2.7 Chapter Summary

This chapter discuss the theory, background in some of the related research on personal brand, sentiment analysis and data science in general. Getting grips on current position of study on specific field allows us to carefully plan our study by using their finding as guidance. This project deals with Twitter as a social media platform to observe personal's brand behaviour. Thus, the literature review mainly covers the

personal's brand behaviour on social media and integrating it with application of sentiment analysis techniques and approach and also some other approach that will be implement in analysing the result from the sentiment analysis.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The study methodology, including the stages and strategies used for the research, will be covered in more detail in this chapter. Four phases make up the total research workflow, and these will be covered in more detail in this chapter. It will be explained in full how each phase of the framework will be applied to this research, including the approaches that will be employed.

3.2 Hybrid approach

A hybrid (lexicon-machine learning) approach for sentiment analysis combines the strengths of both lexicon-based and machine learning-based approaches. Lexicon-based methods rely on pre-defined dictionaries of words and their associated sentiments and are simple to implement, but may be limited in their ability to capture complex sentiments and nuances. Machine learning-based methods can learn and adapt to complex patterns and nuances in language, but require a large amount of labelled training data and may be computationally intensive (Hassan et al, 2017). A hybrid approach can overcome these limitations by leveraging the simplicity of lexicon-based methods and the adaptability of machine learning-based methods, providing more accurate and nuanced sentiment analysis that is also more efficient and easier to implement.

In our proposed hybrid approach, we leverage the power of VADER (lexicon-based) and SVM (machine learning) models. The cleaned dataset will initially be processed through the VADER classifier, which provides sentiment labels. We then split the dataset, using the train set from the VADER classifier to train our SVM model.

By combining the simplicity of lexicon-based methods with the adaptability of machine learning-based approaches, our hybrid approach aims to deliver more accurate and nuanced sentiment analysis that is efficient and easy to implement.

3.3 Operational Framework

The research methodology framework consists of four phases: literature review and problem definition, data collection and pre-processing, experiment and evaluation, and documentation. In the first phase, a comprehensive literature review was conducted to gather existing knowledge and insight that will be needed to run this experiment and discuss the finding in detail. The second phase involve collecting the required Twitter data for sentiment analysis. This includes identifying the target group of Twitter personal brands, scraping their tweets, replies from each tweet and the number of followers of each personal brands. The data are then cleaned to increase efficiency of sentiment analysis process. In the third phase, the pre-processed data is used to perform sentiment classification using the hybrid approach, specifically combining VADER (lexicon-based) and SVM (machine learning) methods. The sentiment classification involves categorizing each tweet and its associated replies into positive and negative sentiment categories. The result from the hybrid sentiment analysis will be evaluated and the finding from the sentiment analysis will be discussed. The final phase of the research methodology framework involves documenting the research process, findings, and conclusions. Whole visualisation of the framework will be shown in Figure 3.1.

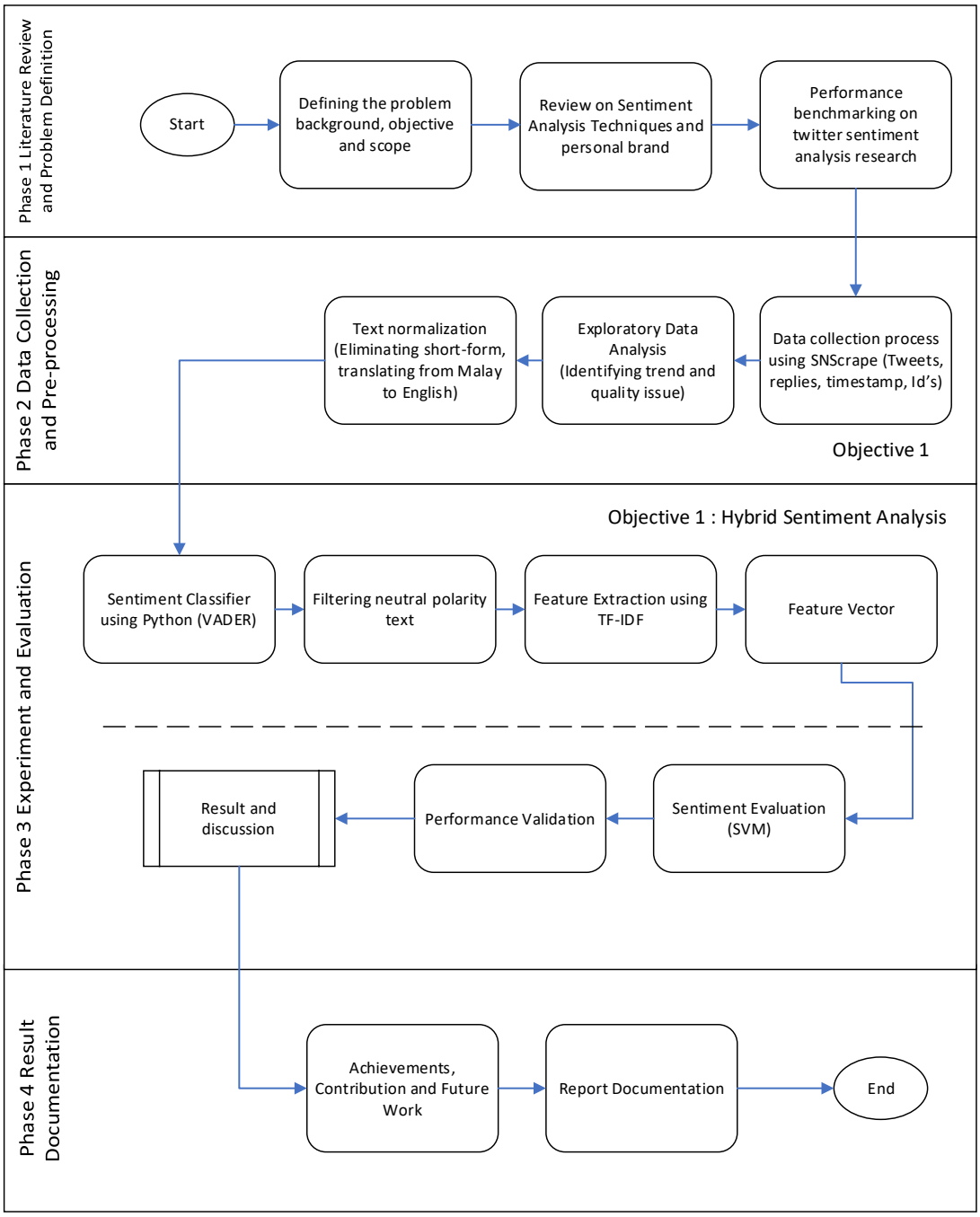


Figure 3.1 Operational framework of the methodology

3.4 Phase 1: Literature Review and Problem Definition

Phase 1 of the project involved conducting a literature review and formulating the research problem, as reported in Chapter 2. This phase included an understanding of the implementation of sentiment analysis techniques in the context of personal branding, as well as the tools and machine learning approaches chosen for this project.

The literature review provided a summary and critical analysis of the existing research on sentiment analysis and its applications in social media and branding, and helped to identify the key challenges and opportunities in this area. This phase also involved defining the research question and hypotheses, as well as the objectives and methodology of the project, which provided a foundation for the subsequent phases of the study. In summary, Phase 1 is a very important first step in conducting a successful and meaningful project on sentiment analysis for a personal brand.

3.5 Phase 2: Data Collection and preprocessing

In this phase, the research study involves the process of scraping necessary data from Twitter and the process of preparing the data for sentiment analysis.

3.5.1 Data Collection

The data collection process is visualized in a flow chart shown in Figure 3.2. The first step of our research process involves collecting data from Twitter using the SNScrape module. The data collection process will focus on gathering tweets and replies that were posted between January 1, 2021, and December 31, 2021. To ensure the credibility and relevance of the data, only tweets from selected personal brand accounts that comply with the requirements are considered. Additionally, further observations will be conducted during the scraping process to ensure the quality and suitability of the collected data. Using SNScrape, important attribute such as username, tweet text, likes count, retweet count, replies count, source of tweet, tweet ID, conversation ID and tweet created was collected as shown in Figure 3.3. Using the list of well-known personal brand, their username was collected and used to scrape the data according to attribute listed. From total of 37 username, 216,856 tweet and replies are scraped and these raw data will be pass through pre-processing phase.

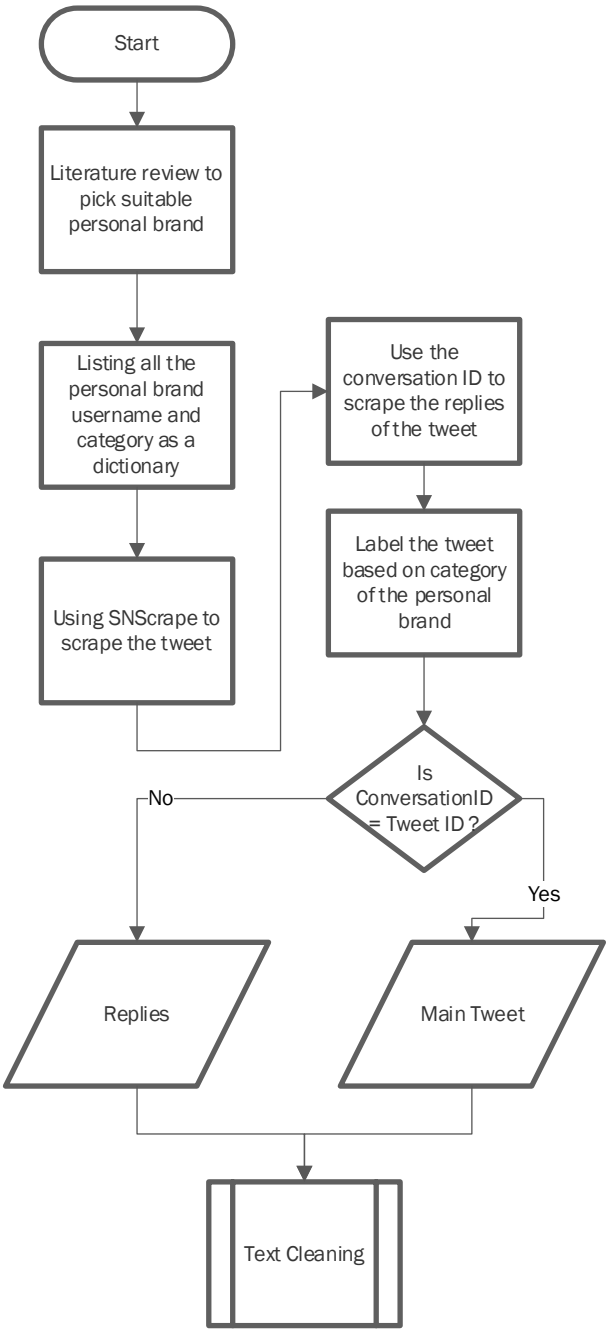


Figure 3.2 Flow chart of data collection

	Username	Text	Date Created	Number of Likes	Number of Retweet	Number of Replies	Source of Tweet	User ID	Conversation ID
0	najibfazail	Wahed Invest\n\nNew all time high return, Alha...	2021-12-29 12:22:49+00:00	296	76	9	Twitter for Android	1476166818255425536	1476166818255425536
1	fedtriyahya	Jumaat terakhir tahun 2021. \n\nSyukur Alhamdu...	2021-12-30 21:34:03+00:00	507	23	6	Twitter for iPhone	1476667927105454089	1476667927105454089
2	Pandelela_R	Hail our Lord Jesus Christ. 🙏 \n#HappyBirthdayJ...	2021-12-25 12:29:04+00:00	207	16	6	Twitter for iPhone	1474718837304029186	1474718837304029186
3	ijaicool	AYUHHH https://t.co/c2bV9gbqI8	2021-12-29 14:07:50+00:00	17	7	2	Twitter for iPhone	1476193244388024320	1476193244388024320
4	DrAmalinaBakri	If you are a Malaysian based in the UK but wou...	2021-12-20 16:39:14+00:00	1886	2393	8	Twitter for iPhone	1472969853787070464	1472969853787070464

Figure 3.3 Raw Data from SNScrape module

In addition to collecting tweets and replies, the number of followers from user profiles are scraped as part of our research process. Figure 3.4 shows the top 10 user which has most follower's count. This additional data provides insights into the popularity and reach of each user profile. By considering the number of followers, the engagement rate can be analyzed in relation to the size of their follower base.

However, it is important to acknowledge that the engagement rate calculated based on the number of followers may not be 100% accurate due to the possibility of fluctuating follower counts over time. Despite this limitation, incorporating the number of followers into the analysis allows for a better understanding of the impact and influence of each user profile within their respective audience. This additional dimension contributes to a more comprehensive assessment of engagement rates and provides valuable insights for our research.

Category	Username	followers_count
Politics	NajibRazak	4168807
Politics	Khairykj	2908740
Celebrity	CTNurhaliza11	2464469
Celebrity	yunamusic	2428232
Celebrity	Shaheizy_Sam	2171403
Celebrity	AaronDwiAziz	1997118
Celebrity	Nor4Danish	1937595
Celebrity	missfazura	1813119
Politics	anwaribrahim	1812218
Entertainer	Azfarheri	1592124

Figure 3.4 Personal brand's followers count

3.5.2 Data pre-processing

Data pre-processing is a vital step in preparing the datasets for machine learning evaluation and application. During the second phase of data collection, various measures were undertaken to ensure the data's quality and suitability for analysis. To standardize the data and ensure consistent language representation, the deep_translator library was imported. Specifically, the GoogleTranslator module from the deep_translator library was utilized to translate Malay words into English. This translation process helped in achieving a standardized language format for further analysis. Additionally, common data pre-processing techniques such as removing stop words, eliminating numbers, removing usernames, removing punctuation, eliminated URL and converting the text to lower case were implemented. These steps successfully cleaned and refined the data, enhancing the accuracy and reliability of subsequent machine learning tasks and analysis outcomes.

3.6 Phase 3: Experiment and evaluation

In this phase, the research study involves an in-depth discussion on the hybrid sentiment analysis, the result evaluation and also discussion on the finding. will visualize the detailed pipeline of the sentiment analysis from the pre-processed data to a data with sentiment label. Firstly, pre-processed data will be pass through a VADER classifier where VADER will measure the sentiment scoring of the text data. This process will produce output that include the value of polarity in each sentiment category of positive, neutral, negative and also the compound value. Using this value, sentiment classification will be done to classify each tweet into positive, neutral and negative based on the compound score. Next, since neutral sentiment is not of any use for the analysis, tweet with compound score between - 0.05 to 0.05 will be removed.

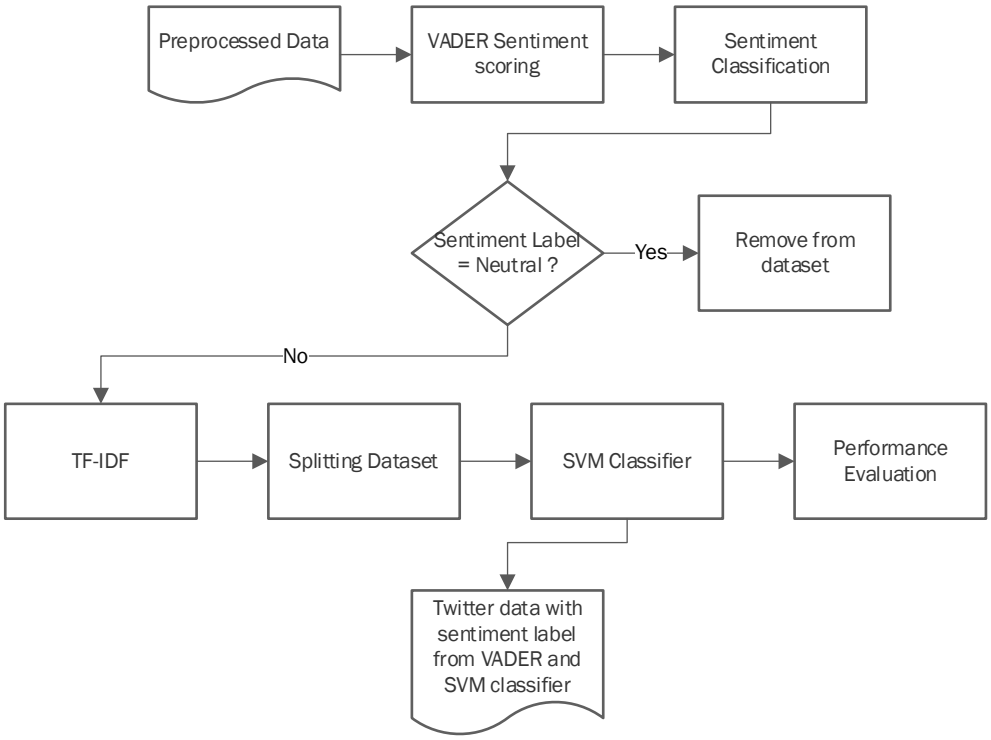


Figure 3.5 Flow of hybrid sentiment analysis

Next, to prepare the data for machine learning, TF-IDF will be use to extract feature required for SVM machine learning. This process tokenizes and vectorize the text data which make it suitable for machine learning process. GridSearch will be used to assist in conducting hyperparameter tuning. Evaluation metrics such as accuracy, precision, recall, and F1-score will be used to assess the performance of the classifier.

3.6.1 Lexicon Classifier

As a baseline for our machine learning training, VADER will be used to classify the cleaned tweet text. VADER is a lexicon-based sentiment analysis tool that was developed specifically with the goal of analyzing the sentiments conveyed in social media content. It has been used extensively for sentiment analysis on platforms such as Twitter and Facebook. VADER uses a dictionary of words and their associated sentiments to analyze the sentiment of a given text. The dictionary includes both

positive and negative words, as well as words that are neutral or that convey a more mixed sentiment. VADER produces a sentiment score for a given text, ranging from -1 (very negative) to +1 (very positive). The score reflects the overall sentiment of the text, taking into account the sentiments expressed by individual words and phrases.

In the analysis, the objective is to comprehend the sentiment expressed in the tweets and their corresponding replies. In our interpretation, when a tweet conveys a positive sentiment, it implies that the author has expressed a positive sentiment. Similarly, a positive sentiment in a reply signifies a positive response or agreement with the issue being addressed in the tweet. Conversely, a negative sentiment in a reply indicates a negative viewpoint or disagreement regarding the topic discussed in the tweet, rather than a negative sentiment towards the tweet itself. Despite this assumption will not be always correct, it will be used in order to analyze the result.

3.6.2 Feature Extraction

The next step in the process is to create a feature vector representation of the input text, which in this case consists of tweets. The feature extractor converts the text into a numerical representation that can be used by machine learning algorithms. In this project, the feature vector that will be used is TF-IDF. This feature vector is then fed into a lexicon-based algorithm, along with pairs of tags indicating the sentiment of the text (e.g., negative, positive, or neutral). The resulting model is then used to perform sentiment classification, which is the automated process of identifying and categorizing opinions in the text as negative, positive, or neutral.

3.6.3 Machine Learning Classifier

In this study, a hybrid approach will be utilized for sentiment analysis, combining both lexicon-based and machine learning methods. Specifically, VADER will be used to label the tweets text data, and Support Vector Machine (SVM) as the machine learning classifier to train and test the model. Support Vector Machines

(SVMs) are a type of supervised learning algorithm that can be used for both classification and regression tasks. The main idea behind SVMs is to find the best decision boundary, or hyperplane, that divides the different classes in the data, which in this case will be the polarity of the text. To map all the data, SVM use kernel which will allow the data to be input into a higher dimensional space where a linear decision boundary can be found which are called hyperparameters. The goal of this process is to fine-tune the SVM model by searching for an optimized set of hyperparameters, the find parameters with the performance and accuracy in sentiment analysis.

In this study, grid search will be utilized to find the most optimized hyperparameters for the Support Vector Machine (SVM) model in sentiment analysis. grid search is a commonly used method for hyperparameter optimization, where a predefined set of hyperparameters is specified, and the model's performance is evaluated for each combination of hyperparameters. To initialize the grid search process, research paper titled "An implementation of support vector machine on sentiment classification of movie reviews" by M Yulietha et al (2018) will be utilized. This paper provides insights into the initial parameter settings used for SVM in sentiment analysis of movie reviews. Referring to the mentioned research paper provides valuable information regarding the initial hyperparameters selected by the authors. This information can serve as a starting point for the grid search process, involving systematic exploration of different hyperparameter combinations. The goal is to find the optimal configuration that maximizes performance and accuracy in sentiment analysis.

3.6.4 Performance Evaluation

The performance of the sentiment analysis classifier was evaluated using several metrics to assess its effectiveness. The primary metric used was accuracy, which measures the percentage of correctly classified texts. Precision and recall were also utilized to evaluate the classifier's performance for each sentiment class, providing insights into its ability to identify positive, negative, and neutral sentiments accurately. Furthermore, the F1 score, a combined measure of precision and recall, was calculated

to provide an overall assessment of the classifier's performance. This comprehensive evaluation approach allowed for a thorough analysis of the classifier's strengths and weaknesses, enabling the identification of areas for further improvement in sentiment analysis.

3.6.5 Result and discussion

In this section, the methods used to manage the data and ensure its suitability for extracting meaningful insights will be discussed. Figure 3.6 provides an overview of the data analysis flow for the purpose of facilitating the discussion.

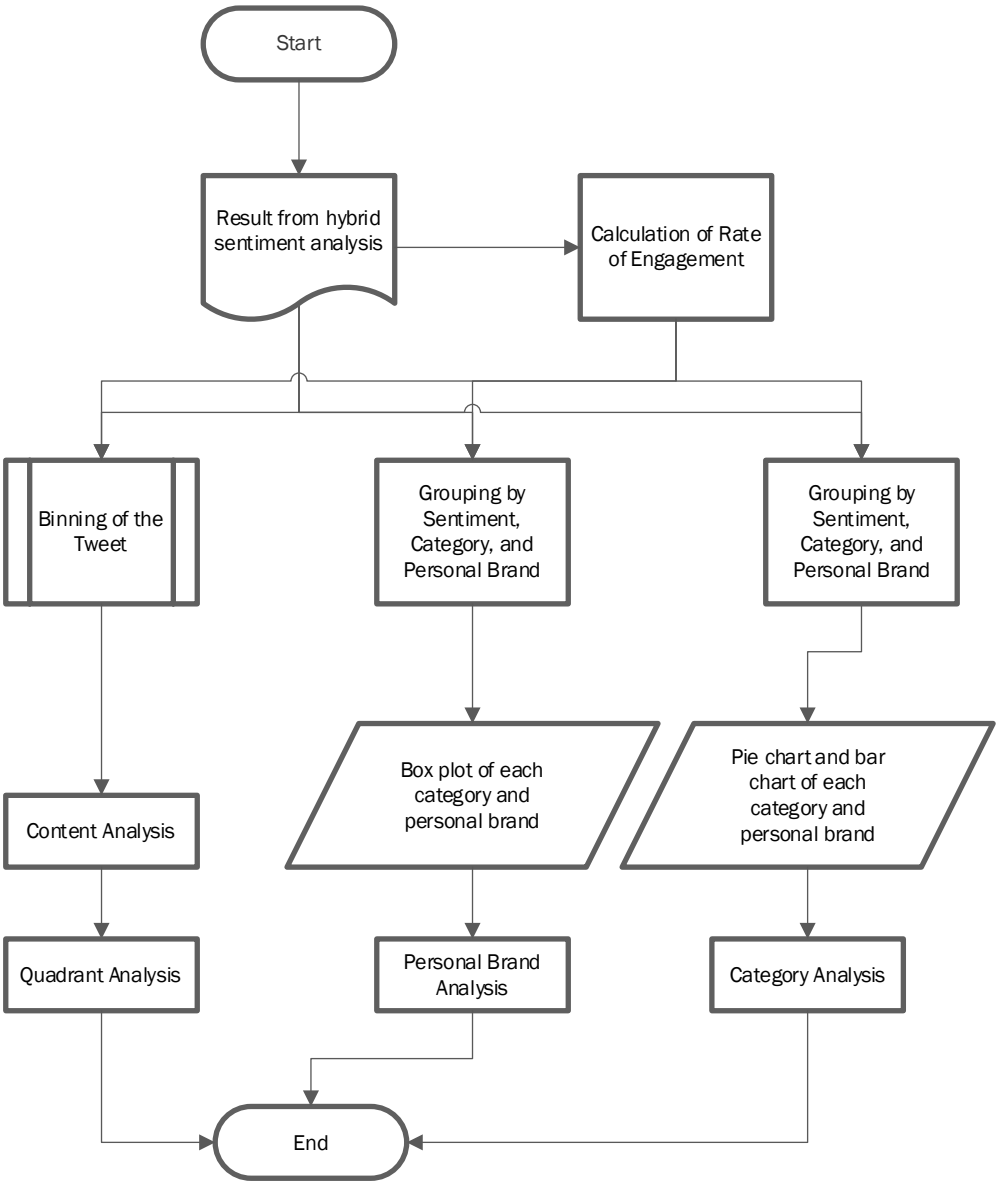


Figure 3.6 Flow chart of analysis of the data

From the output of hybrid sentiment analysis, required detail such as number of followers, likes, replies and retweets will be used to calculate the rate of engagement. This value will be used to discuss the output of the data in content analysis, personal brand analysis and category analysis.

3.6.5.1 Rate of Engagement

The engagement rate serves as a valuable metric for evaluating the effectiveness and impact of a tweet. By considering factors such as the number of likes, retweets, and replies received, as well as the total number of followers of the tweet's owner, the engagement rate can be calculated using the formula provided (3.1). This rate provides insights into the level of interaction and audience engagement generated by the tweet. This formula will give us the engagement rate for each tweet in a form of percentage.

$$Engagement\ Rate = \frac{likes+retweet+replies}{total\ followers} \times 100 \quad (3.1)$$

However, it is important to note that the number of followers collected during the data scraping process may not represent the exact number of followers at the specific time when each tweet was created. The number of followers obtained reflects the count at a particular point in time. As follower counts can fluctuate over time due to various factors, such as people following or unfollowing accounts, the follower count obtained may not be 100% accurate for the precise moment of a tweet's creation.

3.6.5.2 Binning of Rate of Engagement and Sentiment Resultant

In order to gain a deeper understanding the tweet's performance and analyze the sentiment distribution, it is important to categorize the sentiment counts into different groups. Binning the sentiment counts allows us to identify patterns, trends, and the overall sentiment landscape of personal brand's tweets at the same time reducing the complexity of the data. Since different personal brand have different data, the groups are divided based on the statistical criteria of the data itself. For Rate of Engagement, they are categorized as shown in Table 3.1 and for the sentiment resultant will be shown in Table 3.2. According to the approach outlined in Table 3.1, the mean and standard deviation are employed as benchmarks for distinguishing between High Engagement Rate and Low Engagement Rate, as opposed to using the median. One of

the primary reasons for this choice is to account for outliers within the dataset. By including outliers, there is a higher likelihood of capturing tweets with moderately high engagement rates that may have the potential to become viral. This consideration aims to mitigate the risk of having excessively large disparities between the lowest and highest values within each bin.

Table 3.1 Engagement rate category

Category	Range
High Engagement Rate	(mean + standard deviation, ∞)
Low Engagement Rate	(0, mean + standard deviation)

Table 3.2 Sentiment aggregation category

Category	Range
Low Positive	(0, mean + standard deviation)
High Positive	(mean + standard deviation, ∞)
Low Negative	(-mean - standard deviation, 0)
High Negative	(-∞, -mean - standard deviation)

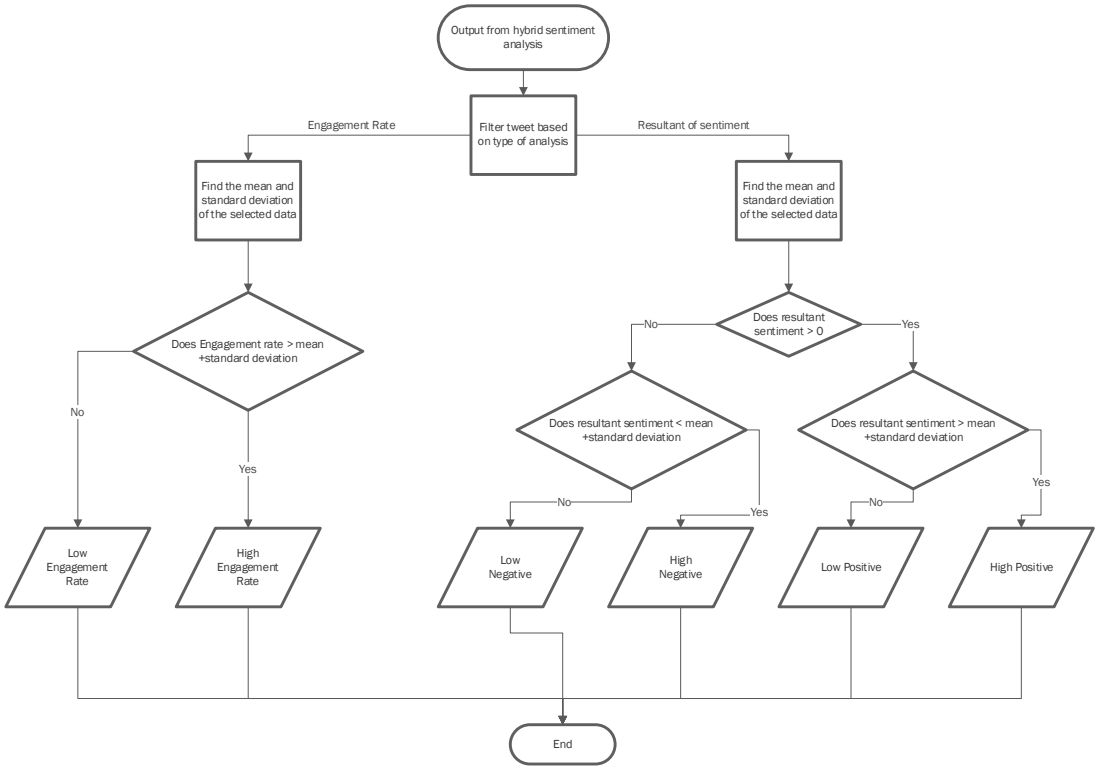


Figure 3.7 Binning of the sentiment resultant and engagement rate

As an overview of this process, Figure 3.7 will visualize the flow of the data on in the process of binning. This binning technique can help define "high" or "low" in relative terms based on each group or specific personal brand's individual performance. This allows us to assess whether a tweet's sentiment count is significantly higher or lower compared to the average sentiment count.

3.6.5.3 Content Analysis

Content analysis is a powerful method used to analyse and categorize textual data, such as social media posts, with the aim of extracting valuable insights and identifying patterns. In the context of our project, content analysis was applied to gain a deeper understanding of the nature of the personal brand's tweets and their performance in terms of engagement rate and sentiment.

The primary objective of content analysis was to categorize the tweets based on their general themes, enabling us to group them according to their main subject matter. This categorization process involved utilizing various techniques, including keyword analysis and sentiment analysis. Keyword analysis allowed us to identify specific keywords or terms within the tweets that represented different topics or themes. This helped us to gain a better understanding of the content areas covered by the personal brand and discern the subjects that attracted higher or lower engagement rates. The categorized data were then organized into a comprehensive table, providing an overview of the engagement rate and resultant sentiment for each tweet. This table served as a valuable resource for identifying patterns, trends, and areas of interest within the personal brand's tweet performance. By analysing this table, meaningful conclusions and insights regarding the effectiveness and impact of their communication strategies can be drawn.

Table 3.3 Content analysis guideline for Politic category

Label	Description
Education System	Tweet that involves giving information or discussion related to education system
Relief Effort	Focus on relief efforts, particularly in response to natural disasters, emergencies, or community support initiatives.
Political Dispute	tweets that involve political disputes or disagreements, differing opinions and arguments related to politics
Government Criticism	criticize or question the actions, policies, or decisions made by the government or government officials.
MUDA Updates	updates or information about the activities, progress, or initiatives of the organization, MUDA.
General Information	tweets that contain general information or announcements that are not specific to a particular theme or category.
Political Discussion	Tweets that initiate or engage in political discussions, debates, or conversations.
Criticism of justice system	tweets that express criticism or concerns regarding the justice system, legal processes, or judicial decisions.

Table 3.3, included in our report, presents a detailed guide on how the labels for the content categories were determined based on the analysis. Table 3.3 only included for Politic category while content analysis guideline for another category will be included in Appendix A.

3.7 Tools and Platforms

In this project, a variety of tools and platforms will be utilized to conduct sentiment analysis on a dataset of text data. The primary tool employed is Python, a widely-used programming language known for its extensive libraries and packages for data analysis and machine learning. Specifically, the SNScrape package will be used to scrape data from social media platforms, SKlearn library for machine learning tasks, the Malaya package for natural language processing of the Malay language, and the VADER package as a lexicon-based classifier to label and extract sentiments from the text data. These tools collectively enable preprocessing, feature extraction, classification, and overall comprehension of the expressed sentiments within the dataset.

3.8 Chapter Summary

In the literature review subchapter, relevant literature on sentiment analysis have been discussed and its application to social media data, specifically Twitter. Different methods and tools that have been suggested for analyzing Twitter data based on how people feel. The experiment and analysis subchapter describes the general procedure for conducting the sentiment analysis on Twitter data for this project. This includes the steps for collecting and preprocessing the data, lexicon-based labelling, as well as the specific sentiment analysis method that will be used. Finally, in the documentation subchapter, the plan for documenting the results of the sentiment analysis, including any relevant data visualizations and statistical analyses are discussed. Overall, the methodology described in this chapter aims to provide a rough approach to conducting sentiment analysis on Twitter data.

CHAPTER 4

EXPERIMENT AND EVALUATION

4.1 Introduction

This chapter will be discussing an in-depth implementation of the research methodology discussed in Chapter 3. Each phase will be elaborated which include the steps in Twitter data scraping, dataset preparation for lexicon sentiment classifier and machine learning classifier.

4.2 Dataset collection

The dataset collection for this project will begin with identifying Twitter users for analysis. A list of Twitter usernames and user IDs will be compiled by manually examining the accounts to ensure that they have relatively good engagement compared to normal users. A tweet from a statistic Twitter account, Politweet.org was also used as a reference to collect more user that suitable for our objective.

Table 4.1 List of usernames according to category

Entertainer	sofyank96
	LuqmanPodolski
	hazemanhuzir
	CeddyOrNot
	brgsjks
	Azfarheri
	Matluthfi90
Politics	chairman_GLC
	Khairykj
	MuhyiddinYassin
	limlipeng
	kuasasiswa
	NajibRazak
	n_izzah
	anwaribrahim
	SyedSaddiq
Technology	XavierNaxa
	fazlihalimmedia
	thefaizzainal
	aribismail
	acaiijawe
Celebrity	altimet
	AfikryAibrahim
	yunamusic
	CTNurhaliza11
	AaronDwiAziz
	missfazura
	Nor4Danish
	Shaheizy_Sam
Entrepreneur	FezzaHussin
	adibhazlami
	afiqnazary
	AmirulMu_min
	richardker
	IZZTAZWAR
	khairulaming

To collect the data for analysis, SNScrape tool will be used to scrape all the posts and replies from each selected user from January of 2021 until December of 2021. It is important to consider what data should be scraped in order to effectively run this project. In this project, the username, tweet text, date of tweet, number of likes, number of retweets, number of replies, the source of tweet, tweet ID and the conversation ID will be scrapped. All this data will be scrape for both main tweet and the replies. Figure 4.1 and Figure 4.2 shows the script required to scrape all the data needed for the project

```
def scrape_tweets(query,limit):
    data = []
    for tweets in sntwitter.TwitterSearchScraper(query).get_items():
        if len(data) == limit:
            break
        else:
            data.append([tweets.user.username,tweets.content,tweets.date,
            tweets.likeCount,tweets.retweetCount, tweets.replyCount,
            tweets.sourceLabel, tweets.id, tweets.conversationId])
    return pd.DataFrame(data, columns=['Username', 'Text',"Date Created",
     "Number of Likes", "Number of Retweet", "Number of Replies",
      "Source of Tweet", 'User ID', 'Conversation ID'])
```

Figure 4.1 Function to scrape tweets from Twitter

```python

def get_replies(tweet_id):
    mode = sntwitter.TwitterTweetScraperMode
    scraper_reply = sntwitter.TwitterTweetScraper(tweetId=tweet_id, mode=mode.SCROLL)
    replies = list(scraper_reply.get_items())

    # Create a list of lists containing the values for each row in the DataFrame
    data_reply = [[tweets.user.username,tweets.content,tweets.date,
    tweets.likeCount,tweets.retweetCount, tweets.replyCount,tweets.sourceLabel,
     tweets.id, tweets.conversationId] for tweets in replies]

    # Create the DataFrame from the data
    df = pd.DataFrame(data_reply, columns=['Username', 'Text',
    "Date Created", "Number of Likes", "Number of Retweet",
     "Number of Replies", "Source of Tweet", 'User ID',
      'Conversation ID'])
    return df
```

Figure 4.2 Function to scrape relies from Twitter

Both main tweet and replies dataset are merged saved as dataframe for further pre-processing. Figure 4.3 displays the sample of raw data set scraped from SNScrape. Due to limitation in quantity of data being able to scrape which is part of Twitter API free user regulation, the scraping process are split into two group. The first group is to scrape the tweet from the Politic and Entertainer category and the second group is Technology, Celebrity and Entrepreneur. After scraping, both datasets are combined using concatenate function in python into one table. Due to the presence of noisy text and abnormalities it is evident that considerable data preparation is necessary. Its quality was very poor and required some pre-processing.

	Username	Text	Date Created	Number of Likes	Number of Retweet	Number of Replies	Source of Tweet	User ID	Conversation ID
0	najibfazail	Wahed Invest\n\nNew all time high return, Alha...	2021-12-29 12:22:49+00:00	296	76	9	Twitter for Android	1476166818255425536	1476166818255425536
1	fedtriyahya	Jumaat terakhir tahun 2021. \n\nSyukur Alhamdu...	2021-12-30 21:34:03+00:00	507	23	6	Twitter for iPhone	1476667927105454089	1476667927105454089
2	Pandelela_R	Hail our Lord Jesus Christ. 🙏 \n#HappyBirthday!...	2021-12-25 12:29:04+00:00	207	16	6	Twitter for iPhone	1474718837304029186	1474718837304029186
3	ijaicool	AYUHHH https://t.co/c2bV9gbqI8	2021-12-29 14:07:50+00:00	17	7	2	Twitter for iPhone	1476193244388024320	1476193244388024320
4	DrAmalinaBakri	If you are a Malaysian based in the UK but wou...	2021-12-20 16:39:14+00:00	1886	2393	8	Twitter for iPhone	1472969853787070464	1472969853787070464

Figure 4.3 Sample of newly scraped data

4.3 Data preparation

4.3.1 Text Normalization

After the data extraction process is completed, the data will be normalized to improve the ability of the translation tool and the lexicon to recognize and process the words in the data. The normalization also involves the process of translating the Malay text to English by using Google Translator module. Figure 4.4 shows the implementation of Google Translator module to the column "Text" and the output will be inserted into a column of "translated_text". The function will automatically detect the initial language and translate it into English language. However, there are some words that Google Translator module cannot detect and translate properly, mostly it involve a typo or a short form word.

```
from deep_translator import GoogleTranslator

# Define the translation function
def translate_text(text):
    return GoogleTranslator(source='auto', target='en').translate(text)

# Use the `apply` method to translate the text in the `"text"` column
df_merged["translated_text"] = df_merged["Text"].apply(translate_text)
```

Figure 4.4 Function to translate text using Google Translator

4.4 Data Pre-processing

Data preprocessing is an additional essential step for ensuring that the text corpus has been cleaned and is acceptable for sentiment classification. The text pre-processing process for this project will involve a series of steps to prepare the data for sentiment analysis. These steps will include removing @usernames, punctuation, URLs, emojis, elements that may not be relevant for the analysis. This will help to ensure that the data is clean and accurately represents the sentiment of the text being analyzed. By carefully pre-processing the text, the accuracy and reliability of our

sentiment analysis results can be improved. Figure 4.5 to Figure 4.7 shows the script that are used to prepare the data for sentiment analysis process.

```
#remove stopwords
def remove_stopwords(text):
    text = ' '.join([word for word in text.split() if word not in (stopwords.words('english'))])
    return text

# Remove url
def remove_url(text):
    url = re.compile(r'https?://\S+|www\.\S+')
    return url.sub(r'',text)

# Remove punctuation
def remove_punct(text):
    table = str.maketrans('', '', string.punctuation)
    return text.translate(table)

# Remove html
def remove_html(text):
    html=re.compile(r'<.*?>')
    return html.sub(r'',text)

# Remove @username
def remove_username(text):
    return re.sub('@[^\s]+','',text)

# Remove emojis
def remove_emoji(text):
    emoji_pattern = re.compile("["
                           u"\U0001F600-\U0001F64F"  # emoticons
                           u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                           u"\U0001F680-\U0001F6FF"  # transport & map symbols
                           u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                           u"\U00002702-\U000027B0"
                           u"\U000024C2-\U0001F251"
                           "]+", flags=re.UNICODE)
    return emoji_pattern.sub(r'', text)

```

Figure 4.5 Function to pre-process the tweet text

```python
# Decontraction text
def decontraction(text):
    text = re.sub(r"won\'t", " will not", text)
    text = re.sub(r"won\'t've", " will not have", text)
    text = re.sub(r"can\'t", " can not", text)
    text = re.sub(r"don\'t", " do not", text)

    text = re.sub(r"can\'t've", " can not have", text)
    text = re.sub(r"ma\'am", " madam", text)
    text = re.sub(r"let\'s", " let us", text)
    text = re.sub(r"ain\'t", " am not", text)
    text = re.sub(r"shan\'t", " shall not", text)
    text = re.sub(r"sha\n't", " shall not", text)
    text = re.sub(r"o\'clock", " of the clock", text)
    text = re.sub(r"y\'all", " you all", text)
    text = re.sub(r"n\'t", " not", text)
    text = re.sub(r"n\'t've", " not have", text)
    text = re.sub(r"\'re", " are", text)
    text = re.sub(r"\'s", " is", text)
    text = re.sub(r"\'d", " would", text)
    text = re.sub(r"\'d've", " would have", text)
    text = re.sub(r"\'ll", " will", text)
    text = re.sub(r"\'ll've", " will have", text)
    text = re.sub(r"\'t", " not", text)
    text = re.sub(r"\'ve", " have", text)
    text = re.sub(r"\'m", " am", text)
    text = re.sub(r"\'re", " are", text)
    return text
```

Figure 4.6 Function to pre-process the tweet text

```python
# Seperate alphanumeric
def seperate_alphanumeric(text):
    words = text
    words = re.findall(r"[^\W\d_]+|\d+", words)
    return " ".join(words)

def cont_rep_char(text):
    tchr = text.group(0)

    if len(tchr) > 1:
        return tchr[0:2]

def unique_char(rep, text):
    substitute = re.sub(r'(\w)\1+', rep, text)
    return substitute

def char(text):
    substitute = re.sub(r'[^a-zA-Z]',' ',text)
    return substitute

# combaine negative reason with  tweet (if exsist)
df['final_text'] = df['negativereason'].fillna('') + ' ' + df['text']


# Apply functions on tweets
df['final_text'] = df['final_text'].apply(lambda x : remove_username(x))
df['final_text'] = df['final_text'].apply(lambda x : remove_url(x))
df['final_text'] = df['final_text'].apply(lambda x : remove_emoji(x))
df['final_text'] = df['final_text'].apply(lambda x : decontraction(x))
df['final_text'] = df['final_text'].apply(lambda x : seperate_alphanumeric(x))
df['final_text'] = df['final_text'].apply(lambda x : unique_char(cont_rep_char,x))
df['final_text'] = df['final_text'].apply(lambda x : char(x))
df['final_text'] = df['final_text'].apply(lambda x : x.lower())
df['final_text'] = df['final_text'].apply(lambda x : remove_stopwords(x))
```

Figure 4.7 Function to pre-process the tweet text

As shown in the Figure 3.3, the dataset currently contains the attribute scraped from Twitter. To facilitate the process of data analyzation, the data was labelled based on its status which is “Main post” or “replies” and also the category of the owner of the conversation ID which are the category state in Table 4.1. By doing this, the data can easily be distinguished and facilitate extraction of insight from the analysis.

4.4.1 Tweet polarity and classification

In this study, VADER (Valence Aware Dictionary for Sentiment Reasoning), a model that uses a human-centric approach, to calculate the polarity of sentiment will be used for each message in our dataset. The text column of the data is given more

attention VADER's classifier function are used to obtain the sentiment and polarity values. These values were then added to a new column in the dataset, allowing us to easily visualize and analyze the overall sentiment of the messages. This label will be a reference data which will help train the machine learning in the next phase. For this study, VADER will assign each text and calculate the polarity of a given text. The polarity of a text ranges from -1 to 1, with -1 being extremely negative, 1 being extremely positive, and 0 being neutral. Figure 4.8 shows python function on the implementation of VADER sentiment classifier onto the “text” column. Figure 4.9 shows the sample of text with the polarity score and the sentiment value included.

```python
def sentiment_vader(sentence):

    # Create a SentimentIntensityAnalyzer object.
    sid_obj = SentimentIntensityAnalyzer()

    sentiment_dict = sid_obj.polarity_scores(sentence)
    negative = sentiment_dict['neg']
    neutral = sentiment_dict['neu']
    positive = sentiment_dict['pos']
    compound = sentiment_dict['compound']

    if sentiment_dict['compound'] >= 0.05 :
        overall_sentiment = "Positive"

    elif sentiment_dict['compound'] <= - 0.05 :
        overall_sentiment = "Negative"

    else :
        overall_sentiment = "Neutral"

    return negative, neutral, positive, compound, overall_sentiment
```

Figure 4.8 Function to analyse the sentiment of text using VADER

	Negative	Neutral	Positive	Compound	Overall Sentiment
0	0.0	1.0	0.0	0.0	Neutral
1	0.08	0.92	0.0	-0.1027	Negative
2	0.0	1.0	0.0	0.0	Neutral
3	0.0	1.0	0.0	0.0	Neutral
4	0.0	1.0	0.0	0.0	Neutral
...	...	...	...	...	...
141485	0.0	0.5	0.5	0.4588	Positive
141486	0.0	0.782	0.218	0.7506	Positive
141487	0.0	0.575	0.425	0.8625	Positive
141488	0.0	1.0	0.0	0.0	Neutral
141489	0.0	0.495	0.505	0.7989	Positive

Figure 4.9 Sample of result from VADER module

The resulting polarity distribution of the tweets and replies are shown in Figure 4.10. As shown, the count of text with neutral polarity is 49.7% which is about half of the dataset. VADER sentiment classifier is not able to classify Twitter text with Malay word, thus the text is labelled as neutral as words that cannot be recognized by the classifier will be categorized as neutral words. Due to neutral text having no benefit for the project, neutral tweet will be effectively removed from the dataset. Other than language barrier, leaving this imbalance category distribution would generate a poorly trained model which result in bad output for this study. To rectify the situation with imbalance dataset, SMOTE will be implemented to oversample the minority category.

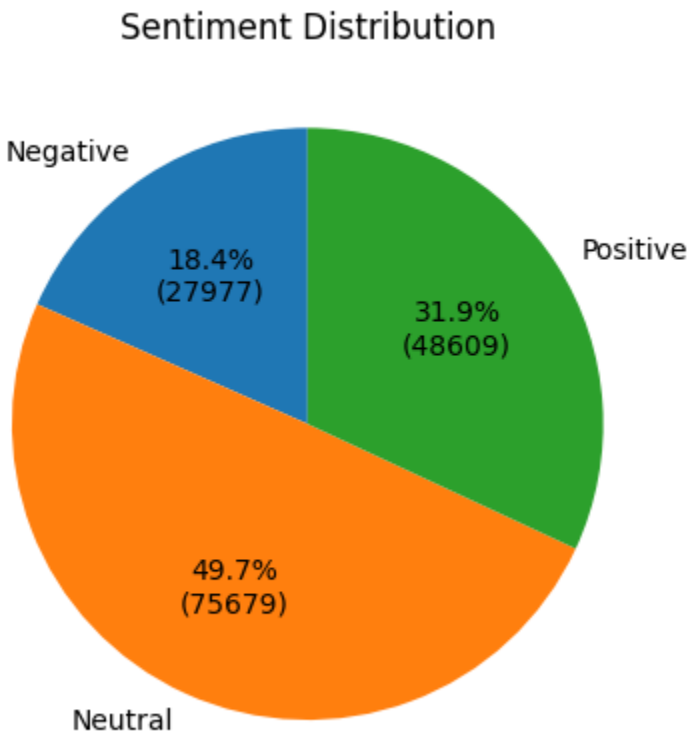


Figure 4.10 Distribution of overall sentiment

4.5 Data Exploratory Analysis

The amount of data collected for this study varies by category, with a total of 232410 tweet and replies from January 2021 to December 2021. The categories included in the study are Entertainer, Politics, Technology, Celebrities and Entrepreneur. The user distribution in each category is shown in Figure 4.11.

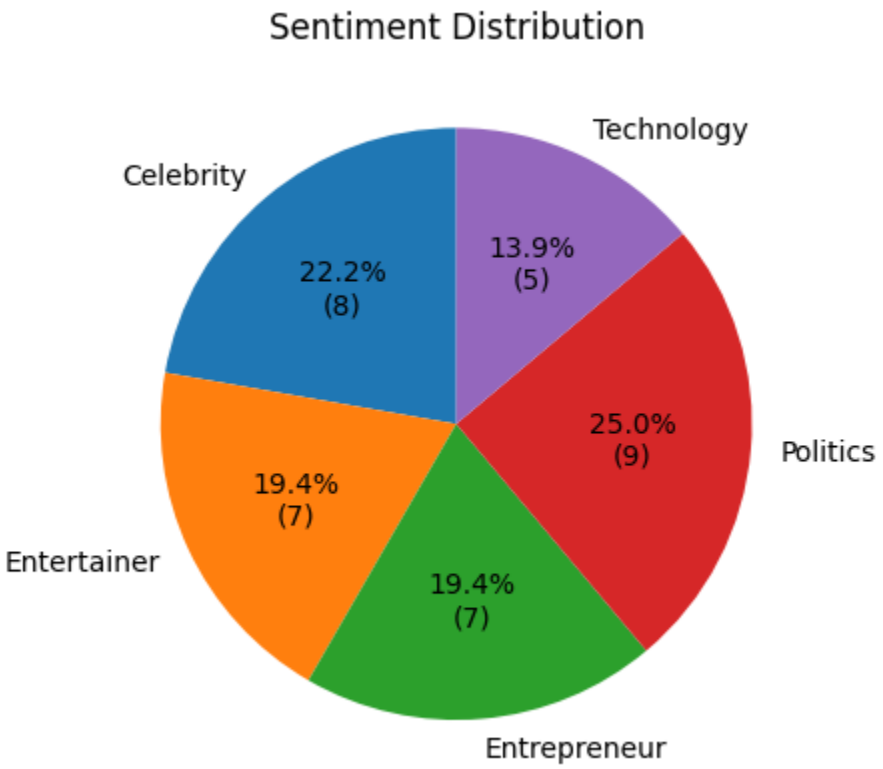


Figure 4.11 Distribution of users by category

The data is in CSV format and consists of 8 attributes for each tweet or reply. Table 4.2 will be presenting the summary of the attributes that will be scaped from Twitter for both reply and tweets.

Table 4.2 Data description

Attribute	Description	Value
Conversation_id	The unique identifier of the parent tweet	1476417842723520000
username	The user's Twitter handle or screen name	XavierNaxa
Number of Likes	Tweet's likes count	1
Number of Retweet	Tweet's retweet count	1
Number of Reply	Tweet's Reply count	1

Table 4.2 Data description (continued)

Source of tweet	Type of device the tweet user used to tweet	Twitter for iPhone
Date Created	Time of the tweet posted on Twitter	2021-12-30 05:00:18+00:00
status	The status of the text whether it is a reply or main post	Main post
Category	The category of the username. If it is a reply, it will refer to the owner of the Conversation ID	Technology
Polarity Score (positive, negative, neutral, compound)	Polarity score generated by VADER classifier	0.5, 1.0, 0.08
Sentiment	Sentiment generated by VADER classifier	Neutral, negative, positive

The dataset distribution is imbalance throughout the category as Figure 4.12 shows that the number of texts from Politics category is the highest which is at 38.6%. Having this bias will not affect the sentiment analysis as the category is not an output of the analysis. However, it is a good practice to take note on the quantity to ensure a better understanding on the dataset and the result.

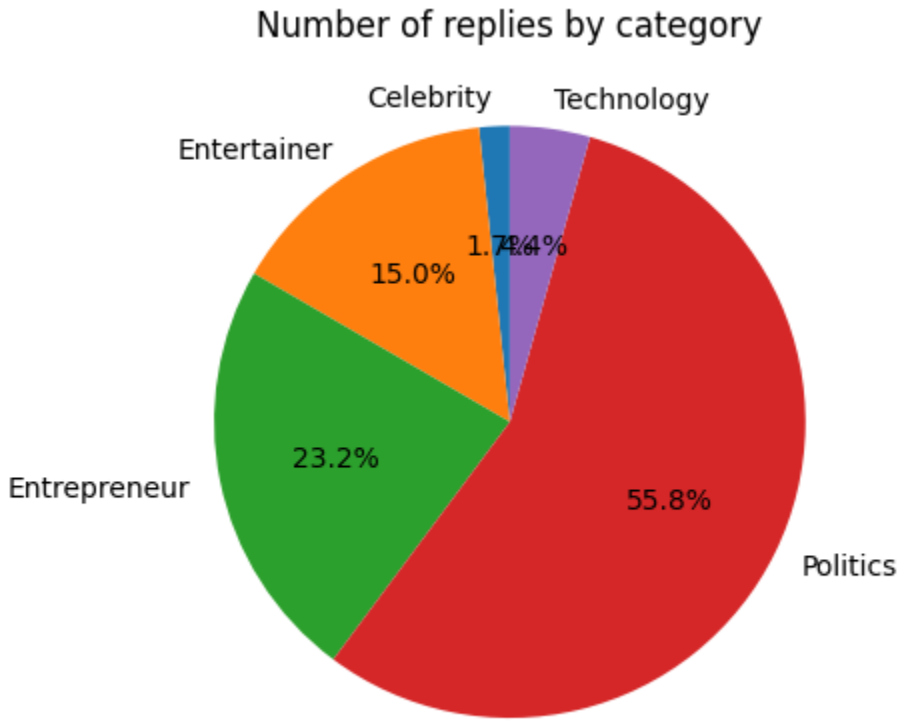


Figure 4.12 Distribution of replies by category

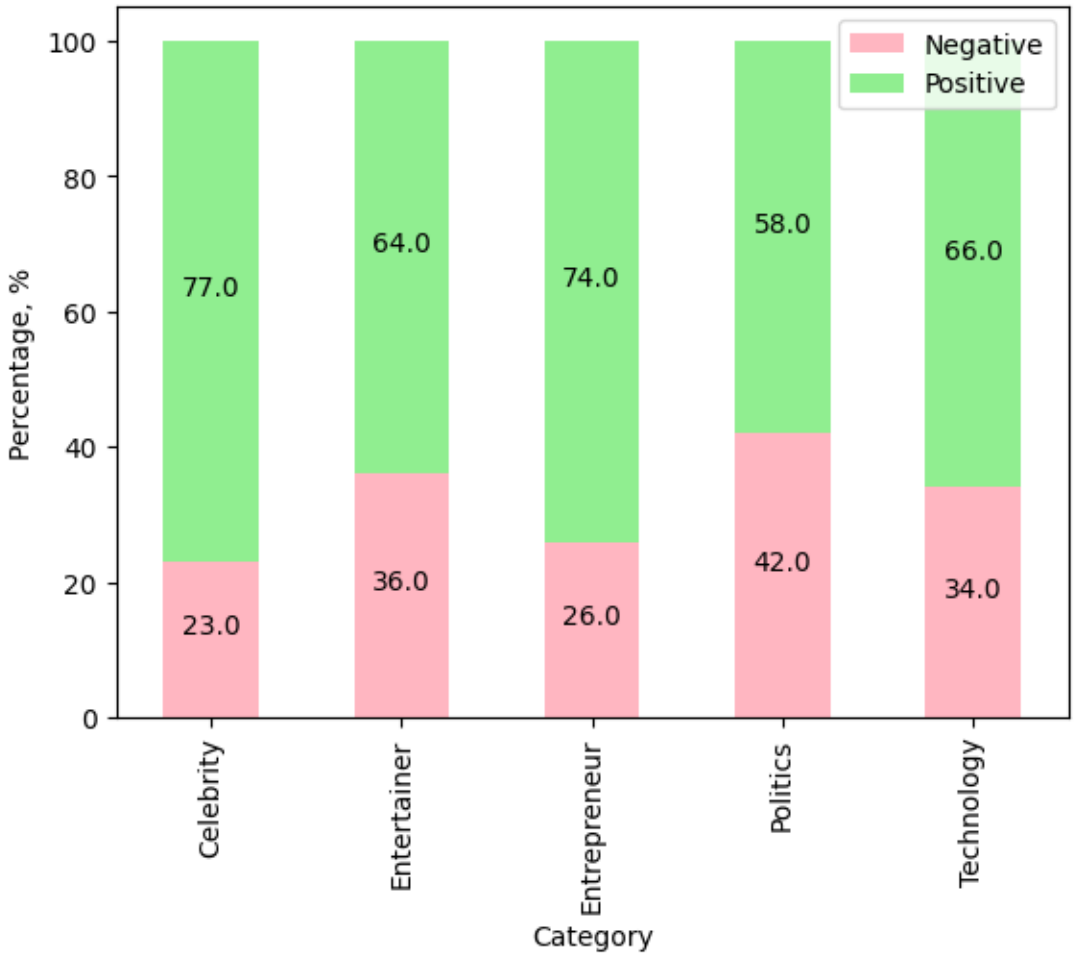


Figure 4.13 Sentiment distribution in percentage

The word cloud for this dataset is divided by 6 categories of user, as shown in Appendix B. The word cloud presented the total frequency of words by the size of the words. The words with the highest frequency over the whole corpus will be having the biggest size while words with low frequency will be having small size scattered around the space. Knowing the frequency of the words help us roughly guess the trend of the word as to which words is an important word or even just a frequently used stop words that are missed during pre-processing.

4.6 Splitting (Training and Testing data)

In order to develop and evaluate our sentiment analysis model, it is crucial to split the pre-processed data into separate training and testing sets. This process ensures that the model is trained on a diverse range of data and evaluated on unseen examples. In our case, the data are split in a ratio of 80% for training data and 20% for testing data, maintaining the proportion of data categories in terms of sentiment across both sets.

Initially, the data is divided into positive and negative sentiment categories based on the sentiment scores obtained from the VADER algorithm. This step allows us to separate the data into two distinct categories representing positive and negative sentiments.

Subsequently, the positive sentiment data is further divided into a training set and a testing set. This is achieved using the train_test_split function from the Sklearn library. By specifying a test_size of 0.2, 20% of the positive sentiment data for testing are allocated, while the remaining 80% is reserved for training purposes. This ensures that the sentiment analysis model receives a diverse range of positive sentiment examples during training.

To create the final input datasets for the model, the training and testing sets for positive sentiment are combined with the negative sentiment data. This merging process is accomplished using pd.concat function, which concatenates the respective sets to create the merged training and testing datasets. These merged datasets, namely X_train_merge, X_test_merge, y_train_merge, and y_test_merge, serve as the input for the subsequent stages of model training and evaluation.

By performing this stratified data splitting procedure helped ensure that our sentiment analysis model is trained on a balanced representation of positive and negative sentiment examples. Moreover, evaluating the model on the separate testing data provides an accurate assessment of its generalization capabilities and performance

on unseen data. This approach enables us to effectively train and evaluate our sentiment analysis model, leading to reliable and robust results.

```python
# Split the data into positive and negative sentiment
pos_data = df_celebrity[df_celebrity['sentiment'] > 0]
neg_data = df_celebrity[df_celebrity['sentiment'] <= 0]

# Shuffle the data and split into training and testing sets with
# equal amounts of positive and negative sentiment
pos_train, pos_test = train_test_split(pos_data, test_size=0.2, random_state=42)
neg_train, neg_test = train_test_split(neg_data, test_size=0.2, random_state=42)
train = shuffle(pd.concat([pos_train, neg_train]))
test = shuffle(pd.concat([pos_test, neg_test]))
```

Figure 4.14 Python code for Stratified sampling

4.7 Feature Extraction

The data that are cleaned, split and processed will be pass through a feature extraction process which in our case will be using TF-IDF. TF-IDF will generate a score that represent the weight of the word. TF-IDF method from Sklearn module will be applied to calculate this parameter. Figure 4.15 shows the function that will calculate and transform the text into numerical form. The output of this process will be in form of matrix that consist of document ID, token ID and the TF-IDF score.

```python
tfid = TfidfVectorizer(use_idf=True)
X_final = tfid.fit_transform(X)
```

Figure 4.15 Python code calculate TF-IDF score

4.8 SMOTE

In order to address the issue of class imbalance in our sentiment analysis dataset, a technique called SMOTE (Synthetic Minority Over-sampling Technique) was applied. Referring to Figure 4.13, the have major class imbalance which could affect our model's ability to predict. By using SMOTE, the representation of the minority class in the dataset can be effectively increased, thus balancing the class distribution.

```
# Perform oversampling on the training data using SMOTE
sm = SMOTE(random_state=42)
X_train_celeb_smote, y_train_celeb_smote = sm.fit_resample(X_train_celeb, y_train_celeb)
```

Figure 4.16 Python code to apply SMOTE

4.9 Performance evaluation

Prepared data will past through our SVM model for sentiment analysis. GridSearchCV was applied in order to find the best hyperparameter for our SVM model as shown in Figure 4.17. GridSearchCV is a systematic approach that exhaustively searches for the best combination of hyperparameters from a predefined grid of parameter values. In this case, the hyperparameters being tuned for the SVM model include the kernel type (linear, poly, rbf, sigmoid), the regularization parameter C (0.1, 1, 10, 100), the gamma parameter (0.01, 0.1, 1, 10), and the use of probability estimates (True).

By fitting the GridSearchCV object to the training data, the code performs the hyperparameter search using cross-validation. The result is the identification of the best combination of hyperparameters that yield the highest performance, as determined by accuracy. This best-performing SVM model can then be used for sentiment analysis on new, unseen data.

```python
from sklearn.model_selection import GridSearchCV

model = SVC()

# Define the hyperparameters to search over
param_grid = {'kernel': ['linear', 'poly', 'rbf', 'sigmoid'],
              'C': [0.1, 1, 10, 100],
              'gamma': [0.01, 0.1, 1, 10],
              'probability': [True]}

# Define the GridSearchCV object with 5-fold cross-validation
grid_search = GridSearchCV(model, param_grid, cv=5, n_jobs=-1)

# Fit the GridSearchCV object to the training data
grid_search.fit(X_train_celeb, y_train_celeb)
```

Figure 4.17 Python code for grid search to train SVM model

The confusion matrix, accuracy, recall, precision and F1-sscore are calculated as shown in Figure 4.18.

```python
from sklearn.metrics import confusion_matrix
# Transform the test data using the same vectorizer used for the training data
X_test_celeb = vectorizer.transform(X_test_celeb)

# Make predictions using the best model obtained from the grid search
best_model = grid_search.best_estimator_
y_pred_celeb = best_model.predict(X_test_celeb)

# Compute the precision, recall, f1-score, and support for positive and negative sentiment
from sklearn.metrics import classification_report

report = classification_report(y_test_celeb, y_pred_celeb)
print(report)

# Compute the confusion matrix
cm = confusion_matrix(y_test_celeb, y_pred_celeb)

# Print the classification report with the confusion matrix
report = classification_report(y_test_celeb, y_pred_celeb)

print("Confusion matrix:")
print("              Predicted")
print("            |  0  |  1  |")
print(f"True    0   | {cm[0][0]}  |  {cm[0][1]}  |")
print(f"        1   | {cm[1][0]}  |  {cm[1][1]}  |")
```

Figure 4.18 Python code to predict the test data

4.10 Descriptive analysis

The output of sentiment from both VADER and SVM are collected and visualized to aid in finding trend in the dataset. Couple of visualization are constructed as initial output. However, only summarized and selected output will be discussed in Chapter 5 to avoid clutter and promotes comparative view.

Engagement rate is numbers that varies across all rows of tweets. To present this data appropriately, box plot will be used to fit all the numbers into the visual without cluttering. The tweet is filtered based on the column “status” and “sentiment_label” and using formula discussed in Section 3.6.5.1, the engagement rate is generated. By using module from matplotlib, the engagement rate can be visualized as a box plot using code as shown in Figure 4.19. This procedure will be applied for each category and each selected personal brand.

```python
import matplotlib.pyplot as plt

# Filter the result_pol DataFrame
filtered_df = result_pol[(result_pol['status'] == 'Main post') & (result_pol['sentiment_label'].isin(['Negative', 'Positive']))]

# Merge the filtered DataFrame with the followers DataFrame to get the number of followers for each user
merged_df = filtered_df.merge(df, on='Username')

# Calculate the engagement rate
merged_df['engagement_rate'] = (merged_df['Number of Likes'] + merged_df['Number of Retweet']
                                + merged_df['Number of Replies']) / df['followers_count']

# Create a boxplot to compare the distribution of engagement rate between negative and positive posts
merged_df.boxplot('engagement_rate', by='sentiment_label', showfliers=False)
plt.xlabel('Sentiment')
plt.ylabel('Engagement Rate')
plt.title('Distribution of Engagement Rate by Sentiment')
plt.show()

```

Figure 4.19 Python code to display box plot for engagement rate

To understand the rough content and trend of each category or personal brand, wordcloud was used to present all the word in from each tweet. Filter was first applied to select only preferred criteria of the tweet and the text are prepared for a wordcloud construction as shown in Figure 4.20.

```python
import pandas as pd
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Filter main_posts_df by username and date range
username = 'XavierNaxa'
start_date = '2021-01-01'
end_date = '2021-12-30'
df_filtered = main_posts_df[(main_posts_df['Username'] == username) &
                            (main_posts_df['Date Created'] >= start_date) &
                            (main_posts_df['Date Created'] <= end_date)]



# Combine all text into a single string
text = ' '.join(df_filtered['translated cleaned'])

# Generate word cloud
wordcloud = WordCloud(width=800, height=800, background_color='white').generate(text)

# Plot word cloud
plt.figure(figsize=(8,8))
plt.imshow(wordcloud)
plt.axis('off')
plt.show()

```

Figure 4.20 Python code to display wordcloud

Engagement rate and polarity of sentiment is two important parameter that personal brand need to look out for. To aid in analysing these parameters, number of engagement rate will be categorized as well as the aggregated sentiment from each tweet to simplify the data. The formula for the categorization is discussed in Section 3.6.5.2 and shown in Figure 4.21 and Figure 4.22.

```python
# Setting the mean and standard deviation of the Engagement Rate
mean_count = syedsaddiqpost["Engagement Rate"].mean()
std_count = syedsaddiqpost["Engagement Rate"].std()

# Define the bin edges for the categories
bin_edges = [0, mean_count + std_count, float('inf')]

# Create labels for the categories
labels = ['Low Engagement Rate', 'High Engagement Rate']

# Use pd.cut to categorize the engagement rate based on the bin edges and labels
category = pd.cut(syedsaddiqpost["Engagement Rate"], bins=bin_edges, labels=labels)

# Assign the "category" column to syedsaddiqpost
syedsaddiqpost["Engagement Rate Category"] = category

```

Figure 4.21 Python code to classify tweet based on engagement rate

```python
#Setting the mean and standard deviation of the sentiment counts
mean_count = sentiment_counts.mean()
std_count = sentiment_counts.std()

# Define the bin edges for the categories
bin_edges = [-float('inf'), -mean_count-std_count,0, mean_count+std_count, float('inf')]

# Create labels for the categories
labels = ['High Negative','Low Negative','Low Positive', 'High Positive']

# Use pd.cut to categorize the positive values based on the bin edges and labels
category = pd.cut(sentiment_counts, bins=bin_edges, labels=labels)

# Map the sentiment category to the syedsaddiqpost DataFrame using Conversation ID
syedsaddiqpost['Sentiment Aggregation Category'] = syedsaddiqpost['Conversation ID'].map(category)
```

Figure 4.22 Python code to classify tweet based on sentiment aggregation

This category is also act as a filter for a wordcloud to find the trend in the tweet posted by selected personal brand that will be discussed in Section 5.3.3. The text found in the wordcloud will be used as observation to determine the nature of each tweet and label them accordingly. To help visualize the content for each category, the category and label for each category are tabulated to have global view of the label of each tweet and the category of tweet as the effect.

4.11 Chapter Summary

This chapter has presented an in-depth procedure and outcome of each process. Each process is discussed from data collection, data pre-processing, data exploratory analysis, splitting, feature extraction, SMOTE, performance evaluation and analysis are discussed in term of output, input and the processes. The result from all this process will be discuss in Chapter 5.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Introduction

In this chapter, the result and discussion of the analysis performed in Chapter 4 will be discussed. The performance of the machine learning and lexicon will be analysed. The result of machine learning is presented in a summarized classification report and the result will be discussed. Next, the output of sentiment based on VADER sentiment classifier will be observed to determine the perception of the audience to the personal brand listed.

5.2 Data Validation

The summary of result obtained from the SVM machine learning classification are presented in Table 5.1. The table is shown to compare the performance of SVM classification across different category of data. The accuracy of machine learning shows relatively good result which range from 82% to 96% having Merged data being the most accurate at while Technology category having the weakest performance. The precision result ranges from 83% to 95% and 82% to 96% for positive and negative classification respectively. Recall result on the other hand is not good news, Celebrity category are having value of recall of 55% while negative is 93%. Other than that, the value ranges from 93% to 85% and 97% to 55% for positive and negative respectively. As the result, the F1-score for Celebrity in negative classification having the lowest value which is at 62% while the positive classification at 90.

Table 5.1 Summary of SVM sentiment analysis report

		Merged	Tech	Politics	Entrepreneur	Celebrity	Entertainer
Accuracy		0.96	0.82	0.94	0.93	0.85	0.93
Precision	**Pos**	0.95	0.83	0.93	0.90	0.87	0.89
	Neg	0.96	0.85	0.94	0.95	0.81	0.94
Recall	**Pos**	0.93	0.85	0.91	0.85	0.93	0.89
	Neg	0.97	0.83	0.95	0.97	0.55	0.94
F1 Score	**Pos**	0.94	0.84	0.92	0.88	0.9	0.89
	Neg	0.96	0.84	0.95	0.96	0.62	0.94

5.2.1 Source of Error

There is multiple possible reason or factor that is affecting a low value of recall for the celebrity. Low of recall in this case would be because the SVM model failed to classify a significant number of true positive samples correctly in the data which in this case, Celebrity category. To address this issue, further investigation is necessary to identify the possible factors that could have contributed to the low recall score. These factors can be broadly categorized into three groups: data-related factors, model-related factors, and lexicon related factors. The following section will discuss each of these factors in more detail to identify potential sources of error and uncertainty in the classification process.

5.2.1.1 Data Related Factor

One of the most possible reason for this case is due to data imbalance. This led to a biased model where the model is more susceptible to predict an object as a majority

class which in this case a positive category. Figure 4.13 shows the data distribution in all data category.

The imbalance of data in Celebrity category and Entrepreneur can lead to bias. However, only Celebrity category is where the recall value is very low on positive category. Theoretically, introducing SMOTE will likely reduce bias in the model. Table 5.2 is the result when SMOTE is being used compare to bare data.

Table 5.2 Comparison of report using SMOTE and not using SMOTE

		Bare Data	SMOTE
Accuracy		0.85	0.85
Precision	**Pos**	0.87	0.88
	Neg	0.81	0.72
Recall	**Pos**	0.93	0.93
	Neg	0.55	0.57
F1 Score	**Pos**	0.9	0.91
	Neg	0.62	0.64

Insufficient training data could also be one of the factors that are affecting the result. Small training dataset led to immature model where it may not have learned enough patterns or variation in the data where it failed to generalize the data which lead to low recall value. Noisy data on the other hand will introduce confusion to the model. Which could make it harder for the model to accurately identify the true positives leading to lower recall value.

Another different factor related to data set is the frequency of keyword in different sentiment polarity. This should focus on the tweet that are negative but misclassified as positive by the SVM model. By examining the misclassified tweets, the analysis of words within them and their frequency of occurrence in both positive and negative labelled tweets can be conducted. This analysis helps determine whether the words in the misclassified tweet are more closely associated with the positive or negative label. By identifying the frequency and association patterns of these words,

valuable insights into the factors contributing to the low recall value in the negative category can be gained.

The occurrence of each word in positive and negative tweet from celebrity training dataset are calculated as shown in Table 5.3. Temporary label is created according to the frequency of occurrence in positive and negative tweet. If the count of occurrence in positive is more on positive, the word will be labelled positive and vice versa. If the count of occurrence is equal, the word will be labelled neutral.

Table 5.3 Word labelling based on occurrence in positive and negative sentiment

Word	Count in Positive	Count in Negative	Label
older	4	0	Positive
country	6	4	Positive
melted	0	1	Negative
takes	2	0	Positive

Now, the label in each word in the misclassified tweet will be calculated and the percentage will be calculated based on how many words of the tweet is labelled as positive, negative and neutral as shown in Table 5.4. These steps help us compare and understand the training dataset and its effect on the testing.

Table 5.4 Percentage Distribution of Sentiment Labels in Text Sentences

Text	Percentage Positive, %	Percentage Negative, %	Percentage Neutral, %
filling independence miserable	33.33	0	66.67
aint funny doc real talk shizzy	50	16.67	33.33
read profile bro reason tired	60	20	20
decorate finding reasons hate rid number heart	57.14	14.29	28.57

The analysis of the misclassified tweet reveals an interesting pattern. It is observed that a majority of the words in these misclassified tweets are more strongly associated with positive sentiment rather than negative sentiment. Based on the analysis conducted, it was discovered that out of the 31 sentences examined, 27 of them had a higher percentage of words associated with the positive dataset. The remaining dataset had an equal percentage of positive and negative words, and interestingly, none of its words were present in the training dataset. This suggests that the majority of the words in these sentences were labelled as positive based on the training dataset. This shows that despite the text is labelled negative by VADER, if the words within the text are commonly associated with positive sentiment in the training dataset, the SVM model might incorrectly predict it as positive. This finding shed light on why the SVM model fails to accurately predict these rows as negative despite their labeled sentiment.

An important lesson learned from this issue is the significance of having a sufficient amount of data. To avoid similar challenges in the future, it is crucial to include a larger and more diverse set of samples representing negative sentiment. By incorporating a broader range of negative sentiment instances, the SVM model can better capture and understand a wider variety of negative sentiment words. This expansion of the training data can significantly enhance the model's ability to accurately classify negative texts and improve overall performance.

5.2.1.2 Lexicon related factors

The polarity of text obtained using VADER Lexicon is acting as a baseline for our data instead of manually labelling the text as positive or negative. It has proven by many studies that the VADER Lexicon is very good at handling Twitter text data where it can incorporate rules or mechanism to handle negations and sentiment modifiers for example; “very good” or “extremely good”.

Figure 5.1 Sample tweet

When VADER analyzes a text, it assigns sentiment scores to individual words based on their presence in the lexicon. Each word is assigned a sentiment polarity score ranging from -1 to +1. The sentiment score reflects the word's positivity or negativity, with higher scores indicating more positive sentiment and lower scores indicating more negative sentiment. VADER also takes into account the context of the words by considering the valence shifters, intensifiers, and negations that appear in proximity to the words. These modifiers can influence the sentiment scores assigned to the words.

The sentiment score for the entire text is calculated by aggregating the individual word scores while considering the context and word order. The resulting sentiment score represents the overall polarity of the text, ranging from highly negative (-1) to highly positive (+1), with scores around 0 indicating neutrality. However, there is some situation where the VADER has struggled to accurately determine a text's polarity.

In the case of the phrase "trust issue" as shown in Figure 5.1, VADER might not identify it as negative polarity because it primarily relies on the sentiment scores of individual words. "trust" is generally considered positive, while "issue" can be context-dependent and not inherently positive or negative. VADER use combination of word-level and sentence-level grammatical rules to determine the sentiment of a text. However, the combination of the word "trust issue" might be missing from the VADER pre-trained lexicon library. Table 5.5 shows the sentiment of the word "trust",

“issue” and “trust issue” to show how VADER handle these words. Combining all three words, including the word "problem," indicates that VADER primarily utilizes word-level analysis. However, it is important to note that VADER does possess the capability to identify the context of words. For example, the word "delicious" is generally considered positive. By adding the intensifier "very" before the word "delicious," it further enhances the positive value of the compound polarity, indicating a stronger positive sentiment. This showcases VADER's ability to take into account contextual information and adjust the polarity accordingly.

Table 5.5 Compound polarity of sample word

Word	Compound polarity
Trust	0.5106
Issue	0.0
Problem	-0.4019
Trust issue	0.5106
Trust issue problem	0.1531
Delicious	0.5719
Very delicious	0.6115

In conclusion, while VADER is effective in many cases, it may not capture more complex sentiments that require a deeper understanding of the text's overall context and structure. In such cases, more advanced sentiment analysis techniques or custom models trained on specific domains may be necessary to achieve more accurate results.

5.3 Result Discussion

This section will discuss the finding obtained based on the trend of sentiment on the dataset. The findings are discussed in perspective of timeline of the tweet, the category of the tweet and also the personal brand that post the tweet. All the sentiment

discussed in this section is based on the results obtained from VADER, unless otherwise stated.

5.3.1 Category Analysis

Main objective of this section is to compare the sentiment of the tweet based on the category. Based on the whole dataset, the number of positive tweets is 48609 while the number of negative tweets is 27978. Figure 5.2 shows the percentage of the replies based on the main tweet. The left pie chart represents the distribution of its replies based on the sentiment of the main tweet. The distribution shows that positive main tweets receive a higher proportion of positive replies compared to negative replies, with 70.3% being positive and 29.7% being negative. On the other hand, negative main tweets receive 50.1% negative replies and 49.9% positive replies. To simplify the discussion for specific category, the result will be presented in Table 5.6 and Appendix C.

.

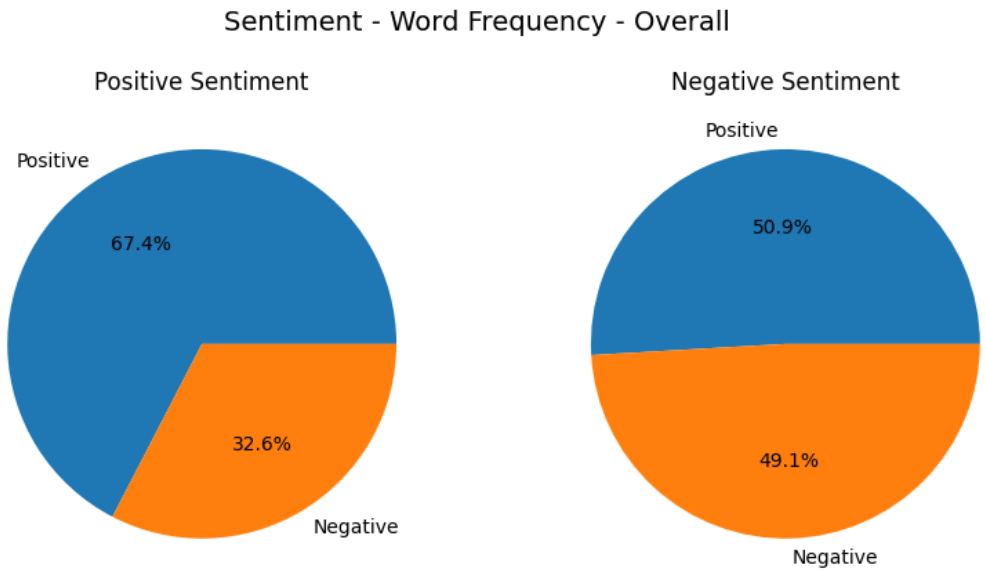


Figure 5.2 Overall sentiment distribution

Based on Table 5.6, the result shows that the percentage of negative tweet and replies on positive main tweet in celebrity category is very low at 25% followed by

entrepreneur at 39.4%. Sentiment of replies on negative main tweet shows the highest negative replies on politic category at 35.9% and 49.2% for positive and negative replies respectively. It is important to note the differences in replies between positive and negative tweet. In Entrepreneur category, the difference in proportion of positive replies on positive tweet compare to negative tweet is about 20% which is the highest, which also can be seen for its negative tweet where the differences are 19%.

As a summary, the distribution of sentiment in Celebrity category shows that positive main tweets are generally associated with positive replies. While Celebrity shows that most obvious portion of positive replies, this is also true for other category where user tend to respond to the tweet positively. High percentage of positive replies on positive main tweet in Entrepreneur category indicating a generally positive sentiment associated with entrepreneurial content. However, the percentage of positive replies to negative main tweets in this category drops by couple percent, suggesting that negative content in this category might receive a slightly lower positive response compared to positive content. Politic category shows higher percentage of negative main tweets compare to another category. Additionally, the percentage of positive replies to negative main tweets is relatively lower indicating a less positive response to negative political content.

In general, it can be observed that the audience reacts to celebrities quite positively. This positive response may be attributed to the fact that celebrities often share content that promotes positivity. Such content may include their achievements, or inspirational messages. By sharing these types of content, celebrities are able to cultivate a positive image and engage with their audience in a way that resonates with them. Positive main tweets from celebrities tend to resonate with their followers, resulting in a positive response. On the other hand, Politic category stands out with a higher percentage of negative main tweets compared to other categories. Politics often involves controversial or divisive topics, which can generate negative sentiment. Political figures may use negative messaging to criticize the government, address challenges, or highlight issues, contributing to the prevalence of negative main tweets in this category and in return might create tense negative emotion to the audience.

Table 5.6 Summary of sentiment distribution

Category	Main Tweet	Replies	Percentage (%)
Technology	Positive	Positive	68.3
		Negative	31.7
	Negative	Positive	59.5
		Negative	40.5
Entertainer	Positive	Positive	70.7
		Negative	29.3
	Negative	Positive	55.9
		Negative	44.1
Entrepreneur	Positive	Positive	79.6
		Negative	20.4
	Negative	Positive	60.6.
		Negative	39.4
Celebrity	Positive	Positive	82.3
		Negative	17.7
	Negative	Positive	75.0
		Negative	25.0
Politic	Positive	Positive	64.1
		Negative	35.9
	Negative	Positive	50.8
		Negative	49.2

Engagement rate is a crucial metric for assessing the effectiveness and impact of social media posts. In this analysis, the distribution of engagement rates between negative and positive sentiment tweets across various categories are compared which includes Technology, Entertainer, Celebrity, Politic, and Entrepreneur. By examining the mean engagement rates, standard deviations, and other statistics, the aim is to gain insights into the audience's response to different sentiments within each category. Table 5.7 provides a summary of the mean, median, standard deviation, maximum, and minimum values of the engagement rate across different categories.

Table 5.7 Summary of Engagement rate across category

Category	Sentiment	Mean Engagement Rate	Median Engagement Rate	Standard Deviation	Maximum Engagement Rate	Minimum Engagement Rate
Technology	Negative	0.0264	0.0295	0.0214	0.0473	0.0010
	Positive	0.3870	0.0126	1.3259	7.5172	0.0002
Entertainer	Negative	0.3218	0.0303	0.5240	1.7561	0.0033
	Positive	0.3650	0.1285	0.9664	4.9806	0.0046
Celebrity	Negative	0.4498	0.4498	0.6266	0.8929	0.0068
	Positive	0.0352	0.0248	0.0421	0.1312	0.0020
Politic	Negative	0.3332	0.0444	1.0340	4.3214	0.0004
	Positive	0.0821	0.0165	0.1379	0.4892	0.0004
Entrepreneur	Negative	0.4134	0.1031	0.6162	1.9099	0.0065
	Positive	0.4597	0.0280	1.1097	4.2877	0.0026

The trend of engagement rate with respect to sentiment of the main tweet shows mixed outcome. When compared to tweets with negative sentiment, positive sentiment posts in the Technology category have a significantly higher mean engagement rate where positive post. The mean engagement rate for posts with positive sentiment is 0.3870 which is 93.8% higher compare to mean engagement rate for tweet with negative sentiment is 0.0264. Positive sentiment tweets have a much higher standard deviation at 1.3259 than tweets with negative sentiment at 0.0214, which suggests a wider range of engagement rates. As a result, it seems likely that tweets with a positive tone in the Technology category will result in a higher volume of interactions overall, but with a greater degree of variability. The high value of standard deviation suggests that while some tweet received high engagement, others may receive relatively low engagement. While positive tweets in the Technology category have the potential to generate a significant amount of engagement, but the results can vary widely. It could be due to various factors such as the content of the tweet, the timing of the tweet, or the specific interests of the audience within the Technology category.

In the Entertainer category, tweets with both positive and negative sentiment have relatively the similar mean engagement rate. The average engagement rate for tweets with negative sentiment is 0.3218, whereas the average engagement rate for tweets with positive sentiment is 0.3650. However, different from mean engagement rate, the difference between positive tweet and negative tweet median engagement rate

is 323.76% where the median engagement rate for positive and negative is 0.1285 and 0.0303 respectively. This suggests that there is a noticeable difference between the central engagement rates for tweets with different sentiment. The positive sentiment tweets have a relatively higher median, indicating that a significant proportion of those tweets received higher engagement rates compared to the negative sentiment tweets. Positive sentiment tweets have slightly higher standard deviation which is at 0.9664, than negative sentiment tweets at 0.5240, indicating less consistency in engagement rates for positive tweet.

In the Celebrity category, the mean engagement rate for negative sentiment tweets is significantly higher compared to positive sentiment tweets. Negative sentiment tweets have a mean engagement rate of 0.4498, while positive sentiment tweets have a much lower mean engagement rate of 0.0352. The median engagement rate is also much higher in negative tweet compare to positive tweet at 0.4498 and 0.0248 for negative and positive respectively. In term of the standard deviation, it is higher for negative sentiment tweets at 0.6266, compared to positive sentiment tweets which is at 0.0421 in the Celebrity category. The higher mean, median, and standard deviation for negative sentiment tweets in the Celebrity category indicate that content with a negative tone tends to generate more overall engagement. This could be due to various factors such as controversy, provocative content, or the tendency of audiences to respond more strongly to negative emotions or gossip in the context of celebrity-related content.

Almost similar trend can be observed in Politic category, tweets with negative sentiment have a higher mean engagement rate than tweets with positive sentiment. Particularly, tweets with negative sentiment have an engagement rate of 0.3332, while tweets with positive sentiment have an engagement rate of 0.0821 which is 305% less. The median engagement rate for negative sentiment tweets is not much higher than that of positive sentiment tweets at 0.0444 and 0.0165 for negative and positive tweet respectively. This shows that negative tweet contains some extreme values that are significantly higher than the majority of the data points. Positive sentiment tweets have a lower standard deviation at 0.1379 and 1.0340 for positive and negative sentiment tweet respectively, indicating less consistency in engagement rates for negative tweets.

In the Politic category, negative sentiment tweets tend to have higher mean engagement rates and exhibit more variability in engagement compared to positive sentiment tweets. This suggests that the audience in the Politic category may be more responsive to tweets with a negative sentiment as it probably put more tempt for the audience to participate in the political debate and share their opinion.

In the Entrepreneur category, the mean engagement rate for both positive and negative sentiment tweets is not significantly different at 0.4134 and 0.4597 for negative and positive tweet respectively. The median engagement rate for negative sentiment tweets in the Entrepreneur category is approximately 5 times higher at 0.1031 compared to positive sentiment tweets at 0.0280. This indicates that a significant proportion of negative sentiment tweets receive relatively higher engagement rates compared to positive sentiment tweets. While the mean may not show a distinct difference, the higher median suggests that there is a subset of negative sentiment tweets that perform exceptionally well in terms of engagement. Furthermore, the standard deviation for positive sentiment tweets is approximately 2 times higher compared to negative sentiment tweets. This indicates that there is greater variability in the engagement rates of positive sentiment tweets in the Entrepreneur category. Some positive tweets may receive significantly higher engagement, while others may receive relatively lower engagement. On the other hand, the engagement rates of negative sentiment tweets in this category show relatively less variation. In short, negative sentiment tweet tend to attract more engagement as shown by the median engagement rate while some positive tweet also receives high engagement while not as frequent as negative sentiment tweet.

In this section, only the sentiment of the tweet and how it effects the engagement rate of the tweet will be considered. In summary, positive sentiment in Technology category tend to boost the engagement rate as positive tweets have a significantly higher mean engagement rate compared to negative tweets. On the other hand, Celebrity and Politic category seems to be benefit more on negative sentiment tweet compare to positive tweet in term of engagement rate. The engagement rate for Entertainer and Entrepreneur category on the other hand does not rely solely on the sentiment of the tweet itself. This could probably because the Entertainer and

Entrepreneur categories are influenced by factors beyond just the sentiment of the tweet.

5.3.2 Personal Brand Analysis

In this section, a personal brand analysis will be conducted by selecting one username from each category who has a substantial number of tweets. The aim is to examine how certain parameters, previously discussed in Section 5.3.1, correlate with specific personal brands. By doing so, insights into how these parameters influence public perception can be obtained and the effective strategies for personal brands to manage them in order to leverage their impact positively can be explored.

Correlating the parameters with individual personal brands allows for a deeper understanding of how these brands are perceived by the public. This analysis helps identify key factors that contribute to their success. This analysis will provide insights into the strategies employed by personal brands to cultivate a favourable image and establish meaningful connections with their audience.

This examination aims to emphasize the significance of personal branding and provide insights into effective strategies for personal brands to navigate the ever-changing landscape of social media. By shedding light on these aspects, the importance of cultivating a strong personal brand can be studied as well as offer valuable guidance for individuals seeking to establish their presence on social media platforms. By understanding the impact of various parameters on public perception and engagement, individuals can refine their personal branding strategies and effectively communicate their desired image to their target audience.

5.3.2.1 Syed Saddiq

For Politics category, the personal brand that are chosen is the username “SyedSaddiq”. Syed Saddiq is a Malaysian politician and youth activist. He has held several significant roles in his political career, demonstrating his commitment to youth

empowerment, education, and sports development. Throughout his career, Syed Saddiq has been vocal about the importance of youth empowerment, education reform, and creating a more inclusive society. He has been recognized for his efforts in advocating for youth rights and political engagement. Syed Saddiq's roles reflect his dedication to uplifting the voices and concerns of the youth population in Malaysia, making him a prominent figure in the country's political landscape.

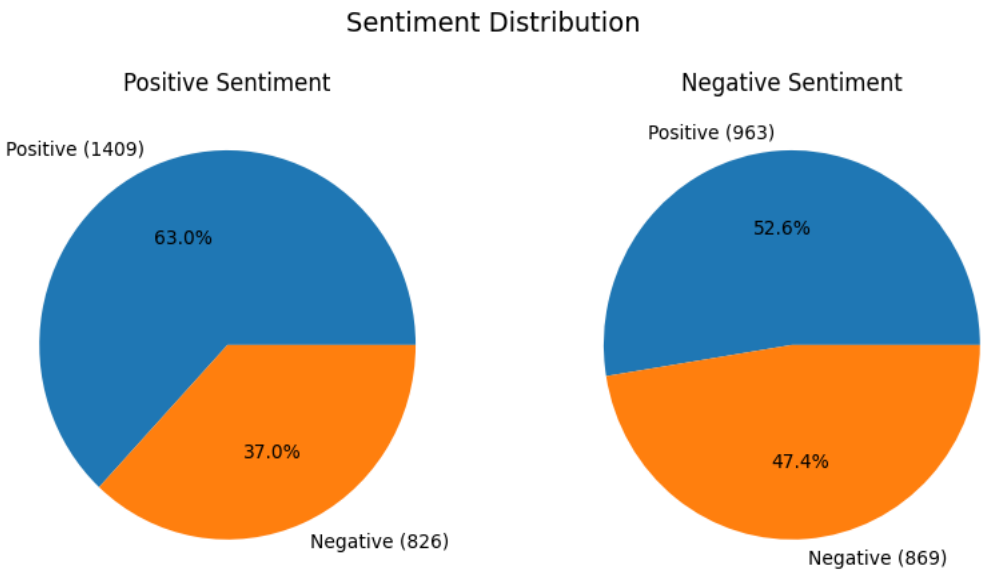


Figure 5.3 Sentiment distribution in replies (Syed Saddiq)

Figure 5.3 show the distribution of reply sentiment based on the sentiment of the main tweet. Based on Figure 5.3, for positive main tweets, approximately 63% of the replies were positive, accounting for 1409 replies, while around 37% of the replies were negative, totalling 826 replies. This indicates a higher proportion of positive sentiment in response to positive main tweets.

On the other hand, for negative main tweets, approximately 52.6% of the replies were positive, amounting to 963 replies, while about 47.4% of the replies were negative, totalling 869 replies. This suggests a relatively balanced sentiment distribution in response to negative main tweets, with a slightly higher proportion of positive sentiment compared to negative sentiment.

These findings indicate that positive main tweets tend to receive a higher proportion of positive replies, while negative main tweets elicit a relatively balanced mix of positive and negative replies. This could imply that positive main tweets have a more positive impact and generate a higher level of engagement with positive sentiments. The presence of positive replies to negative main tweets might suggest efforts by some users to counteract or provide alternative viewpoints.

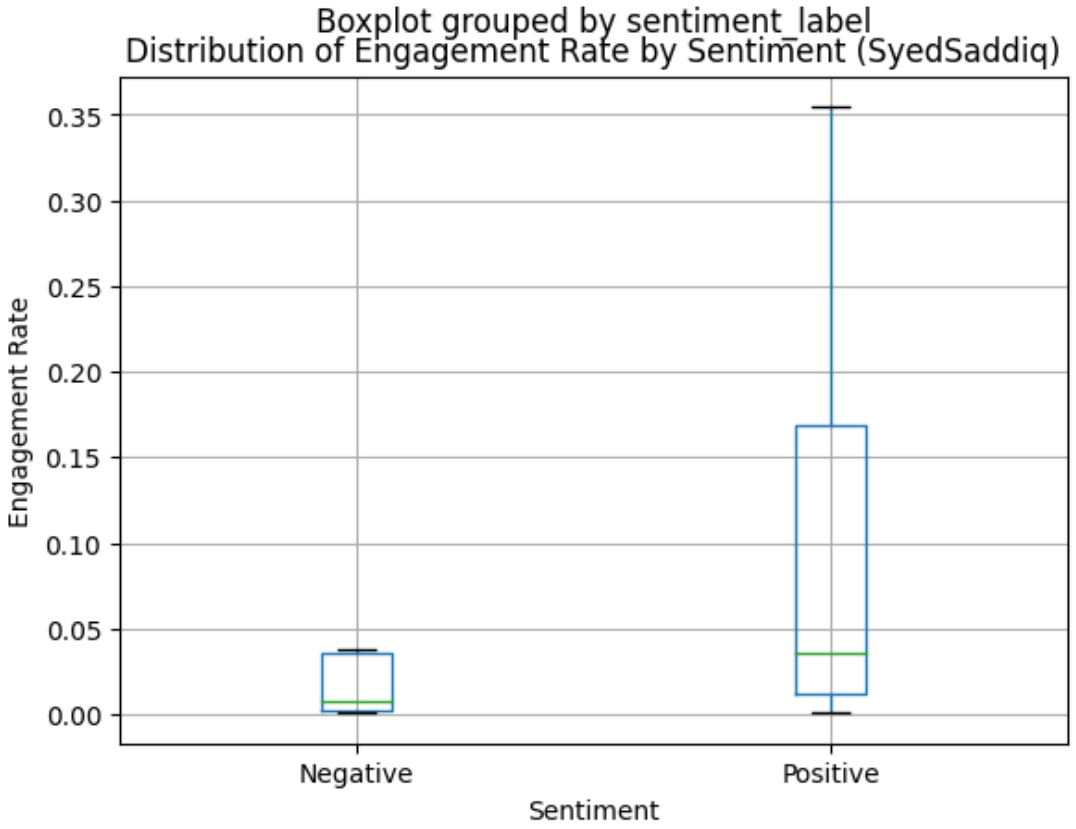


Figure 5.4 Box plot of engagement rate (Syed Saddiq)

Table 5.8 Summary of box plot of engagement rate (Syed Saddiq)

Sentiment	count	Mean	S.D	Min	25%	50%	75%	Max
Negative	16	0.092	0.216	0.001	0.003	0.007	0.036	0.828
Positive	25	0.137	0.222	0.001	0.012	0.036	0.169	0.951

Generally, the engagement rate for Syed Saddiq's tweet show relatively bigger spread for positive tweet as the range from 25% to 75% shows that the spread of engagement rate is bigger for positive tweet compare to negative tweet. The mean engagement rate for positive tweet appears to be 48% higher than negative tweet at 0.137 and 0.092 for positive and negative tweet relatively. The 75^{th} percentile comparison for positive and negative tweets shows that positive tweet received a higher engagement rate. Referring to Table 5.8, the maximum engagement rate for both positive and negative tweet is relatively the similar with very high engagement rate. As this particular case is an outlier, it cannot be statistically discussed in this context. However, for further exploration and analysis, a closer examination of each tweet is necessary. This will be addressed in Section 5.3.3, a detailed discussion of the individual tweets will be studied.

As a prominent figure in national politics, every statement made by Syed Saddiq carries significant weight and can be subject to varying interpretations. In the age of social media, where public perception is greatly influenced by online interactions, it becomes crucial for politicians to be mindful of the potential negative perceptions that can arise from their statements. Studies has found that social media analysis can be one of the methods to predict the result of an election which shows how important an online perception for a politician. With 25 positive tweets and 16 negative tweets, the analysis demonstrates the impact of positive messaging on Syed Saddiq's perceived image. Therefore, it is imperative for him to prioritize using positive language and communication strategies to shape a favourable public perception.

Generally, the pervasive nature of social media amplifies the impact of these statements, making it essential for politicians to employ strategic communication

strategies to cultivate a positive and favourable image. Recognizing the importance of social media as a powerful tool for shaping public perception, politicians must exercise caution and thoughtfulness in their messaging to ensure that their statements are well-received and contribute to building a positive image.

5.3.2.2 Khairul Aming

Khairul Aming is a well-known social media influencer who has made a name for himself in the online world. He gained popularity through his successful business venture selling "sambal nyet" and recently achieved remarkable success. According to a report by Azma ML in 2023, his business recorded an astonishing 13,000 sales in just 2 minutes, marking a significant milestone after three years of hard work.

Khairul Aming initially rose to fame through his YouTube channel, where he shares cooking tutorials and culinary insights. His videos have garnered a substantial following and helped him establish his presence in the online community. Recently, one of his videos went viral, showcasing his heart-warming gesture of appreciation towards his staff. In the video, he took his staff shopping and generously rewarded them with incentives, breaking the stereotype that young people are only interested in self-indulgence.

This viral video not only highlighted Khairul Aming's business success but also showcased his admirable character and thoughtfulness towards his team. His actions have inspired many, challenging the perception that young individuals are solely focused on personal enjoyment. Through his entrepreneurial journey and genuine gestures, Khairul Aming has become a role model for aspiring entrepreneurs and a source of inspiration for the younger generation.

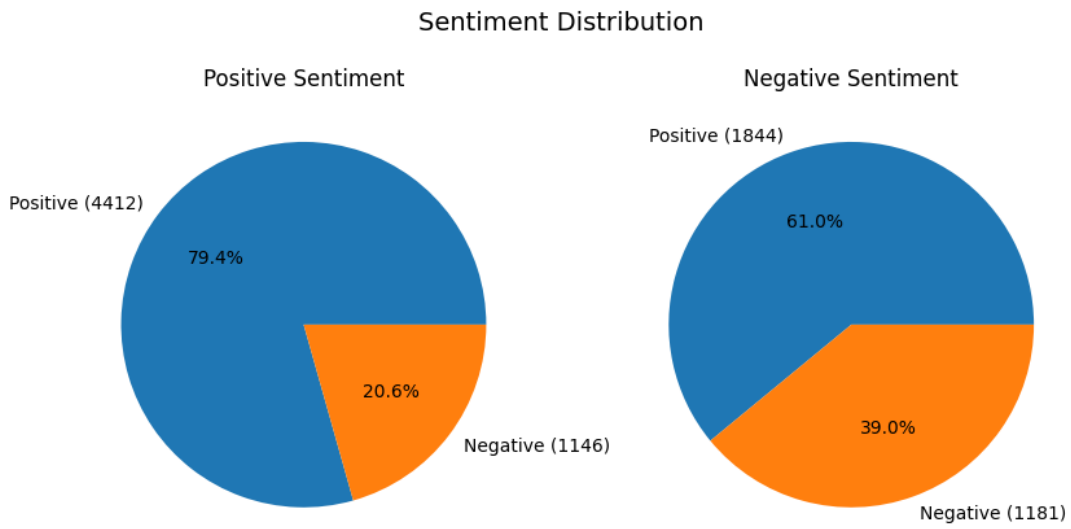


Figure 5.5 Sentiment distribution in replies (Khairul Aming)

Based on Figure 5.5, the distribution of reply sentiment based on the sentiment of the main tweet is examined. For positive main tweets, approximately 79.4% of the replies were positive, totalling 4412 replies, while around 20.6% of the replies were negative, accounting for 1146 replies. This indicates a significantly higher proportion of positive sentiment in response to positive main tweets. Conversely, for negative main tweets, approximately 61.0% of the replies were positive, amounting to 1844 replies, while approximately 39.0% of the replies were negative, totalling 1181 replies. This suggests a relatively balanced sentiment distribution in response to negative main tweets, with a slightly higher proportion of positive sentiment compared to negative sentiment.

These findings suggest that positive main tweets tend to elicit a higher proportion of positive replies, indicating a positive impact and generating a greater level of engagement with positive sentiments. The presence of positive replies to negative main tweets might indicate efforts by some users to counteract or provide alternative viewpoints.

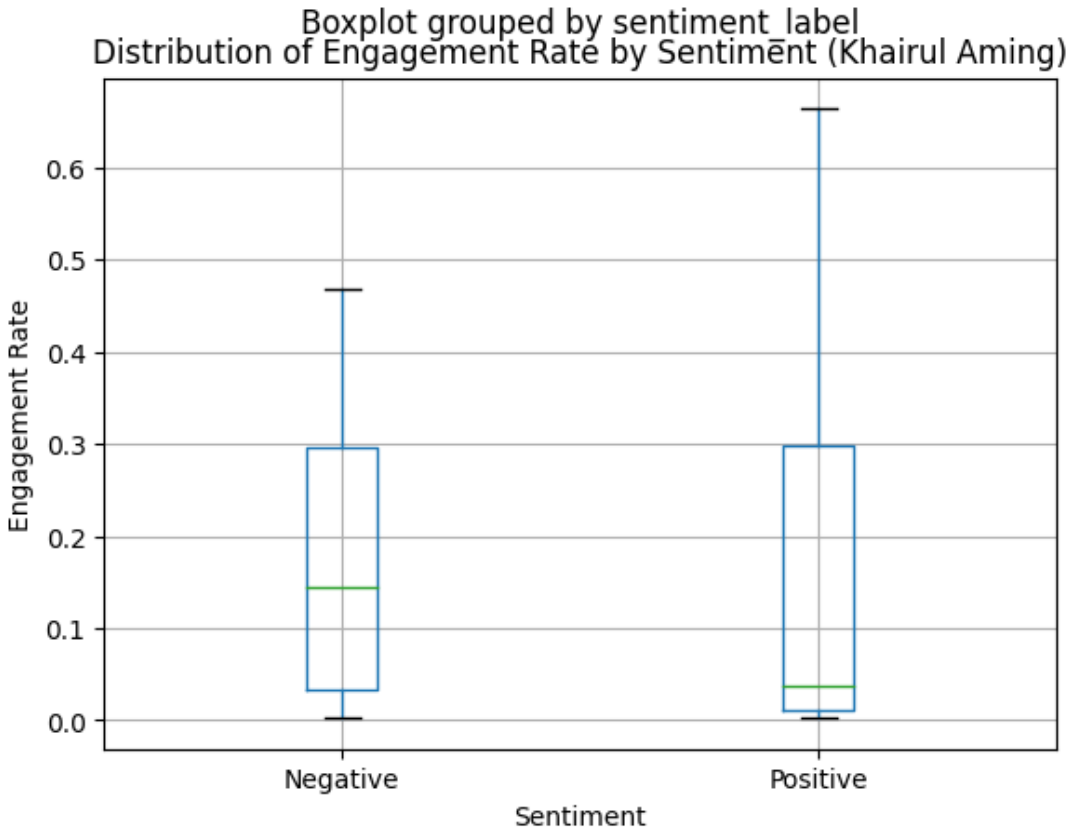


Figure 5.6 Box plot of engagement rate (Khairul Aming)

Table 5.9 Summary of box plot of engagement rate (Khairul Aming)

Sentiment	count	Mean	S.D	Min	25%	50%	75%	Max
Negative	12.0	0.328	0.588	0.003	0.034	0.144	0.296	2.126
Positive	29.0	0.306	0.524	0.002	0.010	0.037	0.298	1.874

The analysis of engagement rates for Khairul Aming's tweets provides interesting insights. A comparison between positive and negative tweets reveals that positive tweets exhibit a wider spread in engagement rates, as evident from the range between the 25th and 75th percentiles. The mean engagement rate for positive tweets is approximately 48% higher than that of negative tweets, with values of 0.306464 and 0.328262, respectively. Examining the 75th percentile, positive tweets receive a slightly higher engagement rate. Table 5.9 presents noteworthy data points: both positive and negative tweets display similar maximum engagement rates, which are remarkably high. However, these outliers warrant further investigation, and their analysis is reserved for Section 5.3.3.

Being a well-known social media influencer with a substantial following, has a significant impact on his audience. The higher proportion of positive sentiment in response to his positive main tweets indicates that his content resonates strongly with his followers, eliciting positive reactions and engagement. The positive sentiments expressed in the replies to his positive main tweets suggest that his audience appreciates his content and business endeavours, resulting in a positive association with his brand.

5.3.2.3 Xavier Naxa

Xavier Naxa is a prominent Twitter influencer known for his expertise in the field of tech news. He actively shares insightful tech advice and provides solutions to various tech-related problems.

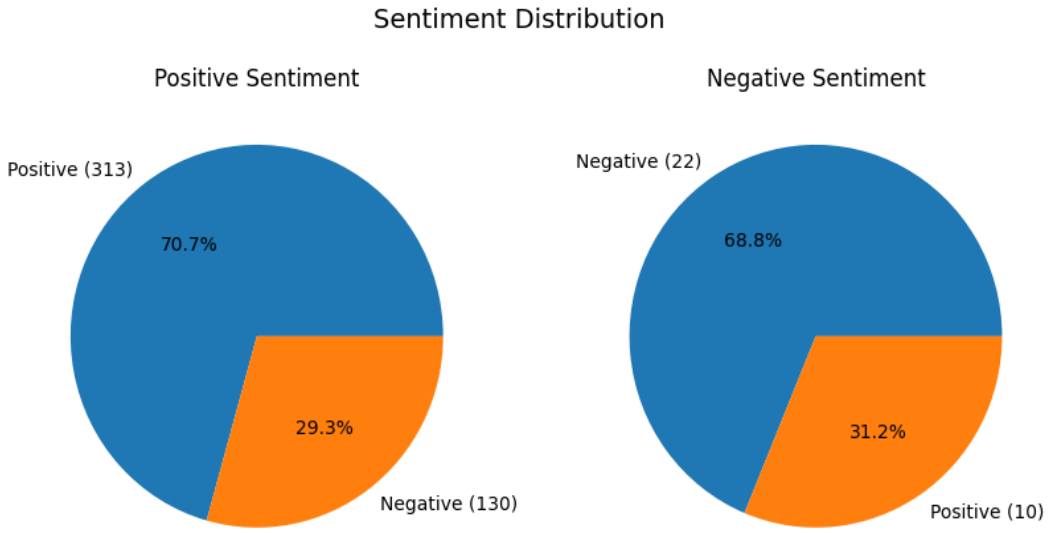


Figure 5.7 Sentiment distribution in replies (Xavier Naxa)

When analyzing the sentiment distribution of Xavier Naxa's tweets and their corresponding replies, intriguing patterns emerge. For positive main tweets, approximately 70.3% of the replies were positive, totalling 313 replies, while around

29.3% of the replies were negative, amounting to 130 replies. This suggests a strong positive sentiment in response to Xavier Naxa's positive main tweets.

Similarly, for negative main tweets, approximately 68.8% of the replies were positive, accounting for 22 replies, while approximately 31.2% of the replies were negative, totalling 10 replies. This indicates a relatively balanced sentiment distribution in response to Xavier Naxa's negative main tweets, with a slightly higher proportion of positive sentiment compared to negative sentiment.

These findings suggest that Xavier Naxa's positive main tweets tend to elicit a higher proportion of positive replies, reflecting his influence and reputation in providing valuable tech advice. However, having only one negative sentiment tweeted by Xavier Naxa made negative proportion of this analysis incomplete as it does not represent a trend for his Twitter post.

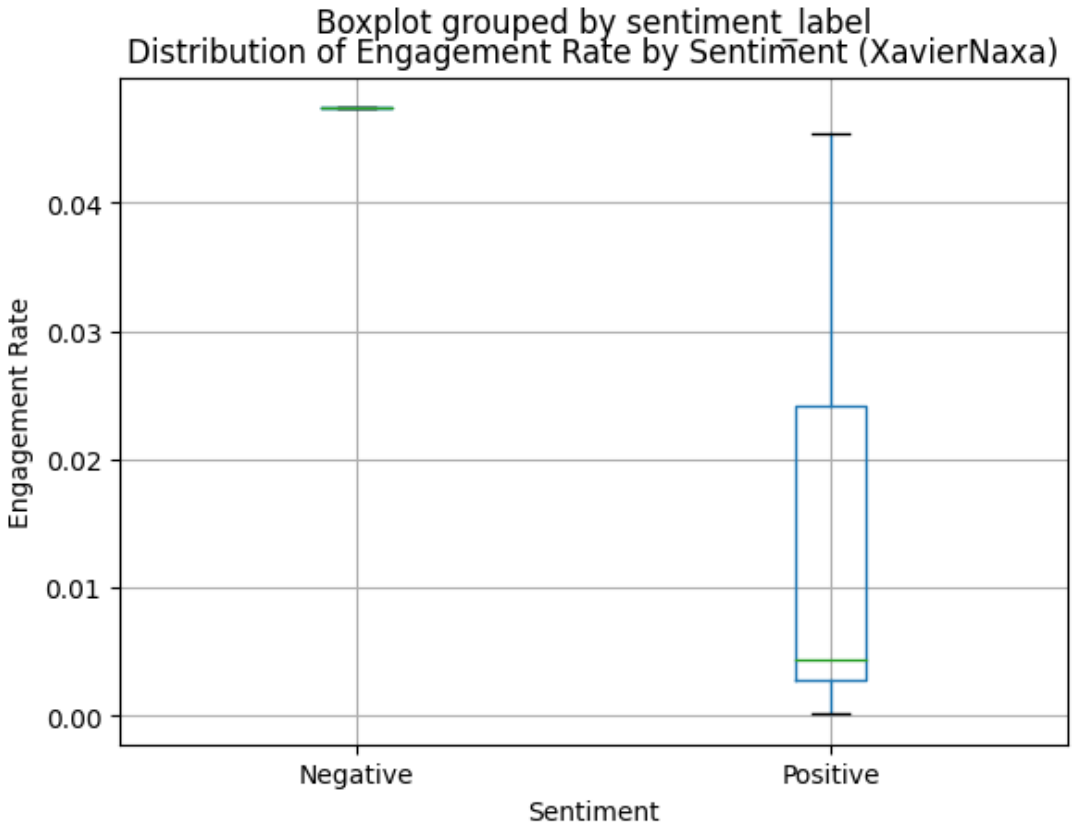


Figure 5.8 Box plot of engagement rate (Xavier Naxa)

Table 5.10 Summary of box plot of engagement rate (Xavier Naxa)

Sentiment	count	Mean	S.D	Min	25%	50%	75%	Max
Negative	1	0.047	NaN	0.0474	0.047	0.047	0.047	0.047
Positive	15	0.023	0.0384	0.0001	0.003	0.004	0.024	0.136

As mentioned, for the negative sentiment, there is only one occurrence, making it difficult to draw definitive conclusions. The mean sentiment score for this single instance is 0.047. However, with just one data point, it is challenging to determine the overall engagement or impact of negative tweets.

In contrast, there are 15 instances of positive sentiment. The mean sentiment score for these instances is 0.023 which is relatively high considering the 75% percentile is 0.024. The minimum engagement level is 0.0001 showing that not all of Xavier Naxa's tweet blew up and receive reaction. The maximum engagement is 0.136 indicates that, within the dataset, there was at least one highly engaging tweet that received a significant amount of positive feedback or responses. This particular tweet could have resonated strongly with the audience, prompting a higher level of engagement in the form of likes, comments, or retweets.

5.3.2.4 Hazeman Huzir

Hazeman Huzir is a prominent TV host and influencer in Malaysia. He gained recognition through his popular show called "Motif Viral," which was produced by the esteemed production team Thinker Studios. Although the "Motif Viral" series has come to an end, Hazeman Huzir continues to be actively involved in other television projects and collaborations. In addition to his television work, Hazeman Huzir has established a presence on YouTube, where he showcases his podcast and engages in activism in various countries. Through his YouTube channel, he shares insightful discussions, interviews, and thought-provoking content related to various topics.

Hazeman Huzir receive a relatively high proportion of positive replies. Out of the total replies to positive tweets, approximately 62.7% (530 replies) are positive, while around 50.9% (409 replies) are negative. This suggests that positive tweets by Hazeman Huzir tend to generate a greater number of positive responses from his audience.

On the other hand, negative tweets by Hazeman Huzir also receive a notable proportion of positive replies. Approximately 37.3% (315 replies) of the replies to negative tweets are positive, while around 49.1% (394 replies) are negative. This indicates that even when sharing negative content, Hazeman Huzir still receives mixed response in term of sentiments. These findings suggest that Hazeman Huzir has managed to cultivate a supportive and engaged audience who actively responds to his tweets, regardless of the sentiment expressed in the original content. The significant number of positive replies to both positive and negative tweets reflect the influence and appeal that Hazeman Huzir holds among his followers.

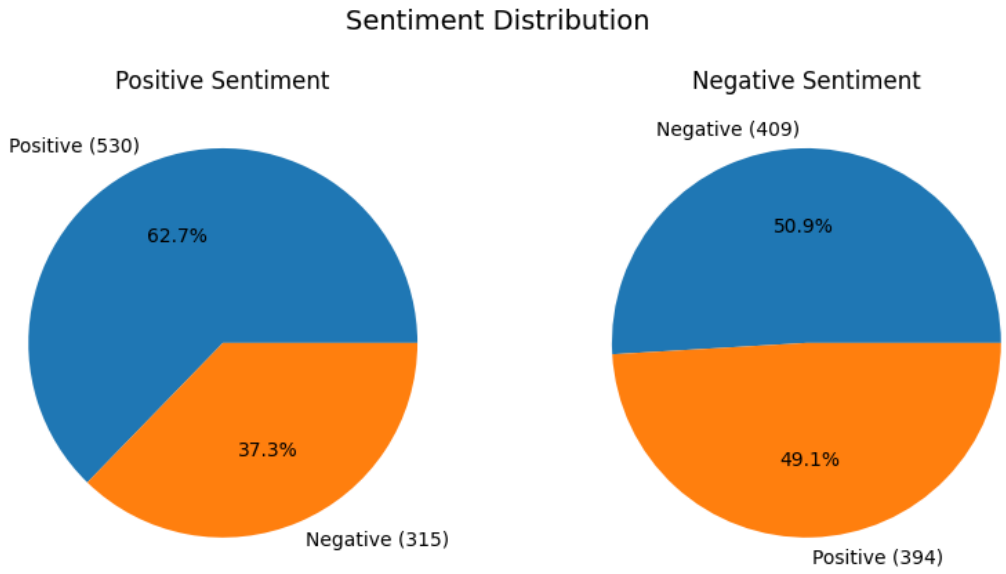


Figure 5.9 Sentiment distribution in replies (Hazeman Huzir)

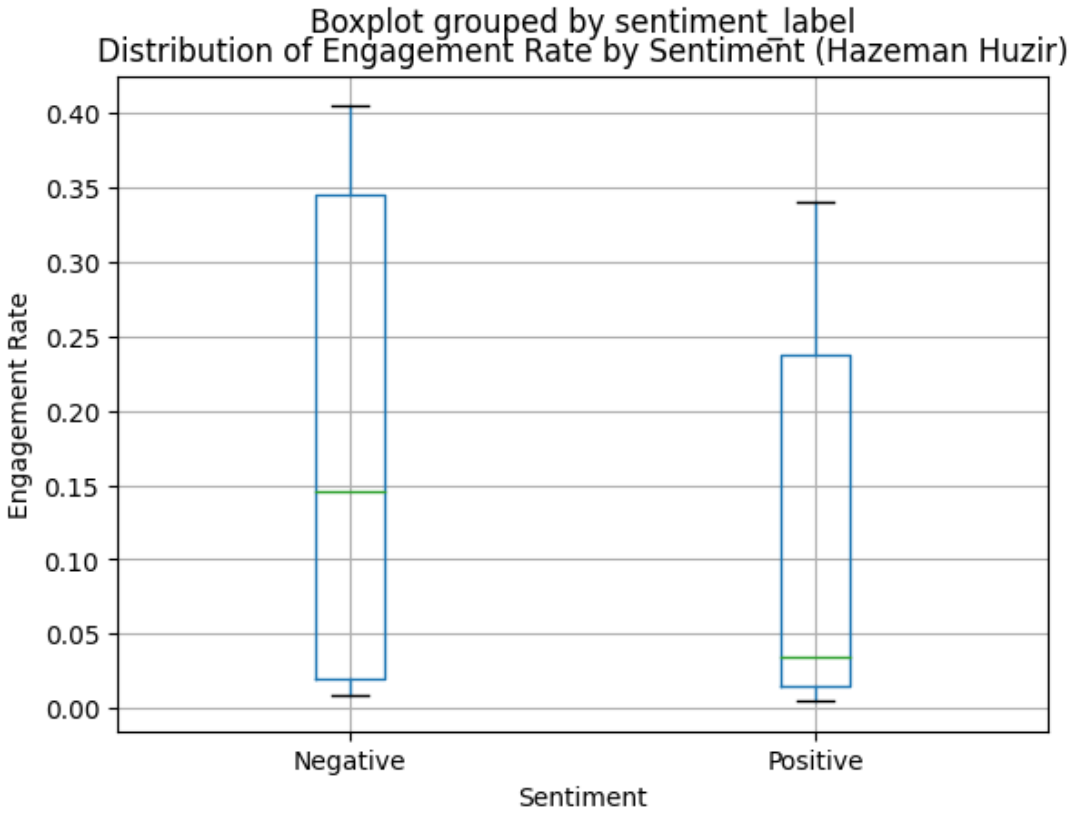


Figure 5.10 Box plot of engagement rate (Hazeman Huzir)

Table 5.11 Summary of box plot of engagement rate (Hazeman Huzir)

Sentiment	count	Mean	S.D	Min	25%	50%	75%	Max
Negative	21.0	0.910	2.278	0.009	0.02	0.146	0.344639	9.154
Positive	20.0	0.236	0.396	0.004	0.01	0.035	0.236950	1.283

These statistics provide insights into the distribution and variation of engagement rates for Hazeman Huzir's tweets. For negative sentiment tweets, the mean engagement rate is considerably higher at 0.910, with a wide standard deviation of 2.279. This indicates that there is significant variation in the engagement received for negative tweets, ranging from relatively low to very high engagement rates. In contrast, positive sentiment tweets have a lower mean engagement rate of 0.236, with a smaller standard deviation of 0.396. This suggests that the engagement received for positive tweets is relatively more consistent and less varied compared to negative tweets.

It is important to note that the maximum engagement rate for negative tweets is substantially higher (9.154) than the maximum engagement rate for positive tweets (1.283). This indicates that some negative tweets by Hazeman Huzir have generated exceptionally high levels of engagement, potentially due to the controversial or attention-grabbing nature of the content from the Motif Viral show.

As a TV host, Hazeman Huzir has successfully tackled negative issues with a respectful approach during his discussions on viral topics. However, the data suggests that there is room for improvement, as his positive sentiment tweets receive less engagement compared to his negative ones. To address this, one strategy he can employ is infusing positivity into his tweets, even when addressing negative issues. By doing so, he can create a more uplifting and encouraging environment for his followers. Additionally, it's important to consider the perspective of his audience. Given Hazeman Huzir's association with the "Motif Viral" show, it is possible that his audience is more inclined to respond to negative issues. Leveraging this understanding, Hazeman Huzir can strategically utilize negative topics to garner greater engagement. However, he must be cautious not to be perceived as promoting negativity or facing potential cancellation. To mitigate this risk, he should ensure that he demonstrates emotional intelligence in his content by approaching sensitive topics with empathy, respect, and understanding. By implementing these strategies, Hazeman Huzir can strike a balance between addressing negative issues responsibly and infusing positivity into his content, thereby fostering a more engaged and positive audience.

5.3.2.5 Yuna

Yuna, a renowned Malaysian singer with a global presence, has gained recognition for her exceptional talent. She initially established her online presence on Myspace as an indie musician. Figure 5.11 reveals that a significant majority of her tweets (81.3%) have received a positive sentiment. This could be attributed to Yuna's strategic use of her Twitter account for professional engagements and sharing positive content with her audience. However, it is worth noting that the analysis is limited due to the small number of tweets available for analysis (only 4, as shown in Table 5.12). Therefore, a comprehensive discussion of Yuna's online presence and its impact on audience reactions cannot be extensively discussed based on the available data.

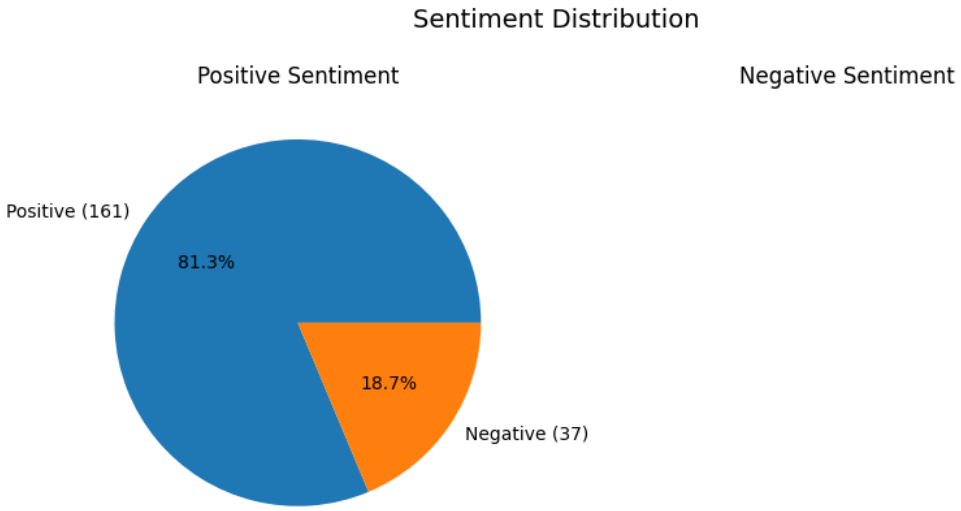


Figure 5.11 Sentiment distribution in replies (Yuna)

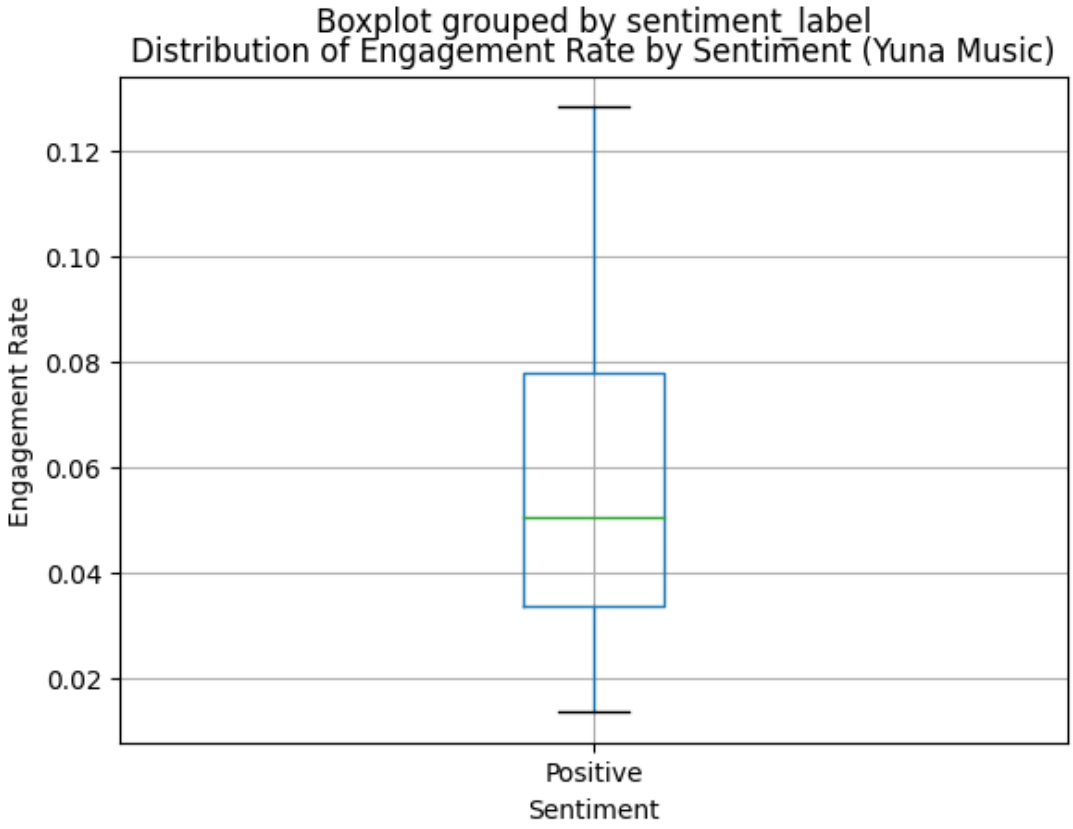


Figure 5.12 Box plot of engagement rate (Yuna)

Table 5.12 Summary of box plot of engagement rate (Yuna)

Sentiment	count	Mean	S.D	Min	25%	50%	75%	Max
Positive	4.0	0.061	0.049	0.014	0.034	0.051	0.078	0.128

5.3.3 Content Analysis

In this section, a content analysis will be conducted using the same usernames as in the previous section. The objective is to identify and analyze interesting patterns from a summarized table, which will provide valuable insights into the nature of the tweets associated with those specific patterns. The primary objective of this analysis is to gain a deeper understanding of the underlying factors that contribute to the observed patterns and their implications. Each table for content analysis is build based on data shown in Appendix A where the translated and cleaned words from tweets are collected in a word cloud and also presented in a bar chart showing the frequency of occurrences. This can roughly give hint of the content of the tweet in each category of parameter.

5.3.3.1 Syed Saddiq

Syed Saddiq's tweets predominantly revolve around political issues, reflecting the nature of a politician's communication. His tweets often address matters of public interest, including social issues, governance, and policy matters. Being a prominent political figure, Syed Saddiq utilizes Twitter as a platform to voice his opinions, share his views on various topics, and engage with his followers and the public. The content of his tweets reflects his role as a political figure and his active involvement in shaping public discourse. Syed Saddiq's tweets occasionally elicit positive sentiment among his followers and the wider audience.

Table 5.13 Summary of content coding (Syed Saddiq)

Engagement Rate	Resultant Sentiment count (Number of tweet)	Keyword	Label
High Engagement Rate	High Positive (4)	Subjects, Religious, University, Malaysia, Years, beach, valley, Sepang	Education system, Relief Effort
	Low Positive (5)	Flood, relief, Sri, items, session, government, mask, people, places, naughty	Relief Effort
	Low Negative (4)	Cha, people, rank, life, sad, drama, suicide, explanation, time, heard	Political Dispute, Government Criticism
	High Negative (0)	-	-
Low Engagement Rate	High Positive (5)	Young, focus, batik, flood, party, sexy, opposition, allowed, passed, muda	Muda, Relief Effort, General Information
	Low Positive (49)	People, minister, Government, flood, parliament, Muar, open, Putrajaya, luxury, time	MUDA updates, Political Dispute, Government Criticism, General Information, Relief Effort, Political Discussion
	Low Negative (29)	People, Government, shut, issue, case, Malaysia, person, open, wait, issues	Government Criticism, Political Dispute, Political Discussion, Relief Effort
	High Negative (2)	Death, punished, poor, journalists, good, question, answer, matter, language, peoples	Criticism of justice system, Political Dispute

Table 5.13 summarize the analysis obtained based on tweet from Syed Saddiq. In term of engagement rate, there is very few of the tweet that can be categorized as high engagement rate. Almost all of the label presence in high engagement rate are also presence in low engagement rate except for Education System which is fair as Syed Saddiq only made one tweet that falls under Education System. Most of the tweet that receive a high engagement revolve around the people which involve Syed Saddiq's team's effort to help people in need during flood and defending the people against Government's bad policy. The fact that there is no tweet that categorized as high negative resultant sentiment count shows that Syed Saddiq's tweet that receive high engagement is a subject that are in favours to the people and often involve debates to discuss and sharing of opinion.

Figure 5.13 and Figure 5.14 shows the tweet that receive high engagement rate and high positive resultant sentiment. As a politician in his youth, many of Syed Saddiq's audience are in his youth. The tweet's content, discussing the longer duration of university studies in Malaysia compared to other countries resonates with his audience because it highlights an issue that directly affects them. Many youths in Malaysia may relate to the tweet's message, as they might have experienced or are currently experiencing the longer duration of university studies in the country. This shared experience creates a sense of empathy and solidarity among the audience, leading to a higher likelihood of engagement and positive sentiment.

Figure 5.13 Sample of tweet

Generally, tweet that received high positive tweet is tweet that has no critical issue or discussion. Tweet labelled with Muda, General Information and Relief Effort often trying to deliver informative content rather initiating a discussion, one of the tweets is shown in Figure 5.14. This tweet provides information about Muda team and their progress on the relief efforts. Although this tweet received low engagement, high positive resultant sentiment suggest that this tweet is well-received by the audience as the tweet is about Muda team helping the community.

Figure 5.14 Sample of tweet

Based on Table 5.13, some of the keyword for low positive and low negative are very engaging such as minister, people, government and luxury. Although there is "flood" as one of the keywords, low positive and low negative tweet often discuss political issue. While it does promote critical discussion, these tweets often involve in political party with different view on certain topic where there can be mixed of sentiment. Despite the tweet expressing dissatisfaction or concern on specific policies or decision made by government or some other politician, the tweet received low engagement rate and this is due to target audience. As mentioned, majority Syed Saddiq's audience might comprise of youth where these group of people might have different interest. Generally, they may be more inclined towards topics like youth empowerment, social issues, or general information rather than political dispute or government related matters. This could explain the lower engagement rate despite the presence of engaging keywords in the tweets.

We have clear trend on Syed Saddiq's tweet where most of the tweet that received high engagement rate and positive sentiment revolved around topic that conform to youth. Syed Saddiq should continue to recognize that his audience predominantly consists of young people. By understanding their interests, concerns, and priorities, he can tailor his content to resonate more strongly with them. While the performance of the tweet could be beneficial, Syed Saddiq still need to diversify the content. Despite some tweet labelled with government criticism and political dispute received lower engagement rate, it is also important to note that some information is crucial to be communicated to the audience. Therefore, Syed Saddiq should strike a balance between addressing youth-centric topics and providing essential information related to government policies, political issues, and social matters.

5.3.3.2 Khairul Aming

Khairul Aming, a well-known influencer and content creator in the culinary realm, has not only achieved success in his online ventures but has also made a significant impact in the business world. With his popular cooking content and the successful sales of "Sambal Nyet", his business has garnered widespread attention and recognition. One crucial aspect that has contributed to Khairul Aming's success is public perception. As an influencer, he has built a strong and positive reputation among his followers and the wider audience. His expertise in cooking, combined with his engaging content and personable demeanour, has earned him trust and credibility in the culinary community. This positive perception has translated into a loyal customer base for his business.

Before discussing the content analysis, it is important to understand Khairul Aming's primary audience. As a content creator and business owner, Khairul Aming's target audience primarily consists of teenagers and young adults who engage with his content and follow his online presence. The analysis of Khairul Aming's tweets reveals distinct trends and patterns worth noting. Overall, Table 5.14 reveals a mixed trend in Khairul Aming's tweet analysis. It is notable that certain themes, such as reaching the audience, COVID-19 updates, and business updates, are present in both the high engagement rate category and the low engagement rate category. This suggests that while some tweets generate high engagement and positive sentiment, others may not resonate as strongly with the audience.

Among the tweet categories that receive positive sentiment, cooking content, business updates, and supporting small businesses stand out. These topics align with Khairul Aming's expertise and business ventures, indicating that his audience appreciates and engages with this type of content. By continuing to share valuable cooking insights, providing timely business updates, and promoting small businesses, Khairul Aming can strengthen his rapport with his followers and maintain their support. On the other hand, certain topics in Khairul Aming's tweets tend to generate negative aggregated sentiment. These topics include discussions related to COVID-19, general information, personal opinions, and reaching the audience. It's important

to note that discussions around sensitive or controversial issues, such as the ongoing pandemic or personal opinions, can elicit mixed reactions from the audience. While some individuals may agree or appreciate the perspectives shared, others may hold contrasting views, leading to negative sentiment. Similarly, when providing general information or trying to reach a wide audience, it's difficult to please everyone, and negative sentiment may arise from differing expectations or interpretations. Engaging in discussions and expressing personal opinions may also invite disagreement and contribute to negative aggregated sentiment. Understanding these trends can help Khairul Aming refine his content and communication strategies to minimize potential negative backlash while maintaining a strong connection with his audience.

Figure 5.15 and Figure 5.16 shows sample of tweet that categorized as “reaching the audience” receiving positive aggregated sentiment. However, Figure 5.15 receive high engagement while Figure 5.16 received low engagement. Upon closer look on the purpose of the tweet, Figure 5.15 serves as personal expression of surprise and gratitude for the high engagement received on a video. It aims to share the user's excitement and appreciation with their audience. In contrast, the Figure 5.16 serves the purpose of seeking recommendations for YouTube channels. It indicates the user's interest in finding entertaining and educational content. Despite these differences, both tweets received positive engagement, suggesting that the user's audience and followers responded positively to their content. The high engagement on the first tweet could be attributed to the personal connection and gratitude expressed, while the low engagement on the second tweet may be due to various factors such as the timing of the tweet or the specific audience's interests.

Figure 5.15 Sample of tweet

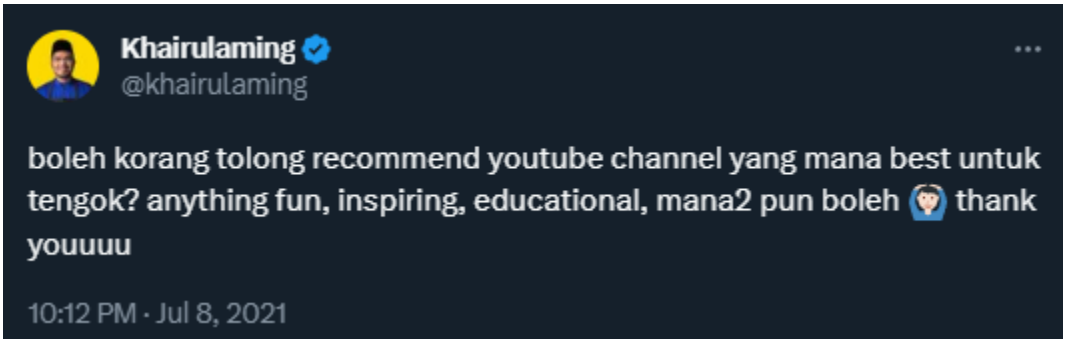


Figure 5.16 Sample of tweet

"Business updates" is indeed a significant component of Khairul Aming's tweets. This category of tweets generally receives positive engagement and a high engagement rate. However, there was an instance where a business update tweet received negative aggregated sentiment. This occurred when Khairul Aming addressed an unfortunate incident where his product was pirated and sold by fake resellers. In Figure 5.17, Khairul Aming shared his frustration regarding the issue, and as a result, the tweet received negative aggregated sentiment. Negative replies to the tweet were likely directed towards the problem itself and the fake resellers involved.

Figure 5.17 Sample of tweet

Table 5.14 Summary of content coding (Khairul Aming)

Engagement Rate	Resultant Sentiment count (Number of tweet)	Keyword	Label
High Engagement Rate	High Positive (5)	Views, video, today, sustenance, years, lot, people, cook, decrease, share	Reaching Audiences, Cooking Content, Business Updates, Supporting Small Businesses, Personal Opinion
	Low Positive (15)	Thing, day, people, today, lot, pls, video, ive, peoples, lives	General Information, COVID 19, Reaching Audiences, Cooking Content, Personal Opinion,
	Low Negative (3)	Please, jangan, panic, buying, guys, weve, harap, yang, terbaik, untuk	COVID 19, Business Updates
	High Negative (0)	-	-

Table 5.14 Summary of content coding (Khairul Aming) continued

Low Engagement Rate	High Positive (19)	Food, thread, free, promote, buy, guys, online, sahur, selling, time	Reaching Audiences, Supporting Small Businesses, General information, Cooking Content
	Low Positive (116)	Sambal, guys, bottles, time, day, people, nyet, tomorrow, today, chicken	Business update, General Information, Cooking Content, COVID 19, natural disaster, Personal opinion, Festive Greetings, Discussion
	Low Negative (18)	People, car, video, phone. Road, hit, driver, stressed, friends, allowed	Reaching Audience, General Information, Personal Opinion, COVID 19, Discussion
	High Negative (0)	-	-

To wrap up Khairul Aming's content analysis, several suggestions can be made to maintain his positive online image. Currently, his tweets have received mostly positive engagement, and he has been successful in delivering content that resonates with his audience at the right time. Khairul Aming has demonstrated proficiency in leveraging viral issues to create engagement while maintaining a positive perspective. Additionally, his efforts to keep his audience engaged with his cooking content have paid off, as evidenced by the high engagement and positive sentiment in that category. Moving forward, Khairul Aming can continue capitalizing on timely topics and viral

trends to maintain audience engagement. By staying up-to-date with current events and understanding his audience's interests, he can consistently deliver content that is relevant and appealing. It would also be beneficial for Khairul Aming to continue diversifying his content and exploring new themes that align with his audience's preferences. This could include incorporating elements of fun, inspiration, and education into his cooking content.

5.3.3.3 Xavier Naxa

As a tech influencer, Xavier Naxa does not need to worry too much on the sentiment of his reply. But engagement is very important as he will use his social media as method to promote store promotion or tech knowledge.

Based on Table 5.15, it is apparent that there is a scarcity of tweets classified as high engagement. Among the limited number of high engagement tweets, one is categorized as having a low positive sentiment, while the other falls under the low negative sentiment category. The tweet associated with low negative aggregated sentiment presents a movie review, focusing on a horror film and its concept. Additionally, there is a tweet addressing a political issue, where the influencer humorously referred to a politician as “full of herself”.

A significant portion of Xavier Naxa's tweets fall into the low engagement rate category. These tweets encompass various topics such as personal opinions, political issues, tech promotions, general information, and his attempts to connect with his audience. Unfortunately, these particular tweets have not garnered substantial engagement from his followers and the wider audience. Despite his efforts to share his perspectives, discuss political matters, promote technology, provide general information, and engage with his audience, these tweets have not generated a significant level of interest or interaction.

Table 5.15 Summary of content coding (Xavier Naxa)

Engagement Rate	Resultant Sentiment count (Number of tweet)	Keyword	Label
High Engagement Rate	High Positive (0)		
	Low Positive (1)	Tired, minister, notice, beauty	Political Issue
	Low Negative (1)	Film, medium, horror, concept, tells, life, shaman, Thailand, close, people	Movie review,
	High Negative (0)		
Low Engagement Rate	High Positive (2)	Coffee, relaxing, twitter, friends, share, mobile, game, entertaining, complex, mind	Discussion, Personal Opinion
	Low Positive (11)	Coffee, yb, iphone, government, year, share, promotion, malay, language, qwerty	Personal Opinion, Political Issue, Tech Promotion, Discussion, Reaching Audiences, General Information
	Low Negative (1)	Government, opposition, solved, flood, issue, times, party, choose, feel, boycotting	Political issue,
	High Negative (0)	-	-

Indeed, political issues appear in both the high and low engagement categories for Xavier Naxa. However, the main difference lies in the magnitude of the viral nature of these issues. In the case of high engagement, Xavier Naxa gained significant attention by calling out YB. DS. Rina for attempting to look pretty without any

apparent context attached to the tweet. This satirical comment received support from several Twitter users. This tweet tapped into the negative public perception surrounding YB. DS. Rina Harun, particularly due to her approach in providing relief support for flood victims, which has been widely criticized.

Figure 5.18 Sample of tweet

As a tech influencer, Xavier Naxa's followers and the wider audience have certain expectations regarding his tech-related content. However, the data suggests a lack of engagement in his tech-related tweet. One particular tweet stands out in Figure 5.19, which focused on an offer to change phone batteries specifically for iPhone 8 and iPhone 8 Plus. The primary reason for the low engagement can be attributed to the issue of relevancy.

In this case, the offer may not have resonated with Xavier Naxa's followers and the broader audience due to its limited scope. The tweet's content targeted a specific subset of iPhone users who owned the iPhone 8 and iPhone 8 Plus and were actively seeking battery replacement services. It's possible that at the time of the tweet, the demand for battery replacement for these particular models was relatively low, or Xavier Naxa's target audience consisted of users with different smartphone models.

To increase engagement despite the tweet being solely for promotion, Xavier Naxa can take a strategic approach by providing more information and context surrounding the promotion. Instead of simply announcing the offer, he can create an informative thread related to battery health and related topics. By doing so, Xavier

Naxa can gradually introduce the promotion within the context of valuable information.

For example, Xavier Naxa can start the thread by discussing the importance of maintaining optimal battery health for smartphones. He can share tips on how to prolong battery life, common battery issues, and the benefits of regular battery maintenance. Throughout the thread, he can highlight the significance of battery replacement as a solution to address potential battery-related problems.

As Xavier Naxa shares this valuable information, he can strategically weave in the promotion for battery replacement services. For instance, he can mention the specific offer for iPhone 8 and iPhone 8 Plus users and explain how it aligns with the importance of maintaining battery health. By providing relevant information and connecting it to the promotion, Xavier Naxa can create an emotional attachment for his audience as they read through the thread.

This approach not only increases engagement but also adds value to the promotion itself. It positions Xavier Naxa as a knowledgeable and helpful resource in the tech space, building trust and credibility among his followers. Additionally, by incorporating storytelling techniques or personal anecdotes related to battery issues, Xavier Naxa can further engage his audience and make the promotion more relatable.

As a conclusion, Xavier Naxa's effort to connect with the audience in different topics has proven beneficial in subtly showcasing his existence and generating a bit more engagement. He has successfully created several threads related to movie reviews, types of coffee, and general information, which have resonated with his followers and sparked discussion. However, some of his tweet is too straightforward which does not create room for further engagement.

Figure 5.19 Sample of tweet

5.3.3.4 Hazeman Huzir

As an influencer transitioning into informative content and humanitarian work, it is crucial to prioritize audience perception. Maintaining a positive image and reputation becomes paramount, as it influences the ability to engage effectively with the audience and attract potential stakeholders for business collaborations. Balancing informative content and engaging with the audience will be instrumental in capturing their attention and fostering meaningful connections.

Table 5.16 Summary of content coding (Hazeman Huzir)

Engagement Rate	Resultant Sentiment count (Number of tweet)	Keyword	Label
High Engagement Rate	High Positive (2)	People, great, today, initiative, fellow, citizens, social, media, energy, money	COVID-19, General Information
	Low Positive (1)	Entertaining, Malaysians, trouble, covid, lockdown, factories, ministers, money, night, kind	Influencer Dispute
	Low Negative (3)	Epidemic, people, covid, stupid, feeling, variant, worse, notice, deaths, plague	COVID-19, Natural Disaster, Personal Reflection
	High Negative (0)		

Table 5.16 Summary of content coding (Hazeman Huzir) (Continued)

Low Engagement Rate	High Positive (3)	Assalamualaikum, guys, wanted, ready, adopt, baby, vaccine, bangi, convention, good	COVID-19, General Information
	Low Positive (14)	People, body, eat, education, fish, guys, level, netizens, continue, thing	Political Issue, Knowledge sharing, Product promotion, COVID-19, Influencer Dispute,
	Low Negative (24)	People, time, life, stable, economy, live, Israel, lot, die, basic	Political Issue, COVID-19, Insurance issue, Influencer Dispute, Israel Issue
	High Negative (0)	-	-

Generally, a mixed distribution of tweet categories and labels can be found in Hazeman Huzir's analysis. However, a consistent trend that emerges is the presence of the label "COVID-19" across various categories. Hazeman Huzir has tweeted about COVID-19 in both high and low engagement scenarios, and the aggregated sentiment associated with these tweets can be positive or negative. COVID-19 is a topic that affects people's lives on a global scale, and it has generated a range of emotions and opinions among individuals. As a result, discussions and updates related to COVID-19 can elicit both positive and negative sentiments, depending on factors such as the nature of the information shared, personal experiences, and the overall context of the tweets.

One other frequent tweet is “influencer dispute”. As influencer, it is hard to not involve in these situations. However, issue sometime can be beneficial as these often attract attention and engagement from followers and other users who are interested in the drama or want to voice their opinions on the matter. While some tweets may have a positive sentiment, indicating support or agreement with the issue, others may have a negative sentiment, reflecting disagreement or criticism. These disputes can be polarizing, with different group of followers and users taking sides and expressing their views. Thus, Hazeman Huzir need to be careful when addressing another influencer.

Hazeman Huzir, being a YouTuber, primarily utilizes the platform to share his viewpoints and provide informative content on various topics. The labels "Personal Reflection" and "Knowledge Sharing" are commonly associated with his videos, indicating that these types of content are prevalent within his YouTube channel. One of the possible factors that drive high engagement for the tweet involving personal reflection is emotional connection. Personal reflection videos often tap into personal experiences, emotions, and storytelling, which can create a strong emotional connection with the audience. Viewers may find them relatable and engaging, leading to more comments, likes, and shares. In contrast, knowledge sharing videos may focus more on conveying factual information or instructional content, which may not elicit the same level of emotional response.

As an influencer and someone who is active as an activist helping people around the world, his positive image is valuable. Hazeman Huzir need to balance between engagement and his perceived image. To optimize his content strategy, Hazeman Huzir needs to strike a balance between sentiment and engagement in his videos. While personal reflection videos tend to generate higher engagement due to their emotional connection, unique perspectives, and storytelling, it's important for Hazeman Huzir to also focus on increasing engagement in his knowledge sharing videos.

5.3.3.5 Yuna

Yuna, a renowned Malaysian singer with a global presence, understands the significance of her image and networking, particularly in her connections with music producers and industry professionals. However, data in Table 5.17 suggests that Yuna has dedicated minimal effort to creating content on her Twitter account. Despite this, a few valuable insights can be obtained from her tweets. In general, her tweets receive limited engagement, encompassing topics such as reaching out to her audience, sharing birthday wishes, and offering personal reflections. The lack of engagement may be attributed to the tweets lacking actionable elements or calls to action. Notably, Yuna received significant engagement for her tweet expressing support for Nabila Azreen, a young Malaysian sprinter donning the hijab, as she made her Olympic debut at the Tokyo Games. This tweet gained traction due to the trending nature of the event, with Yuna's words of praise even making it into NST news.

Table 5.17 Summary of content coding (Yuna)

Engagement Rate	Resultant Sentiment count (Number of tweet)	Keyword	Label
High Engagement Rate	High Positive	-	-
	Low Positive (1)	Heart, Olympics, womens, morning, personal, nabila, azreen, omg, woo	National Sport Appreciation
	Low Negative	-	-
	High Negative	-	-
Low Engagement Rate	High Positive	-	-
	Low Positive (3)	Bday, songs, forgot, tweet, wishes, nov, ive, answers, love, memories	Reaching audience, Birthday wishes, Personal Reflection
	Low Negative	-	-
	High Negative	-	-

5.4 Conclusion

This chapter discuss on the analysis of sentiment analysis, highlighting irregularities found in SVM sentiment analysis results in the Celebrity category. A comparison of sentiment analysis results across categories is conducted, providing insights into variations in sentiment distribution. Selected personal brands are analysed to understand their engagement rate, sentiment, and theme of content.

The objectives of the research have been fully achieved through these analyses, contributing to the enhancement of sentiment analysis in personal branding and celebrity categories. As a wrap, positive and negative tweet does plays important role in determining the replies in term of the polarity of the sentiment and the engagement rate. In building personal brand without considering what category are they in, they need to balance the need to have high engagement rate and discussing negative issue as the cancel culture has been a huge factor in a downturn of specific person or a brand.

Furthermore, the analysis of sentiment and engagement rate in different categories highlights the importance of content analysis in shaping personal branding strategies. By conducting a thorough examination of the content being shared, individuals and brands can gain insights into the types of messages, themes, and topics that resonate most with their audience. This understanding allows them to craft compelling and relevant content that aligns with the preferences and interests of their target audience.

The result from content analysis is summarized as shown in Figure 5.20 by looking at the similarities in each category and the finding are used to construct a quadrant analysis. Based on this historic data, the criteria listed in Figure 5.20 for each category are the factor that influence the position on which quadrant specific tweet are on. This analysis allows us to understand the underlying patterns and factors that influence the engagement and sentiment of tweets within different categories. By understanding the patterns and factors that influence the engagement and sentiment of tweets within different categories, we can make suggestions for improvement and provide guidance for building a personal brand.

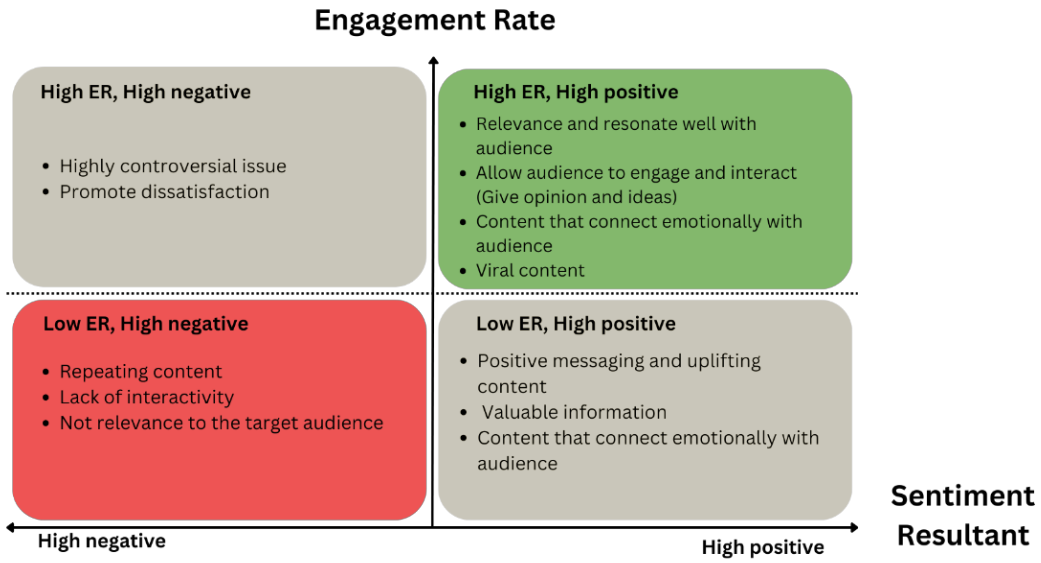


Figure 5.20 Similarities of tweet criteria from each quadrant

As discussed, we can classify certain tweet into one of the quadrants on Figure 5.21 depends on its value of sentiment resultant and the engagement rate, and what action the personal brand should apply to improve their tweets. Generally, tweet with high rate of engagement and high positive sentiment resultant would be the best one. Additionally, Figure 5.21 listed possible action that one can do, if their tweet is categorized in other quadrant other than high rate of engagement and high positive resultant sentiment.

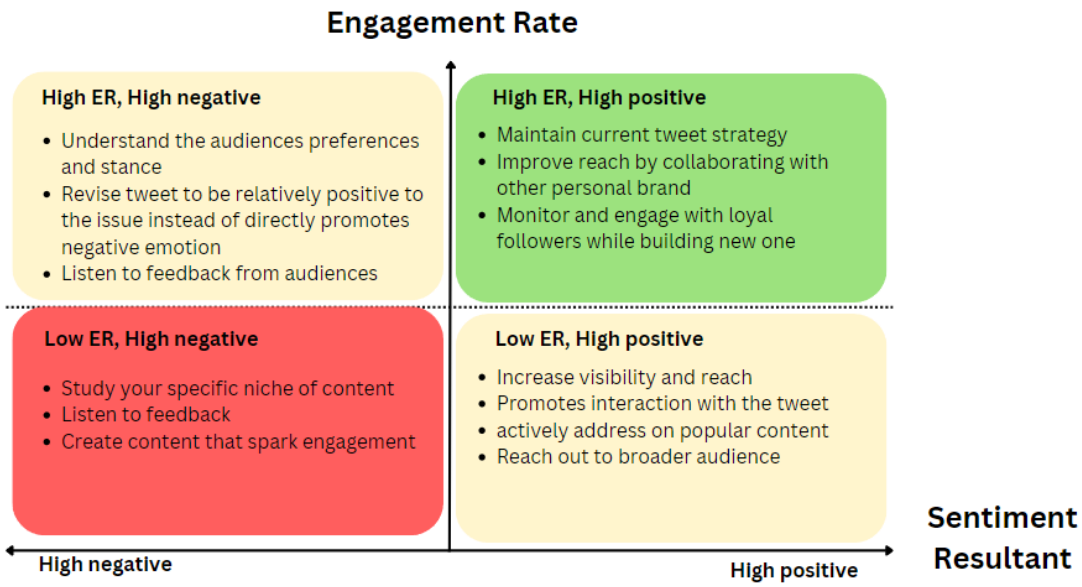


Figure 5.21 Strategies suggestion for each quadrant

In high rate of engagement and high positive resultant sentiment, the tweet has already in a very good position. To maintain this success, it is recommended to maintain the current tweet strategy, as it has proven effective in generating engagement. Additionally, the personal brand can collaborate with other personal brands or influencers to further improve reach and visibility. By monitoring and engaging with loyal followers and actively building new ones, the personal brand can foster a strong and dedicated fan base, resulting in sustained high engagement and positive sentiment.

When facing a low engagement rate but high positive sentiment, it is crucial to focus on increasing visibility and reach. The personal brand can achieve this by implementing strategies such as targeted advertising or partnerships with social media influencers in related fields. To promote interaction with tweets, call-to-action strategies can be employed, such as encouraging comments, likes, or shares. It is also important to address popular content that resonates with the target audience and adapt the content accordingly. Regular analysis and optimization of posting times, hashtags, and content topics can further enhance engagement.

In cases where there is a high engagement rate but also a high negative sentiment, it is important to understand the audience's preferences and stance. By

conducting sentiment analysis and identifying the root causes of negative sentiment, the personal brand can revise tweets to be relatively positive about the issue instead of directly promoting negative emotions. Active listening and responding to feedback from the audience can help address concerns and improve sentiment. Building a strong and open dialogue with followers can lead to better understanding and mitigation of negative sentiment.

For personal brands experiencing low engagement but high negative sentiment, understanding the specific niche of content is key. By studying the niche and target audience, the personal brand can tailor content to address their needs and preferences, thus increasing engagement. Listening to feedback from the audience is crucial, as it provides valuable insights for improvement. Creating content that sparks engagement, such as thought-provoking or controversial topics, can help attract attention and encourage followers to share their opinions.

CHAPTER 6

CONCLUSION AND RECOMMENDATION

6.1 Objective Achievement

The objective of this project discussed in Chapter 1 have been accomplished. The objective that includes to classify the sentiment of the tweet, to measure the engagement rate of the tweet and to study the relationship between sentiment of the tweet and engagement rate of the tweet.

6.1.1 Objective 1

The first objective is to perform sentiment classification on the tweets and their replies using hybrid method. The tweets were scraped using SNScrape and Twitter API based on specification discussed in Section 3.5.1. These tweets are then organized to be label based on which categories of personal brand are they belong to, and also whether they are a tweet or a reply. These labels are important to ensure filtering and grouping of the tweet which is crucial for analysis of the tweet. The tweets are then cleaned to remove stopword, remove username, remove URL and to translate the tweet into English language using Google Translator Package. The tweets have been classified as either positive or negative using the VADER sentiment classifier. The result from VADER classifier will be recorded into another column which will later be serve as an input variable for the SVM model. These tweets are then subjected to further processing, including splitting, vectorization, and will using SVM. The performance of the SVM model is evaluated using various metrics such as confusion matrix, accuracy, recall, precision, and F1-score.

The result from these hybrid sentiment analysis process shows that the highest accuracy achieve is when all the category of the personal brand are combined which is 96% while the accuracy of each category varies from 82% to 94% with Technology

category and Politic category with lowest and highest accuracy respectively. The value of precision and recall are relatively high except for Celebrity category in Negative sentiment where the value recorded at 55%. The possible source of error has been discussed and investigated. As a summary of the discussion, the low recall value was cause by lack of data in term of volume.

6.1.2 Objective 2

The second objective of this research is to measure the engagement rate of each tweet in the dataset. The method to calculate the engagement rate have been discussed in Section 3.6.5.1 where the formula takes number of followers of the personal brand into consideration. The engagement rate are one of the crucial information that will be used for analysis of objective 3.

6.1.3 Objective 3

The third objective is to discuss the relationship between the sentiment of the tweets, the replies and the engagement rate. The discussion for this objective is divided into 3 part which are category analysis, personal brand analysis and content analysis.

The category analysis is where we discuss the effect of sentiment and engagement rate between the category. As a summary, the data shows that Positive tweets generally receive positive replies in the whole category especially Celebrity where the percentage is the biggest. The least percentage can be found in Politic category. Despite that, Politic category receive relatively balanced sentiment on replies. Negative tweet still receives positive replies, but the highest percentage of negative replies on negative tweet is Politic category followed by Entertainer category. In term of engagement rate, the data shows that the Politic, Celebrity, and Entrepreneur category received relatively higher engagement when tweeting with a negative sentiment while Technology and Entertainer received a higher engagement when tweeting positive sentiment.

Personal brand analysis involves delving into each selected personal brand to examine their sentiment and engagement rate. In this analysis, the same parameters are measured and compared, but the discussion mainly revolves around the specific personal brand's background and characteristics. The first personal brand, Syed Saddiq, is a young politician in the Political category. The data reveals that positive tweets from Syed Saddiq receive more positive replies and higher engagement. Therefore, Syed Saddiq should focus on reorganizing his sentences to convey a positive tone. Next, Khairul Aming, an entrepreneur and cooking content creator, receives an equal proportion of positive replies and high engagement for both positive and negative tweets. Hazeman Huzir, a TV host, content creator, and activist known for his volunteer work, generally receives a higher portion of positive replies regardless of the sentiment of his tweets. However, the results show that his negative tweets generate more engagement compared to his positive ones. As an activist, maintaining a positive image is crucial, and posting negative tweets without damaging that image is risky. Thus, Hazeman should ensure that his negative tweets always align with his audience's perspective to avoid potential backlash. Unfortunately, there is insufficient data available for Xavier Naxa, with only one negative tweet, and Yuna, who has no negative tweets at all.

Content analysis involves categorizing and analyzing the themes of tweets from the selected personal brands based on sentiment and engagement rate. The categories of sentiment and engagement rate ("high" and "low") are analyzed separately to understand the patterns exhibited by each category. The objective of this study is to examine the sentiment of tweets and replies regarding a selected group of Twitter celebrities and personal brands to understand their perceived image and messaging effectiveness. Positive and negative tweets are analyzed to compare the sentiment across different categories of personal brands. The impact of this sentiment on the engagement rate within each category and for specific personal brands is also examined.

To further explore the relationship between sentiment and replies, the content of the tweets is coded to understand the type of content and its influence on sentiment and engagement rate. The results indicate that tweets with high engagement rates tend

to be relevant and resonate well with the specific audience, inciting engagement and emotional connection. They often address viral issues and provide valuable information. Conversely, tweets with low engagement rates tend to repeat content, lack interactivity, and are irrelevant to the target audience. Moreover, tweets receiving a relatively high number of positive replies are usually positive, uplifting, and emotionally engaging, while tweets generating more negative replies typically touch on highly controversial topics or express dissatisfaction.

6.2 Limitation of Project

In this project, the main issue is the quantity, quality and availability of the data. Initially, SNScrape was available as a third-party package that would allow anyone to extract Twitter data without Education Twitter API. However, upon finding the Celebrity category dataset is lacking in quantity and quality, another scraping activity is not possible due to change in policy which probably executed after change in Twitter CEO.

The formula of engagement rate that are calculated for each tweet require the number of likes, number of retweets, number of replies and number of followers for specific tweet and at the specific time that particular tweet was created. However, SNScrape or Twitter API in general do not have the ability to scrape the number of followers on that specific time. Thus, the number of followers is scraped as number of followers on the day of scraping takes place instead of when tweet is created. One other alternative as applied by Twitter was to use the number of impressions which is the number of times the tweet was opened. This value is still a value that would changes overtime which would make it not a fixed value. However, this is one of the data that are allowed to be scraped by Twitter API.

Other than that, since most of the tweet extracted are in in complete Malay language, translating the text from Malay to English is not 100% accurate. The method to reconstruct the poorly typed word in Malay language are not widely available for our usage.

6.3 Suggestion for Future Works

The project idea is a very beneficial field as the internet community has awakened on the strength of social network. This field of project can be beneficial as personal business or even as a psychology studies. Couple of the possible additional studies can be done is to expand the dataset. Instead of limiting to only one type of data which is text, other form of data which is videos or audio also can be used to collect a more holistic data and additionally, reaching different community such as Instagram or Tiktok. Using a wider range of dataset in term of user candidate might also give a clearer trend and insight.

Next, instead of a broad analysis, consider focusing on specific areas of research or study within the realm of personal branding and sentiment analysis. For example, investigate the sentiment associated with specific products or services, or dive into specific topics or themes within tweets to uncover more evidence for certain communication strategies or trends online. This targeted approach can yield deeper insights and contribute to specialized knowledge in specific domains.

Since the majority of the extracted tweets are in the Malay language, future research can focus on improving language processing techniques specific to Malay. This can involve developing language models or algorithms that better handle the nuances and unique characteristics of the Malay language. This advancement can enhance the accuracy of sentiment analysis and improve the overall quality of the findings.

Engaging with industry experts, social media professionals, or marketing practitioners can provide valuable insights and guidance. Collaborations can offer real-world perspectives, access to proprietary data, and expertise in personal branding and social media strategies, leading to more impactful research outcomes and practical applications.

6.4 Conclusion

As a conclusion, this research offers practical benefits where it will help discover insight into how specific personal brand or niche of personal brands are perceived on Twitter or social media in general. By understanding the sentiment of tweets and replies, brands can gain valuable feedback on their communication strategies, identify areas for improvement, and make data-informed decisions to enhance their online presence and reputation.

REFERENCES

Alina Andreevskaia, Sabine Bergler. (2008). When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging. *Proceedings of ACL-08*: HLT, pages 290–298.

A. Mudinas, D., Zhang, M., & Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining* (pp. 1-8). ACM. New York, NY, USA.

B. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *In Proceedings of the tenth ACM international conference on Knowledge discovery and data mining* (pp. 168-177). Seattle.

Bing Liu. (2012). Sentiment analysis and opinion mining. Sentiment Analysis and Opinion Mining | *Synthesis Lectures on Human Language Technologies*. (2012, May). Retrieved January 11, 2023, from

Boiy, E., Hens, P., Deschacht, K., & Moens, M. F. (2007). Automatic sentiment analysis in on-line text. *In Proceedings of the 11th International Conference on Electronic Publishing* (Vienna, Austria

Brems, Cara & Temmerman, Martina & Graham, Todd & Broersma, Marcel. (2016). PERSONAL BRANDING ON TWITTER: How employed and freelance journalists stage themselves on social media. *Digital Journalism*. 5. 10.1080/21670811.2016.1176534.

Chaovalit, P., & Zhou, L. (2005). Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 112c-112c.

Chirobocea-Tudor, Olivia. (2017). The Good and the Bad of the Corpus-Based Approach (or Data-Driven Learning) to ESP Teaching. *Scientific Bulletin "Mircea cel Batran" Naval Academy*. XX. 364-371. 10.21279/1454-864X-17-I1-058.

De la Torre-Díez, I., Díaz-Pernas, F. J., & Antón-Rodríguez, M. (2012). A content analysis of chronic diseases social groups on Facebook and Twitter. Telemedicine Journal and E-health: The Official Journal of the American Telemedicine Association, 18(6), 404-408. doi:10.1089/tmj.2011.0227

E. Cambria, (2016)., Affective Computing and Sentiment Analysis. *IEEE Intelligent System*s, vol. 31, no. 2, pp. 102-107, Mar.-Apr. 2016, doi: 10.1109/MIS.2016.31.

E. Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 271-278

García-Díaz, P., Sánchez-Berriel, I., Pontiel-Martín, D., & González-Ávila, J. L. (2023). A novel flexible feature extraction algorithm for Spanish tweet sentiment analysis based on the context of words. Expert Systems with Applications, 212, 118817. doi: 10.1016/j.eswa.2022.118817

Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017, October). Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. *In 2017 International Conference on Information and Communication Technology Convergence* (ICTC) (pp. 138-140). IEEE. doi: 10.1109/ICTC.2017.8190959.

Hassan, M. (2022, October 4). Content Analysis - Methods, Types and Examples. Retrieved from https://researchmethod.net/content-analysis/

Humphreys, Lee & Gill, Phillipa & Krishnamurthy, Balachander. (2014). Twitter: a content analysis of personal information. Information. 17. 10.1080/1369118X.2013.848917.

Https://aliabdaal.com/

Jeremy, D. (2021). The best python sentiment analysis package: 1 huge common mistake. Retrieved from https://towardsdatascience.com/the-best-python-sentiment-analysis-package-1-huge-common-mistake-d6da9ad6cdeb

Kralj Novak P, Smailović J, Sluban B, Mozetič I (2015) Sentiment of Emojis. *PLoS ONE 10*(12): e0144296. https://doi.org/10.1371/journal.pone.0144296

Muhammad, A., Wiratunga, N., & Lothian, R. (2014, November). A Hybrid Sentiment Lexicon for Social Media Mining. *IEEE 26th International Conference on*

De la Torre-Díez, I., Díaz-Pernas, F. J., & Antón-Rodríguez, M. (2012). A content analysis of chronic diseases social groups on Facebook and Twitter. Telemedicine Journal and E-health: The Official Journal of the American Telemedicine Association, 18(6), 404-408. doi:10.1089/tmj.2011.0227

E. Cambria, (2016)., Affective Computing and Sentiment Analysis. *IEEE Intelligent System*s, vol. 31, no. 2, pp. 102-107, Mar.-Apr. 2016, doi: 10.1109/MIS.2016.31.

E. Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 271-278

García-Díaz, P., Sánchez-Berriel, I., Pontiel-Martín, D., & González-Ávila, J. L. (2023). A novel flexible feature extraction algorithm for Spanish tweet sentiment analysis based on the context of words. Expert Systems with Applications, 212, 118817. doi: 10.1016/j.eswa.2022.118817

Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017, October). Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. *In 2017 International Conference on Information and Communication Technology Convergence* (ICTC) (pp. 138-140). IEEE. doi: 10.1109/ICTC.2017.8190959.

Hassan, M. (2022, October 4). Content Analysis - Methods, Types and Examples. Retrieved from https://researchmethod.net/content-analysis/

Humphreys, Lee & Gill, Phillipa & Krishnamurthy, Balachander. (2014). Twitter: a content analysis of personal information. Information. 17. 10.1080/1369118X.2013.848917.

Https://aliabdaal.com/

Jeremy, D. (2021). The best python sentiment analysis package: 1 huge common mistake. Retrieved from https://towardsdatascience.com/the-best-python-sentiment-analysis-package-1-huge-common-mistake-d6da9ad6cdeb

Kralj Novak P, Smailović J, Sluban B, Mozetič I (2015) Sentiment of Emojis. *PLoS ONE 10*(12): e0144296. https://doi.org/10.1371/journal.pone.0144296

Muhammad, A., Wiratunga, N., & Lothian, R. (2014, November). A Hybrid Sentiment Lexicon for Social Media Mining. *IEEE 26th International Conference on*

Statusbrew. (2023, May 26). How To Calculate Social Media Engagement Rate In 2023. Retrieved from https://statusbrew.com/insights/calculate-social-media-engagement-rate/

Sullivan, M. E. (2017, November). Showing your public face: Does screening social media assess residency applicants' professionalism? *American journal of obstetrics and gynecology*. Retrieved from https://pubmed.ncbi.nlm.nih.gov/28784415/

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the Association for Computational Linguistics* (pp. 417-424).

Verma, S. (2022). Sentiment analysis of public services for smart society: Literature review and future research directions. *Government Information Quarterly*, 39(3), 101708. https://doi.org/10.1016/j.giq.2022.101708

Will Koehrsen, (2018). A Feature Selection Tool for Machine Learning in Python. Retrieved from: https://towardsdatascience.com/a-feature-selection-tool-for-machine-learning-in-python-b64dd23710f0

Yulietha, I & Faraby, S & Adiwijaya, Kang & Widyaningtyas, W. (2018). An implementation of support vector machine on sentiment classification of movie reviews. Journal of Physics: Conference Series. 971. 012056. 10.1088/1742-6596/971/1/012056.

Appendix A Content Analysis Guideline

Khairul Aming

Label	Description
Reaching Audience	To reach and engaging with the target audience
Cooking content	Provides content related to cooking, recipes, picture of cooked food, or bought food
Business Updates	Any updates and announcement related to Khairul Aming's business activities
Supporting small businesses	Focus on thread or single tweet that promotes or provide medium to promote other small businesses
Personal Opinion	General personal opinion, thoughts, or perspective on general topic
COVID-19	Information, updates or discussion related to COVID-19 pandemic
Natural Disaster	Information, responses or discussion regarding natural disaster
Festive Greetings	Any greeting, wishes, or celebration related tweet

Label	Description
Political Issue	Involves discussions, news, or opinions related to political matters
Movie Review	Contains reviews, critiques, or discussions about movies or films
Discussion	Initiates or engages in general discussions on various topics.
Tech Promotion	Promotes or advertises technology-related products, services, or innovations
Personal Opinion	General personal opinion, thoughts, or perspective on general topic
Reaching Audience	To reach and engaging with the target audience
General Information	Provides general information, updates, or announcements on diverse topics.

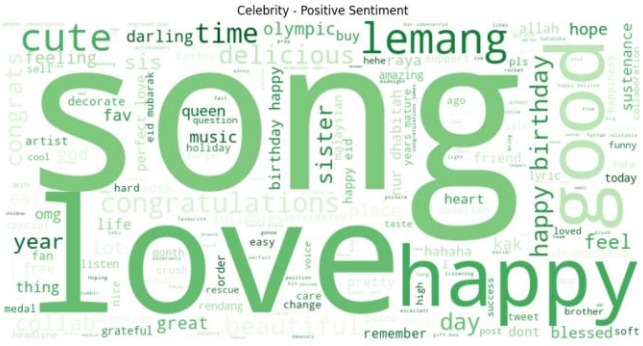
Label	Description
General Information	Provides general information, updates, or announcements on diverse topics.
COVID-19	Involves discussions, news, opinions, or updates related to the COVID-19 pandemic.
Influencer dispute	Contains debates, or controversies surrounding influencers or influencer-related matters.
Natural disaster	Focuses on discussions, news, or experiences related to natural disasters.
Personal Reflection	Contains personal thoughts, reflections, or experiences on various topics.
Political Issue	Involves discussions, news, or opinions related to political matters.
Knowledge Sharing	Shares information, insights, or expertise on specific topics.
Product Promotion	Promotes or advertises products or services
Insurance Issue	Focuses on discussions or news related to insurance matters
Israel Issue	Involves discussions, news, or opinions specifically related to the country of Israel

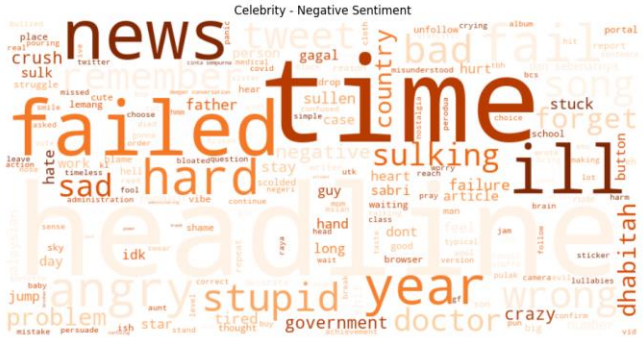
Yuna

Label	Description
National Sport Appreciation	Celebrates and shows appreciation for a specific national sport or sports event.
Reaching Audience	Focuses on strategies, tips, or discussions related to reaching and engaging with a target audience.
Personal Reflection	Contains personal thoughts, reflections, or experiences on various topics.
Birthday Wishes	Includes greetings, well-wishes, or celebrations related to someone's birthday.

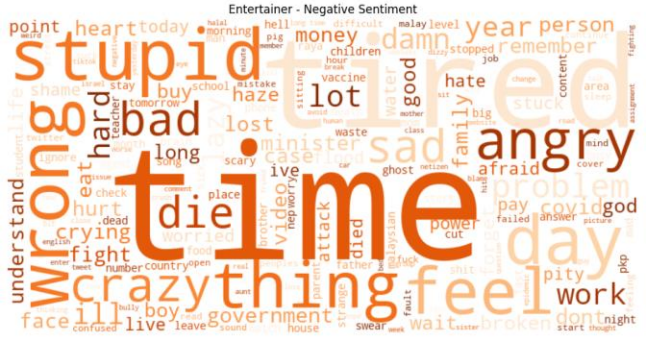
Appendix B Wordcloud

Celebrity



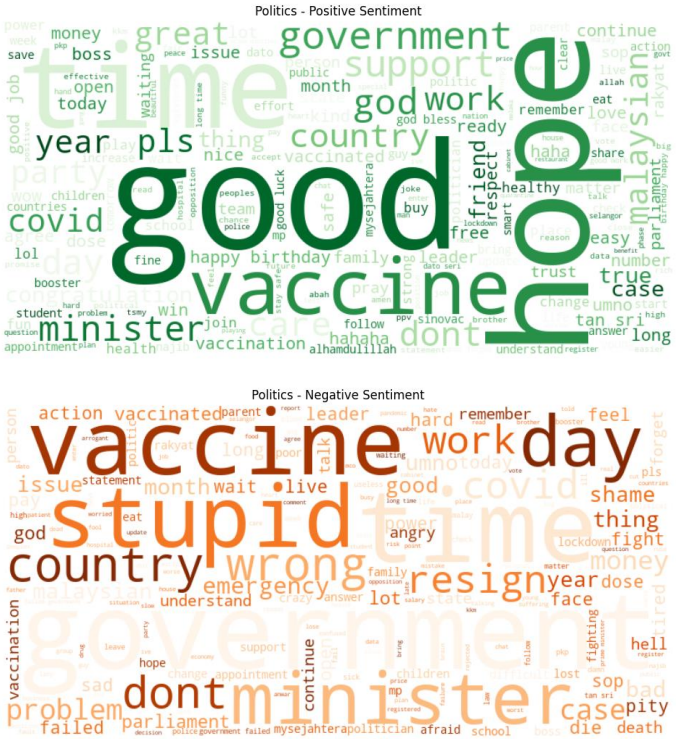


Entertainer



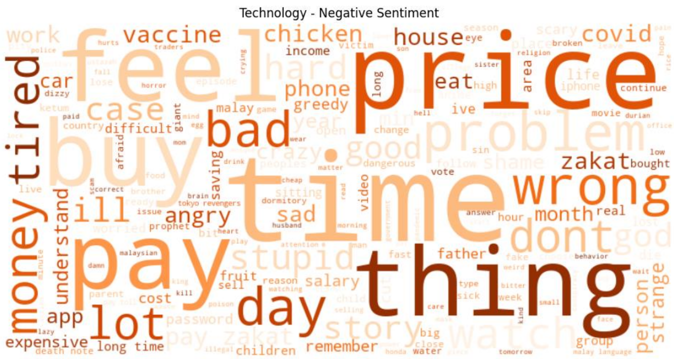
Entrepreneur

Politic

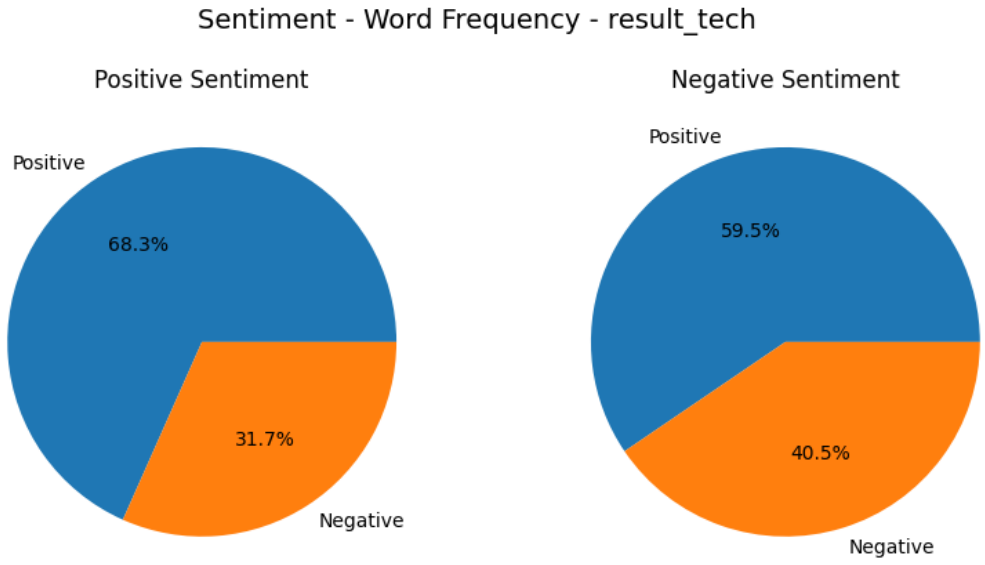


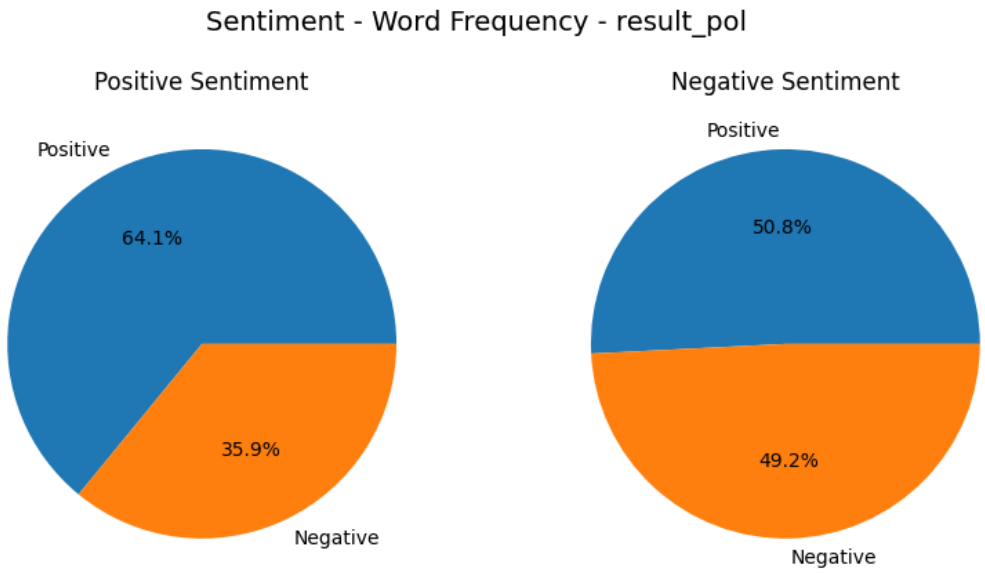
Technology

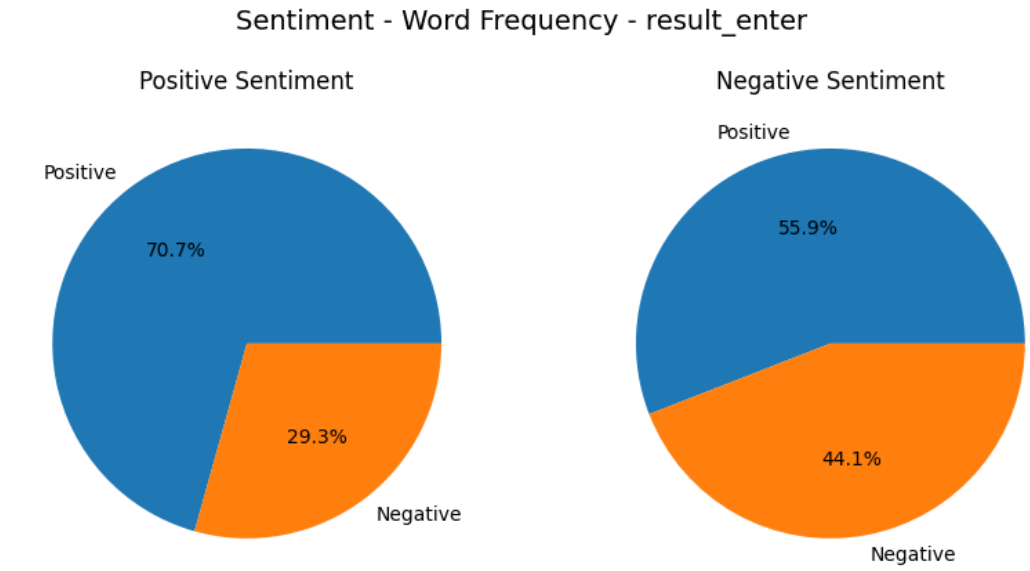

Technology - Positive Sentiment
good
buy
time
price
god
haha
feel
beautiful
phone
lot
house
care
easy
play
share
work
cheap
long
fun
continue
case
support
ready
hope
area
video
game
iphone
clean
nice
dont
eat
food
delicious
expensive
coffee
thing
person
great
save
open
place
taste
free
sell
giant
zakat
bought
day
kind
hahaha
friend
watch
year
agree
likes
love
mother
fresh
sustenance
mattress
alhamdulillah
ive
pay
water
month
car
app
mosque

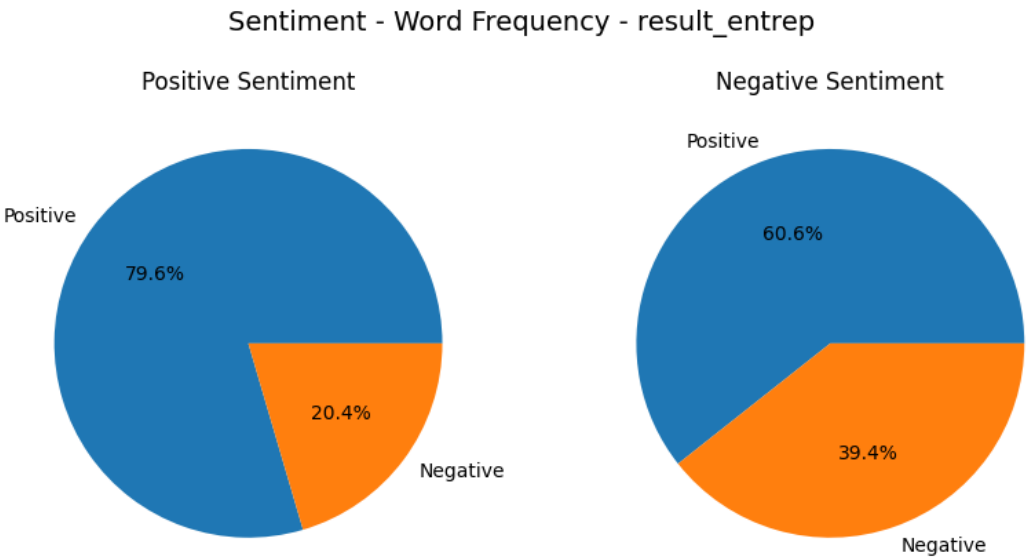

Technology - Negative Sentiment
price
time
thing
pay
buy
feel
problem
wrong
money
tired
dont
day
lot
bad
ill
vaccine
chicken
house
covid
zakat
stupid
story
watch
good
person
strange
angry
sad
understand
case
phone
hard
eat
greedy
salary
month
app
expensive
long time
children
remember
father
crazy
car
work
income
difficult
fruit
sell
cost
password
pay zakat
water
group
real
area

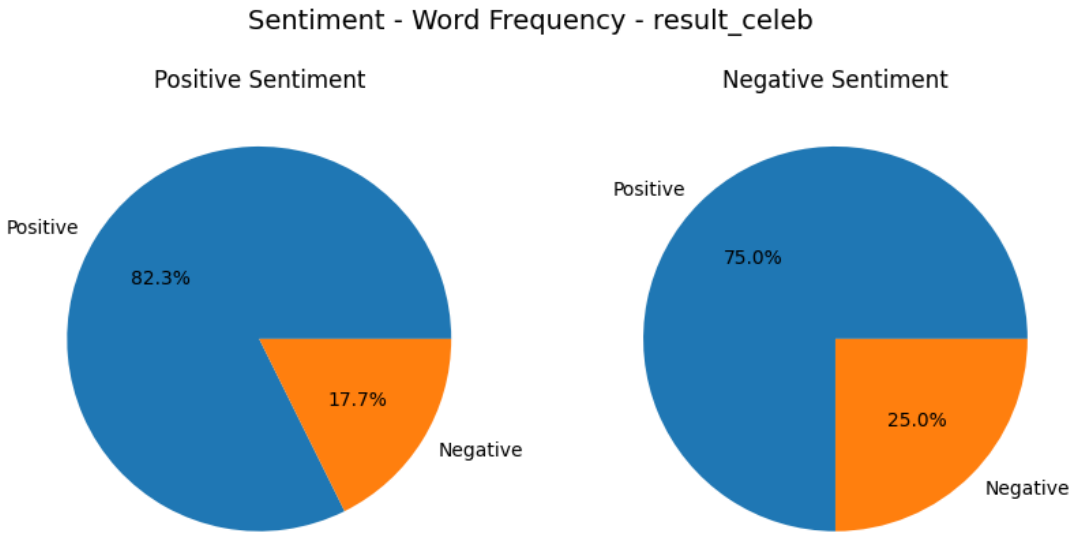
Appendix C Sentiment Distribution of Replies






Sentiment - Word Frequency - result_enter
Positive Sentiment
Positive
70.7%
29.3%
Negative
Negative Sentiment
Positive
55.9%
44.1%
Negative


Sentiment - Word Frequency - result_entrep
Positive Sentiment
Positive
79.6%
20.4%
Negative
Negative Sentiment
Positive
60.6%
39.4%
Negative


Sentiment - Word Frequency - result_celeb
Positive Sentiment
Positive
82.3%
17.7%
Negative
Negative Sentiment
Positive
75.0%
25.0%
Negative