Universiti Teknologi MARA

Sentiment Analysis of Malaysian Insurance Companies (SAMIC): A Visualization using Support Vector Machine Algorithm

Nur Farhana Binti Ahmad

Thesis submitted in fulfilment of the requirements for Bachelor of Computer Science (Hons.) Faculty of Computer and Mathematical Science

February 2021

SUPERVISOR APPROVAL

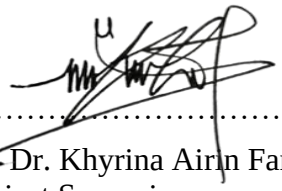
Sentiment Analysis of Malaysian Insurance Companies (SAMIC): A Visualization using Support Vector Machine Algorithm

By

NUR FARHANA BINTI AHMAD
2019314693

The thesis was prepared under the supervision of the project supervisor, Ts. Dr. Khyrina Airin Fariza Binti Hj Abu Samah. It was submitted to the faculty of Computer and Mathematical Sciences and was accepted in partial fulfilment of the requirements for the degree of Bachelor of Computer Science (Hons).

Approved by



…….

Ts. Dr. Khyrina Airin Fariza Binti Hj Abu Samah
Project Supervisor

FEBRUARY 20, 2021

STUDENT DECLARATION

I clarify that this thesis and the project to which it refers is the product of my own work and that any idea or quotation from the work of other people, published or otherwise are fully acknowledged in accordance with the standard referring practices of the discipline.

……

NUR FARHANA BINTI AHMAD

2019314693

FEBRUARY 20, 2021

ACKNOWLEDGMENT

Alhamdulillah, praises and thanks to the Allah S.W.T, the almighty, for his showers of blessings that I could complete this thesis within the given time frame. First and foremost, I would like to express to my supervisor, Ts. Dr. Khyrina Airin Fariza Binti Haji Abu Samah, my deep and sincere gratitude for her invaluable guidance, patience, enthusiasm, and immense knowledge. I could not have imagined having a better supervisor for my CSP600 and CSP650 subjects. I am immensely thankful for what she has taught me, and it was a great privilege and honor to be under her guidance. Throughout all the research and writing of this thesis, her continuous encouragement supported me. She has taught me as clearly as possible the importance of properly formatted and representation of documentation. I would also like to thank her for the friendship and her great sense of humor.

Besides, I would like to thank my lecturer, Ts. Dr. Shafaf Binti Ibrahim for providing assistance and willingness to assist in any way that she could throughout the project's completion. I am extending my heartfelt thanks to my examiner, Madam Nor Fadilah Tahar@Yusoff, for taking her time to read my thesis despite her busy schedule. I appreciate the valuable feedback, constructive remarks, and corrections that have undoubtedly led to this thesis's betterment.

I would like to acknowledge with gratitude my family's support and love. I am incredibly grateful to my family for their constant source of support for inspiration, valuable prayers and spiritual support throughout my life, and all their sacrifices for my future education. Finally, I would like to thank my fellow friends for enlightening me with my project's first glimpse. This thesis would not have been possible without the help of a person who could not be mentioned here but has well played his part to inspire me by providing his valuable assistance and time to complete this thesis.

ABSTRACT

Insurance is one of Malaysia's largest and most important financial services and is seen as a growing sector that makes tremendous progress and plays a socio-economic role. Many insurance companies operate in Malaysia, with 44 companies registered in the list of authorized insurance companies and Takaful operators. The complexities of insurance purchasing issues include assessing financial needs and choosing an insurance policy from which the established companies became confusing and challenging for potential policyholders as they will be entering into a long-term investment. They need to allocate adequate time to review each insurance company's offer to make a wise decision. Thus, this study proposed implementing a web application to visualize Malaysia's best insurance companies. Twitter was used as a source of data in this project. The tweets extracted using dates and keywords were analyzed as it is one of the metrics that will advance insurance companies' online presence. The data was first pre-processed, and the model is run on real-world data once the development of the machine learning models is completed. The model evaluation is conducted using the support vector machine (SVM) classifier, and two machine learning models have been built for classification purposes built based on the datasets in English and Bahasa Malaysia for multi-class text classification to represent the sentiment generation. Real-world data analysis results are visualized in a dashboard to make the analysis outcomes readable and understandable for the policyholders. Additionally, the web application also includes real-time monitoring of tweets sentiment that uses the data processing infrastructure and sentiment analyzer to evaluate public sentiment changes in response to user search keywords. Testing phases have shown that the classifier successfully classified tweets' sentiment with 90% accuracy and achieved all the project objectives. The future work that can be put into this project is to increase the number of neutral datasets for multi-class classification.

TABLE OF CONTENTS

CHAPTER FOUR: DESIGN AND DEVELOPMENT

CHAPTER FIVE: RESULTS AND DISCUSSION

CHAPTER SIX: CONCLUSION

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DM	Data mining
SA	Sentiment Analysis
NLP	Natural Language Processing
ML	Machine Learning
TF-IDF	Term Frequency-Inverse document frequency
RF	Random Forest
ANN	Artificial Neural Network
SVM	Support Vector Machine
GA	Genetic Algorithm
NB	Naïve Bayes
KNN	K Nearest Neighbour
BTF	Brand Talkable Favourability
SDLC	System Development Life Cycle
UI	User Interface
NLP	Natural Language Processing
SUS	System Usability Score

CHAPTER 1

INTRODUCTION

This chapter outlines the background and purpose of the study. It also gives details of the problem statement, objectives, scope, and significance of the study 'Sentiment Analysis of Malaysian Insurance Companies (SAMIC): A Visualization using Support Vector Machine Algorithm'.

1.1 Background of Study

Insurance is one of Malaysia's biggest and most important financial services and is seen as a growing industry with substantial progress and plays a socio-economic role in the economy (Bao, Ramlan, Mohamad, & Yassin, 2018). The insurance industry, including Islamic insurance, is a financial institution that offers risk management services, induces liquidity, diversifies financial losses, and promotes investment in the economy. According to the Malaysia Financial Sector Assessment Program (2018) report, Malaysia's insurance sector accounted for about 6% of financial sector assets, generally divided into life insurance and general insurance. In general, insurance companies can be categorized as health and medical insurance, auto insurance, accidental insurance, life insurance, critical illness insurance, fire and home insurance, and other services.

Many insurance companies are operating in Malaysia, offering a broad range of products and services tailored to Malaysians' needs from all walks of life. A total of 44 companies have been listed as the Licensed Insurance Companies and Takaful Operators list (Bao et al., 2018). Choosing which insurance company to invest in can be challenging as the client will be entering into a long-term investment. The survey conducted by YouGov in April 2019 finds that 45% of social media users in Malaysia use it as a platform to research

services, review, or find new products to buy. Through online communities like the one that exists on Twitter, reviewers can directly impact customer decision-making. Consequently, insurance companies should value how social media can influence clients' perception that affects their purchasing intention. Companies that adopt a strategic approach to social media use will benefit from those who do not use social media (Sum & Nordin, 2018).

The primary approach to be used in this project is data mining (DM), as it analyzes vast amounts of raw data into useful information and uses classification data mining techniques to mine big data. Sentiment analysis (SA) was used to computationally identify and categorizing opinions expressed by the user, to determine whether the user's attitude towards a particular insurance company is positive, negative, or neutral to help insurance companies define policies and offer better services. There are two primary techniques used in SA, which are Natural Language Processing (NLP) based and Machine Learning (ML) based (Yaakub, Latiffi, & Zaabar, 2019). NLP is a field of application and research that studies how computers are used to comprehend and manipulate text or natural language for specific purposes. NLP helps computer systems learn and manipulate natural languages to perform required tasks (Yaakub et al., 2019). ML is another approach used in sentiment analysis, in which the Support Vector Machine algorithm was used for classification tasks. More importantly, it provides higher accuracy and maintains the accuracy of a large proportion of data.

Thus, this study aims to visualize the performance of leading insurance companies in Malaysia through Twitter Sentiment Analysis using the Support Vector Machine (SVM) algorithm. The visualized results could be used in maximizing customers' satisfaction and to ensure retention. The Support Vector Machine algorithm was chosen as it is proven to produce a high learning quality in terms of accuracy. In this way, insurers will be able to proactively target new potential markets and resolve customer issues more effectively.

1.2 Problem Statement

As numerous Malaysian companies provide insurance, the customers faced difficulties choosing and evaluating each insurance company in purchasing decisions. According to Sum and Nordin (2018), the complexities of insurance purchasing issues include evaluating financial needs and choosing an insurance package from which established companies were confusing to customers. To make a wise decision, they need to allocate an adequate amount of time to evaluate each insurance company has to offer. According to the survey conducted towards insurance policyholders, 96.8% out of 62 respondents who owned insurance policy face difficulties in choosing the insurance companies to purchase an insurance policy. Most of the respondents also agreed to refer to online reviews before purchasing an insurance policy. Despite Google My Business (GMB) is a tool that provides online reviews of an insurance company, yet most of the published reviews are not as reliable as any of the bad reviews can be deleted, or there might be shill reviews in which shills might be paid or the company's employees (Deng & Wang, 2016).

Today, insurance companies are operating in a challenging and borderless business world, focusing on improving service to secure increased profits, and reduced costs are vital. In the survey conducted, 86.4% of respondents expected to have good customer service. Nevertheless, many insurance companies overlook the importance of serving their customers well. As a result, customers experience their interactions disappointed and dissatisfied, reducing the insurance industry throughput and performance drastically (Nuruzzaman & Hussain, 2018). Research has shown that nearly 75% of customers have experienced poor customer service from service providers (Nuruzzaman et al., 2018). Nowadays, many Insurance Companies have incorporated Customer Relationship Management (CRM) platforms that support social media. However, these platforms do not integrate sentiment analysis to extract essential information in their CRM. There is no effort done to organize the data and mold it into a purposeful record. Signs of dissatisfaction and chances for business expansion are frequently missed entirely.

Finding clients as an insurance agent can be challenging. The top three insurance industries are AIA, Great Eastern, and Prudential due to their long operation period in the country (Muhamat, Karim, Mainal, Alwi, & Jaafar, 2018). Although these top three companies have a significant number of insurance agents between 15,000 to 19,000 nationwide, most insurance agents have experienced the same struggle in expanding their businesses (Muhamat et al., 2018). Social media are used as a direct marketing tool these days, but a consistent commitment of time is required as it is harder than generic marketing to become successful. Thus, SA is significant as it gives insight into the target audience and allows the insurance agents to conduct competitive research by evaluating their competitor's strengths and weaknesses.

Based on the problems discussed, this study proposed implementing 'Sentiment Analysis of Malaysian Insurance Companies (SAMIC): A Visualization using Support Vector Machine Algorithm'. The tweets extracted were analyzed as it is one of the metrics that will advance insurance companies' online presence. Next, a new machine learning model is built based on the datasets in English and Bahasa Malaysia, in which bilingual sentiment analysis can be performed.

1.3 Project Objectives

The three primary objectives of this project are:

1. To design a web application system that can visualize the best insurance company in Malaysia.
2. To develop the designed system through Twitter sentiment analysis using the Support Vector Machine algorithm.
3. To test the functionality and usability of the system.

1.4 Project Scope

Target Audience

The targeted users are three of Malaysia's biggest Insurance Companies: AIA, Great Eastern, and Prudential.

Features of the system

The system is a web-based application written in python. The scope covers Malaysia's top insurance company's visualization using the support vector machine algorithm as the classifier model. The data is obtained from a post made on the Twitter social platform in bilingual languages, English and Bahasa Malaysia, using Twitter scraping tools. Besides, to prepare the data, data pre-processing steps, including data cleaning, data transformation, data reduction, and data discretization, need to be done. Term frequency-inverse document frequency (TF-IDF) is used in information retrieval and text mining.

1.5 Significance of the Study

The significance of this project is:

1. This system enables insurance company service a policyholder better through customer service. Some of the tweets in the analysis may be related to claims and based on the analysis's content. This information could be used to resolve potential concerns or questions related to the claims process. The questions may also be related to policy provisions or information. Detecting and proactively addressing these problems allows the insurer to provide exceptional customer service to their policyholders.

2. This system facilitates a deeper understanding of customer sentiment about the company. Customer reaction to company advertising campaigns can be evaluated to enable marketing programs more effective.

3. This project would also enable the insurance company to obtain competitive intelligence from several different perspectives, potentially providing insight into competitive issues.

1.6 Thesis Outline

This chapter discussed the project background, problems, objectives, scopes, and project significance. Social media has become an integral part of our daily lives and has become a valuable source of resource for insurance companies. Customers' sentiment towards the insurance company's services provides valuable insights and helps organizations formulate effective business strategies. Hence, this system is proposed to analyze customers' feedback and visualized the data obtained.

CHAPTER 2

LITERATURE REVIEW

The concepts and approaches used by papers, journals, and articles have been explored and compared to complete this study in a detailed manner. Several subjects that contribute to the supporting evidence of the project were highlighted in order to validate this research. Figure 2.1 visually represents the outline of this chapter.

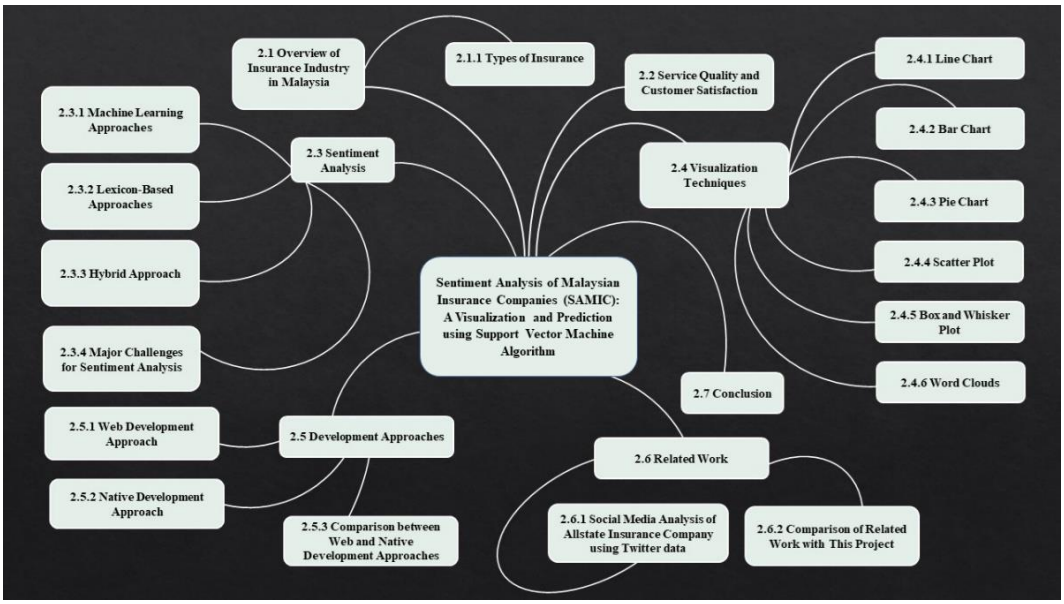


Figure 2.1 The Outline of Chapter 2

2.1 Overview of the Insurance Industry in Malaysia

The insurance companies in Malaysia are under the supervision of Bank Negara Malaysia (BNM). BNM points out that Malaysia's insurance industry has undergone one of its most essential transformations since the Malaysian insurance regulatory system overhauled in the early 1990s (BNM, 2016). The insurance industry was struggling with intense competition. Consequently, it has become vital for insurance companies, whether local or international, to

be competitive and profitable in this field (BNM, 2016). Any risk that can be quantified can potentially be insured. An insurance policy will specify in detail which of the policy covers risks.

There were few selections of insurance types offered by the insurance companies, and each plan has a different package or premium. Nevertheless, the plan chosen depends on the policyholders' situation, age, lifestyle, and the type of insurance policy needed, such as retirement, health, or education purposes (Mcmaken, 2020). Life insurance, general insurance, and medical insurance are common types of insurance business in Malaysia.

Life insurance helps to cover risks such as early death, sickness, and permanent injury. It pays out the insured or beneficiary a certain amount of money as compensation if anything unfortunate happens to the insured (Smith & Hayhoe, 2009). Additionally, the insurance provider will also pay the treatment costs based on the insurance plan and provide financial aid to the policyholders.

General insurance comprises householder insurance, motor insurance, and personal accident insurance (Zulkifly, Kasim, & Bidin, 2019). In the context of householder insurance, the policyholders will receive money if a natural disaster, fire destroy their personal property or house, and other cases included in the insurance plan, as well as the protection of policyholders against claims made by third parties in certain related events (Smith et al., 2009). If the policyholders' motor vehicle is involved in an incident, they may be claimed for damage or loss by means of motor insurance. According to the insurance plan purchased, personal accident insurance claims were obtainable when the policyholder became disabled due to an accident or passed away.

Medical and health insurance provides emergency coverage for hospitalization, recovery, and treatment (Zulkifly et al., 2019). Having health insurance is necessary to receive timely medical care and improve health and lives (Bovbjerg & Hadley, 2007). Besides, if the policyholder is admitted to

the hospital, the insurance company will pay medical expenses, including hospital fees, rooms, and facilities.

Insurance companies are classified into two categories, which are Islamic insurance, recognized as Takaful Insurance, and Conventional Insurance (Muhammad, Karim, & Jamal, 2012). In the case of Takaful Insurance, the company is subject to Syariah laws and regulations. It comprises three concepts, including mutual obligation, cooperation with each other, and protection from all kinds of difficulties as the premium is based on the tabarru' concept of contribution and donation (Hamid, Osman, & Nordin, 2009). Conversely, conventional insurance does not comply with the provisions of Syariah because it includes other non-conclusive elements in the insurance contract, such as uncertainty and interest (Muhammad et al., 2012). Although there are two different insurance categories, all insurance policies have the same purpose by helping to mitigate and reduce the financial burden on policyholders against loss of property and livelihoods. The comparison between Takaful and Conventional Insurance are shown in Table 2.1.

Table 2.1 Comparison between Takaful and Conventional Insurance

	Takaful	Conventional insurance
Contract	A combination of the Tabaru' contract (donation) and profitsharing contract	Exchange of sales and purchases between the insurer and the insured
Policyholders responsibility	Policyholders contribute to the scheme and mutually guarantee one another under the scheme	Policyholders shall pay premiums to the insurer
Liability of the insurer	The takaful operator serves as the administrator of the scheme and pays the takaful benefits using the takaful funds.	The insurer is liable to pay insurance benefits as agreed by the shareholders of the assets
Investments of funds	The assets of the takaful funds will be invested in the Shari'ah-compliant instrument	There is no limitation other than that imposed on prudential reasons

(Source: Muhammad et al., 2012)

According to the latest data released by YouGov BrandIndex (2019) on their website, the leading insurance companies in Malaysia are Prudential, AIA, Great Eastern Life, Etiqa, and Allianz. All of the insurance companies in YouGov BrandIndex have been ranked based on the index score, which is the measurement of the overall company's performance, including the average impression, customer satisfaction, service quality, recommendation, and reputation. The index rankings chart displayed the highest average index scores for 12 months from July 1, 2018 to June 30, 2019. All scores are based on a representative sample of the general population over 18 years old. Prudential, AIA, and Great Eastern Life were the top 3 companies that achieved the highest index score, which is why these companies are chosen as the main scope of this project. Besides, these companies also provide a Takaful insurance policy. Figure 2.2 shows the index rankings of insurance companies in Malaysia.

Top Index Rankings

Rank	Brand	2019 Score
1	Prudential	26.5
2	AIA	21.3
3	Great Eastern Life	20.1
4	Etiqa	17.6
5	Allianz	12.3

Figure 2.2 The Index Ranking of Malaysia Insurance Companies
(*Source*: YouGov BrandIndex, 2019)

The next section will discuss in-depth some evaluation criteria of the index score as well as why insurance companies can use the outcomes of this project to maximize their business performance.

2.2 Service Quality and Customer Satisfaction

The influence of service quality on customer satisfaction has been discussed across various sectors providing services (Kumar, Shah, & Manab, 2018). Service quality and customer satisfaction are the primary concern of service organizations in the increasingly intensified competition for customers in today's customer-centric era (Wang, Chi, & Yang, 2004). The quality of service has a significant impact on customer satisfaction. Both the service and customer orientation are inseparable and must be handled simultaneously.

Bitner and Hubbert (1994) described service quality as the overall customers' perception of the organization's superiority or inferiority and the services offered. From a managerial viewpoint, service quality helps managers understand the importance of making sure that efforts are being made to get the service right for the first time and to meet or surpass customer expectations while providing the service (McDougall & Levesque, 2000). The quality of service is essential. If an insurance company produces a quality insurance plan without delivering excellent services, it does not guarantee that the organization can keep a good reputation (Kumar et al., 2018). In addition, considering the quality of service will offer a good corporate image to the insurance company.

Customer satisfaction refers to the customer's response towards service delivery. In general, customer satisfaction occurred when the customer compared the service's actual performance with their supposition of the service. The differences would result in the form of disconfirmation, either positive, negative, or zero disconfirmation (Angelova, Zekiri, & Jusuf, 2011). Customer satisfaction is inextricably linked to the quality of service analysis. The confluence between the customers' expectations and the customer's reception has resulted in satisfaction (Kumar et al., 2018). Good and positive feedback on the industry's performance from satisfied customers helped develop a good reputation for the industry.

Therefore, in order to sustain and improve the reputation of the social insurance sector, the company should continue to focus on improving customer satisfaction. It is crucial to understand the association between service quality and customer satisfaction to preserve a good reputation for any organization. Sentiment analysis makes it much easier to obtain valuable information about the service quality and customer satisfaction of insurance companies without having to read through every piece of content, as insurance companies are provided with an organized, systematic, and subtle way of monitoring social media reviews (Roosevelt & Mosley, 2012). Critical areas of interest can be established using the power of analytics and optimizing the time spent by concentrating the analysis on these areas. As such, it is acceptable to include the study of sentiment analysis of insurance posts from a Malaysian perspective. Sentiment analysis will be discussed extensively in the next section.

2.3 Sentiment Analysis

Sentiment analysis (SA) is an area of study within natural language processing (NLP) that recognizes and captures opinions within the text. Apart from evaluating opinions, the SA system also evaluates the emotions of the users, such as polarity, to identify if the opinion expressed is positive, negative, or neutral, and the holder of the opinion identifies as the person or group providing the opinion (Ahmad, Aftab, Muhammad, & Ahmad, 2017). Public sentiment about a product, services, or other topics can be beneficial for different commercial applications. The unstructured information collected by SA will be converted automatically into structured data.

The opinion may be divided into three major types, which are regular opinions, comparative opinions, and suggestive opinions that indicate a single entity or multiple entities (Shayaa, Jaafar, Bahri, Sulaiman, Wai, Chung, Piprani, & Al-Garadi, 2018). The regular opinion referred to a single entity and used primarily to describe positive or negative perspectives for an item in particular. Apart from that, comparative opinions compared or display a correlation

between more than one entity and are primarily used for strategic intelligence because they help to elucidate the relationship between multiple entities (Shayaa et al., 2018). The next subsection term frequency-inverse document frequency used in the data preparation and a few SA approaches, such as machine learning, lexicon-based approach, and hybrid approach, will be discussed.

2.3.1 Term Frequency Inverse Document Frequency (TF-IDF)

As part of the feature extraction technique, the term frequency-inverse document frequency (TF-IDF) is used and measured before it is passed to the machine learning model. TF-IDF is a numerical statistic widely used in text mining and information retrieval to shows the significance of keywords to specific documents or ranks the frequency of terms used to classify or categorize particular documents (Qaiser & Ali, 2018). Term frequency calculates the number of times a word appears in a document, divided by the total number of words in that document. On the other hand, the logarithm of the number of documents in the corpus, divided by the number of documents in which the particular word appears, is computed as inverse document frequency (Srividya & Sowyjanya, 2019). Thus, TF provides a term's frequency, and IDF measures the importance of a term. The algorithm for conducting the TF-IDF is illustrated in Figure 2.3.

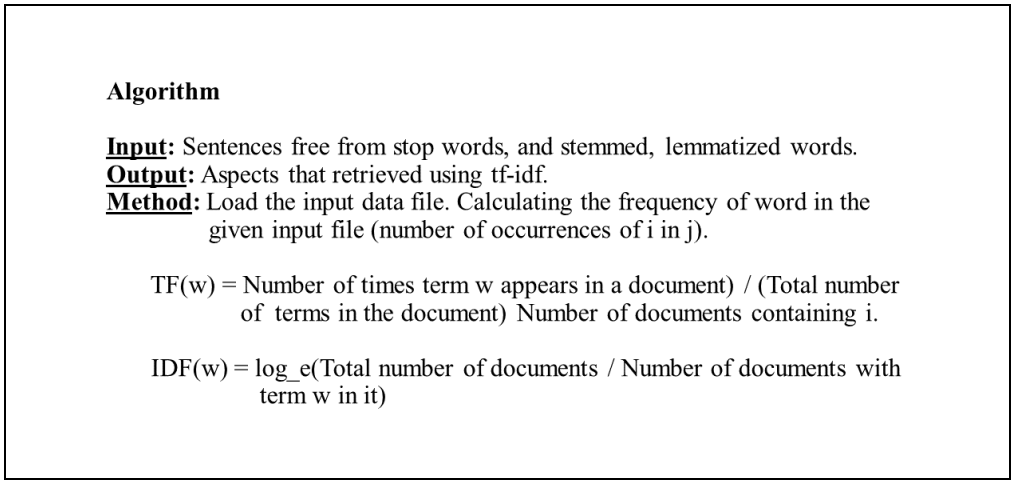


Figure 2.3 TF-IDF Algorithm
(*Source*: Qaiser & Ali, 2018)

2.3.2 Machine Learning Approaches

The classification technique to classify text into classes used in Machine Learning-based approach is including unsupervised learning and supervised learning (Salman, Khan, & Iqbal, 2019). Unsupervised learning does not have the right goals and depends on clustering. The model works on its own to discover information and deals mainly with unlabeled data. Complex processing tasks are typically used by unsupervised learning (Shayaa et al., 2018). In Supervised learning, the labels are provided to the model. These labeled datasets are trained to produce relevant outputs. The selection and extraction of a particular set of features used to detect sentiment determine this learning method's success. The machine learning algorithms applicable to SA are primarily part of the supervised classification (Sailunaz & Alhajj, 2019). In machine learning techniques, data sets needed include a training set or test set. Training sets are used to train the classifier model. Once a supervised classification technique has been selected, features in sentiment classification inspected will indicate how the documents are represented. The only drawback of a supervised machine learning approach is the necessity to provide an instance of sufficiently rigorous training to make the algorithm efficient and highly credible to classify each instance of the data (Ahmad et al., 2017). The next subsection discussed several types of machine learning algorithms.

2.3.2.1 Artificial Neural Network

The Artificial Neural Network (ANN) is a mathematical modeling approach intended to replicate how the human brain learns and consists of input, output layers, and in most cases, a hidden layer (Tang, 2017). The ANN is an outstanding algorithm to find patterns that are too complicated as the artificial neurons are competent in carrying out parallel computations and vast analogous to process information and representation of knowledge (Thangaraj & Sivakami, 2018). The architecture of the ANN layer, as depicted in Figure 2.4, enables the algorithm to adapt its inner structure with practical use. Thus, the algorithm is more applicable to problems that are nonlinear in nature.

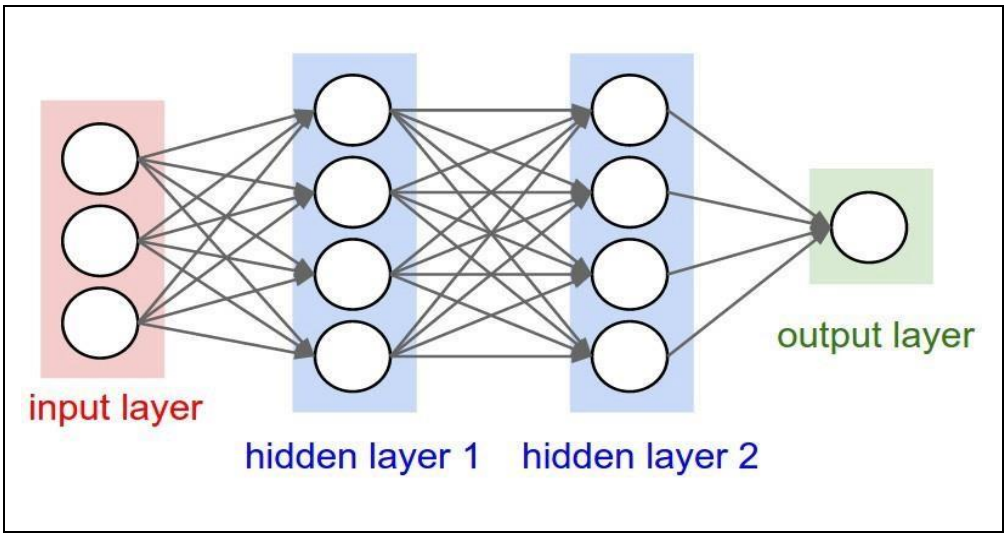


Figure 2.4 The Architecture of ANN Layer
(*Source*: Zerium, 2018)

2.3.2.2 Random Forest

The Random Forest (RF) is a regression and classification approach that employs the ensemble of decision trees' rapid growth during the training process (Alsolamy, Siddiqui, & Khan, 2019). Ensemble learning, also known as bagging, is a method that can build multiple classifiers and also produce aggregated results. Define random forest is a classifier composed of structured tree classifiers{h(x, Oy), y=1,...} with (Oy) is individually identical distributed random vectors, and that each tree casts a voting unit for the class at input x (Tang, 2017).

Figure 2.5 shows that the algorithm creates trees on the subset of data and later combines all trees' output. Consequently, overfitting problems and the variance in decision trees are reduced and therefore improves the accuracy. The algorithm is more complicated than the decision tree as it requires more computational power and resources (Reis, Baron, & Shahaf, 2018). Moreover, the algorithm also has a more extended training period, as many trees are generated.

The classifier carries out classification by ensembles from random partitions (CERP) that is explicitly structured to handle high dimensional data sets that

can predict the random portion of the entire collection of predictions, where multiple classifiers are combined to improve prediction as compared to the original classifier (Shayaa et al., 2018). The ensemble-based classifier is developed with logistic regression trees (LR-T CERP) and classification trees (C-T CERP). C-T CERP less dependent on the threshold than LRT CERP, but the results of both these classifiers showed high accuracy.

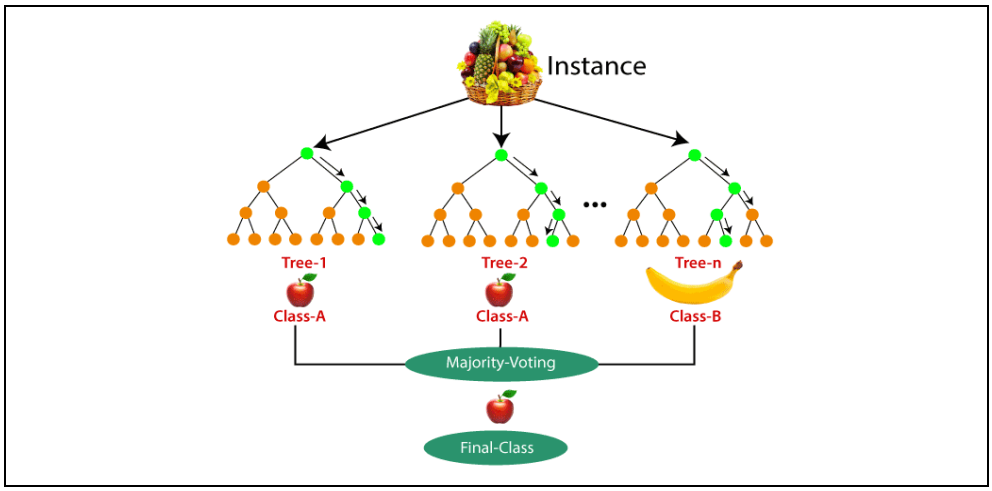


Figure 2.5 Random Forest Structure
(*Source*: Koehrsen, 2017)

2.3.2.3 Support Vector Machine

The support vector machine (SVM) is a supervised machine learning model that used associated learning algorithms to analyze data used for classification analysis (Gupta, Negi, Vishwakarma, Rawat, & Badhani, 2017). The algorithm produces the most efficient result in traditional text categorization by locating the best possible distinctions between positive and negative training samples (Dey & Bandyopadhyay, 2016). There are numerous extensions available for the algorithm to enhance its effectiveness and more adaptable to real-world requirements. Soft Margin Classifier is one of the SVM extensions that classifies the majority of the data. It excludes any outliers and noisy data as the data is occasionally linearly noticeable for multi-dimensional problems and linearly segregated (Shayaa et al., 2018). The non-linear classifier is a further extension of the SVM in which kernel is used to maximize hyperplanes' margin. SVM and its extensions are used for binary classification

functions. However, for the multiclass problems, the Multi-Class SVM extension is used with labels built for objects that are drawn from a finite collection of multiple elements (Shayaa et al., 2018). SVM can manage linear separation at high dimensional non-linear input by using an appropriate kernel such as various kernel functions.

2.3.2.4 Genetic Algorithm

Genetic Algorithm (GA) is a search algorithm based on the developmental concepts of natural selection and genetics (Samah, Badarudin, Odzaly, Ismail, Nasarudin, Tahar, & Khairuddin, 2019). A few genetic operators are used in GA, such as crossover, recombination, mutation, and selection. These operators form an integral component of the GA for the problem-solving strategies (Sailunaz & Alhajj, 2019). Variants of genetic algorithms were applied to a diverse range of optimization tasks, such as graph coloring and pattern recognition. The GA is preferable over traditional optimization techniques as it performs better than other traditional methods in most optimization problems (Samah et al., 2019). Furthermore, the population traverses the search space in several directions simultaneously, making it an excellent way of parallelizing the algorithms in deployment. Nevertheless, new population selection criteria need to be carefully done as there is no guarantee of finding global maxima, making it difficult to converge the algorithm (Ahmad et al., 2017).

2.3.2.5 Naïve Bayes

Naïve Bayes (NB) is a supervised classifier used to measure the data likelihood, whether positive or negative, by entailing a minimum collection of data for training, which is used to foresee the parameters required for classification activities (Shayaa et al., 2018). The classifier model is useful for large amounts of data and is based on the Bayes theorem (Ahmad et al., 2017). The probabilities p of two events, A and B, represented as P(A) and P(B), and

the conditional probability of event A by event B is represented as P(A | B). Thus, the Bayes' formula equation is in Eq. (1).

$$P(A\,|B) = \frac{P\,(B|A).\,P(A)}{P(B)} \tag{1}$$

The classifier has been employed in numerous supervised machine learning applications as it is speedy compared to the other complicated algorithms and performs well with high-dimensional data such as text classification and spam detection. However, the algorithm might not be valid all the time because of the assumptions of attributes being independent, as there is no correlation between features (Shayaa et al., 2018).

2.3.2.6 K Nearest Neighbour

The K Nearest Neighbour (KNN) algorithm is easy to deploy as it is an algorithm that performs an instance-based classification, learning by operating through a non-parametric process to store both instances and inputs (Tang, 2017). New inputs are classified by applying similarity measures such as the Euclidean, Minkowski, and Manhattan distance (Shayaa et al., 2018). The KNN algorithm presumes that similar things exist in close proximity by calculating the distance between points on a graph, as illustrated in Figure 2.6, and hinges on the assumption of it to be useful.

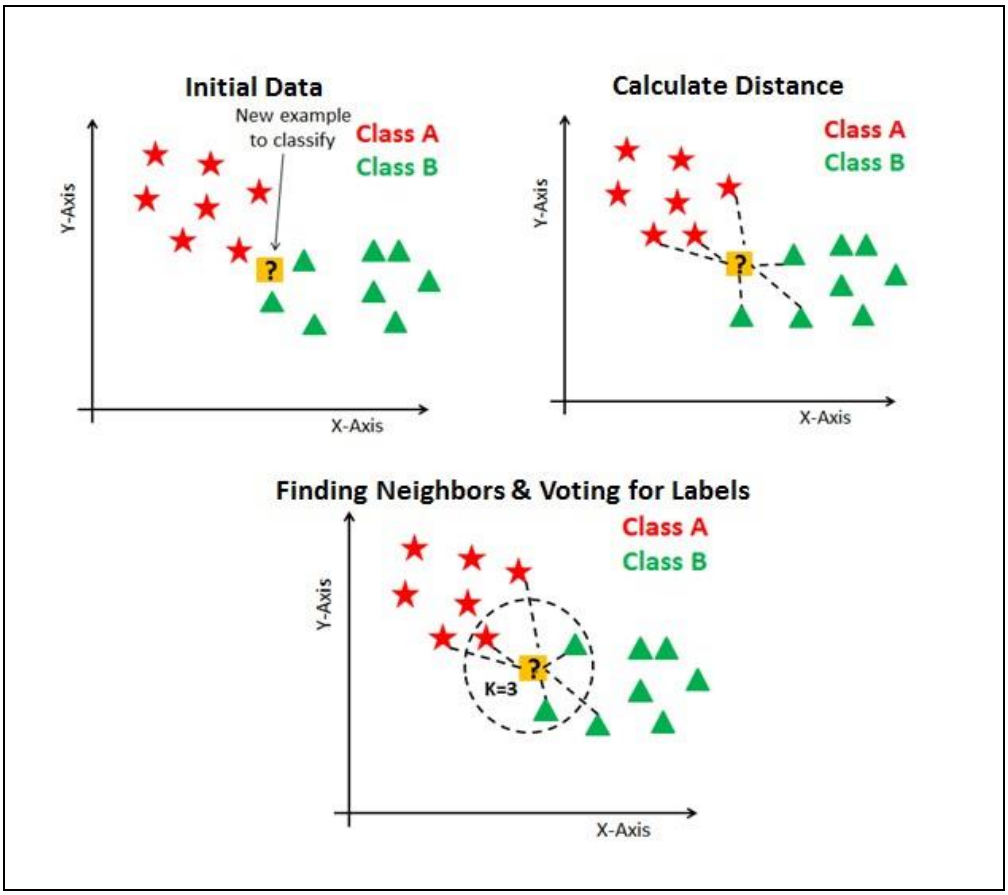


Figure 2.6 Typical Concepts of How the KNN Algorithm Operates
(*Source*: Navlani, 2018)

2.3.2.7 Comparison between Machine Learning Algorithms

The advantages and disadvantages of machine learning algorithms listed in Table 2.2 are identified to select the most suitable algorithm for analyzing the data collected.

Table 2.2 Summary of Comparison between Machine Learning Algorithms

Algorithm	**Source**	**Advantages**	**Disadvantages**
ANN	(Tang, 2017), (Thangaraj & Sivakami, 2018), (Zerium, 2018)	• Ability to complete tasks that cannot be done by a linear system • ANN will continue to function if the neural network aspect fails due to its parallel nature, making it widely used in many sentiment applications	The algorithm needs training data to function, and processing time for large neural networks are longer
RF	(Alsolamy, Siddiqui, & Khan, 2019), (Tang, 2017), (Reis, Baron, & Shahaf, 2018),	• Future selection does not need to be applied as Random Forest uses many features. • It runs efficiently on large data sets and produces a highly accurate model	Require more computational power, resources, and a more extended training period as there are many trees generated to produce accurate ensembles
SVM	(Gupta, Negi, Vishwakarma, Rawat, & Badhani, 2017), (Dey & Bandyopadhyay , 2016)	• The kernel-based framework is compelling and easily adaptable. • Functions very well in practice, irrespective of the size of the samples	If the number of features is much greater than the number of samples, the algorithm is likely to give poor performance
GA	(Samah, Badarudin, Odzaly, Ismail, Nasarudin, Tahar, & Khairuddin, 2019), (Sailunaz & Alhajj, 2019)	• Able to handle complex problems, including stationary or nonstationary, and applicable to global scenarios	Slow performance, no convergence guarantee, and can converge prematurely
NB	(Shayaa et al., 2018), (Ahmad et al., 2017)	• Fast performance and high computational complexity. • Usually entails minimal data sets for training to estimate the parameters needed for classification purposes	The model assumes the dependency of attributes
KNN	(Tang, 2017), (Navlani, 2018)	A simple and robust algorithm. The algorithm requires no training time.	As the volume of data increases, the algorithm becomes significantly slower makes it an impractical choice in environments where predictions need to be done rapidly

Based on the evaluation of various sentiment analysis algorithms, it is concluded that SVM is the most effective algorithm to be used for this project. This is because the algorithm runs efficiently on large data sets and offers better results compared to other traditional classification algorithms (Luthfi & Lhaksamana, 2020). Past research involving SA usage in social media data has shown that the SVM algorithm provides the most accurate text classification problems. It is achieved by constructing a hyperplane for the nearest trained examples with maximum Euclidean distance. The hyperplane of SVM is fully resolved by a comparatively limited subset of the trained data sets treated as support vectors (Bhavitha, Rodrigues, & Chipunkar, 2017). Besides, the outstanding generalization capability of SVM, together with its optimal solution and discriminative power, makes it widely used in the area of text mining. The algorithm is also memory efficient due to its advantage of kernel mapping to high-dimensional feature spaces. (Wang, Zhou, Jin, Liu & Lu, 2018). Other approaches that may be used in SA but will not be used in this project are discussed in the next section.

2.3.2 Lexicon-Based Approaches

The lexicon-based approach uses a sentiment dictionary that contains terms of opinion and matches them with the data to identify the polarity between opinions. Sentiment scores an opinion word defined how the dictionary's words are assigned as positive, negative, and neutral. In general, a bag of words represents a piece of text in the lexicon-based approaches. Following this message representation, sentiment values in the dictionary were applied to all positive and negative words or phrases in the message (Shayaa et al., 2018). The following are two sub-classifications for this method.

2.3.2.1 Dictionary-based

The dictionary-based approach used the compiled and annotated terms manually (Sailunaz et al., 2019). Firstly, the terms of opinion from a review text is discovered. Next, the synonyms and antonyms from the dictionary are

scanned. The commonly used dictionaries are WordNet and SentiWordNet (Ahmad et al., 2017). The inability to deal with context-specific orientations and domain is the main drawback of this approach.

2.3.2.2 Corpus-based

The Corpus-based approach used a collection of structured text to help with automatic processing, in which repeated words found in a document are considered co-occurring terms (Alsolamy, Siddiqui & Khan, 2019). It helps find terms of opinion in a context-specific orientation where the approach started with a set of opinion words. Next, other opinion words in a vast corpus are observed using statistical techniques such as Latent Semantic Analysis (LSA) or semantic techniques such as SentiWordNet 3.0, which is the most effective dictionary (Tang, 2017). SentiWordNet 3.0 is a lexical resource publicly available composed of 'synsets' each is correlated with a positive and negative numerical score ranging from 0 to 1, where the score was allocated to the WordNet automatically (Tang, 2017). The latter applicable approach for SA will be explained in the following section.

2.3.3 Hybrid Approach

The Hybrid Approach uses both machine learning approaches and dictionary-based. A lexicon-based approach for sentiment scoring is employed by following a classifier's training with assigned polarity to the entities in the latest found reviews (Shayaa et al., 2018). The hybrid approach is widely used as it incorporates the best of both approaches, a high precision from a robust supervised learning algorithm and the reliability from the lexicon-based approach. A few SA challenges, such as technical and non-technical challenges, have been described in the next subsection.

2.3.4 Major Challenges for Sentiment Analysis

In this section, the technical challenges, which are the challenges associated with creating technical sentiment and non-technical challenges associated with the implementation of SA of the big data, have been addressed.

2.3.4.1 Technical Challenges

1) Heterogeneous Properties of Big Data

   A vast amount of data in diverse and heterogeneous formats are significant features of big data and are generated due to data collection from various mediums used by web communities' users (Ghaleb & Vijendran, 2017). The sentiment classifier is expected to work adequately with such heterogeneous properties of big data from the different data sources, unlike traditional sentiment classifiers, which deal with data from a single source, such as the company's online review (Tang, 2017).

2) Analyzing Uncertain, Incomplete, and Sparse Data

   Data sparsity is among the characteristics of big data in which the data contains much noise as a result of the extensive use of acronyms and misspellings. This occurrence may affect the precision of the sentiment classification. Moreover, data volume restricts the filtering of the relevant data from non-relevant data, which may undermine the sentiment analysis results (Neves-silva, Gamito, Pina, & Campos, 2016). Thus, sentiment classifiers must accurately predict missing information so that more detailed data can be provided and a classifier that can classify the sentiment polarity of a post.

3) Semantic Associations in Multi-Sources of Data Fusion

   Analyzing semantic associations of activity from data sources such as Twitter, Instagram, Facebook, and YouTube may provide greater insight and understanding for the overall picture of sentiment (Shayaa et al., 2018).

Creating semantic associations among text, image, and video data will be substantially improved after an event analysis from different sources (Alaoui & Gahi, 2019). However, it may be challenging to implement a semantic association based on models for filling the semantic gap between heterogeneous data sources.

2.3.4.2 Non-technical Challenges

1) Does Sentiment Analysis of Social Media Data Help in Assembling Business Strategies

   Big data has become a key component for businesses to get insights into the customers' opinions about their products or services. Massive user-generated data provided by social media brings exceptional processing opportunities for many applications involving a thorough understanding of the public sentiment about products, people, and events (Drus & Khalid, 2019). SA can be applied to clients' captured opinions to boost the standard of their products or services. However, more research is needed to grasp factors working along with sentiment factors to acquire an exceptional understanding of the reviews, such as the competitors' information, consumer confidence index, and country economic growth that may influence the conclusive decisions (Shayaa et al., 2018). Therefore, future research needs to incorporate SA approaches into these factors to generate useful strategies for a fully automated decision-making system.

2) The Influence of the Post

   There is no substantial influence on the direct number of followers (Tang, 2017). For example, a user X that has many followers made a negative comment on brand A in the online reviews section, and a user Y with only several followers made a positive comment on that same brand. Although the number of positive and negative comments of brand A is the same, the negative comment will have a more significant influence since user X has more followers that can be swayed by the comment stated as compared to

user Y. Thus, future studies should take into account the impact of sentimental polarity within the linked user networks. A precise measurement of sentiment can be provided by influence measurement and integrated SA in which both the polarity of the post and its ability to influence the opinions of consumers' on a broader scope will be considered (Shayaa et al., 2018). In the next section, several visualization techniques will be addressed.

2.4 Visualization Techniques

Data visualization is the practice of presenting the data in a graphical or pictorial form. Top management decision-makers are able to see analytics as visually portrayed, making it possible to understand complex concepts and identify new patterns or structures. Visualization-based data discovery techniques enable business owners to create completely different data sources in order to invent custom analytical views (Samuel & Anthonia, 2016). Throughout the years, many visualization techniques have been developed to represent and analyze large-scale information, and these methods include usability, interactivity, and interface features, making it easy to deploy. Data visualization techniques will be explained in the next subsection.

2.4.1 Line Chart

As illustrated in Figure 2.7, the line chart shows the relationship between variables on the chart and is frequently used to compare many items simultaneously. A few standard icons and symbols are used to represent the data points in the line chart. The stacking lines in a line chart demonstrate a contrast between multiple variable patterns and are ideal for displaying a variable shift (Samuel et al., 2016). A line commonly joins data points, and the line chart is simply an expansion of a scatter plot (Womack, 2015).

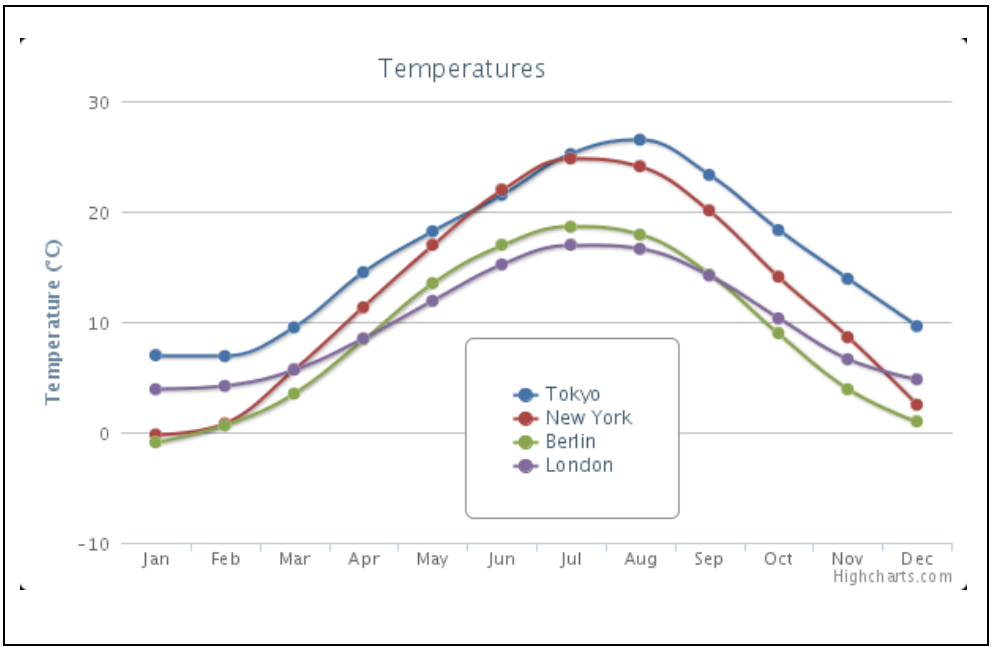


Figure 2.7 An Example of Line Chart
(*Source*: Highcharts, 2020)

2.4.2 Bar Chart

The bar chart is also referred to as the bar graph, which distinguishes objects from different groups. The horizontal bars and verticals bars are two types of bar charts in which bars illustrate different group values (Samuel et al., 2016). An example of a bar chart can be observed in Figure 2.8.

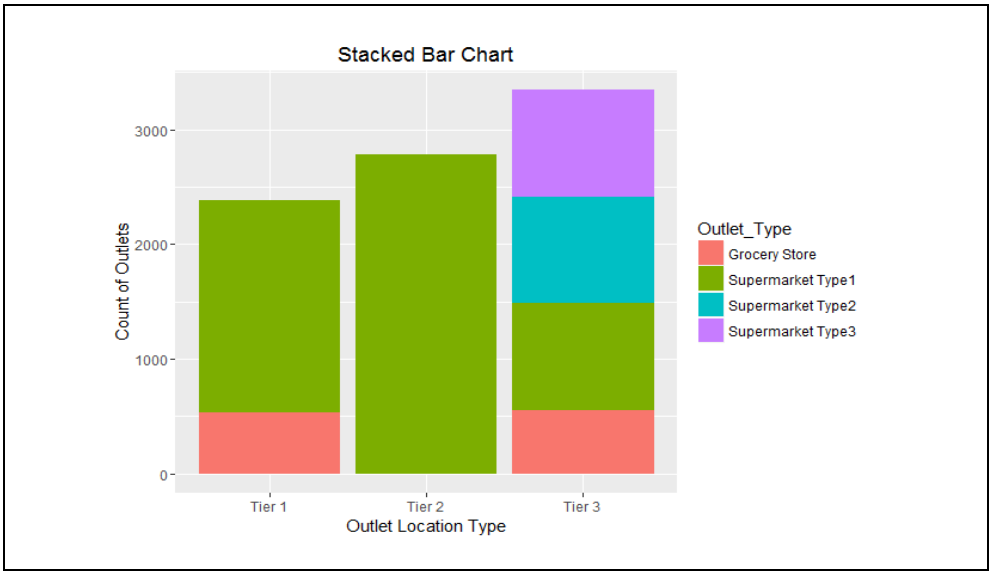


Figure 2.8 A Stacked Bar Chart Example
(*Source*: Vidhya, 2017)

The bar chart is frequently used to visualize discrete data and a single data series in which the connected data points are clustered together. However, if there are many values to be presented, it could be challenging to distinguish between the bars.

2.4.3 Pie Chart

A pie chart, also referred to as a circle graph, displays statistics on information and data in a form that will be easy to interpret, called the 'pie-slice' form, and the different sizes of the slice signify the presence of an element (Khan & Khan, 2016). The most effective way to use the pie chart is to describe the data's content, a few elements, and the percentages (Samuel et al., 2016). This visualization technique functions well to compare a pie chart segment to the pie chart segments (Khan & Khan, 2016). Figure 2.9 provides an example of a Pie Chart.

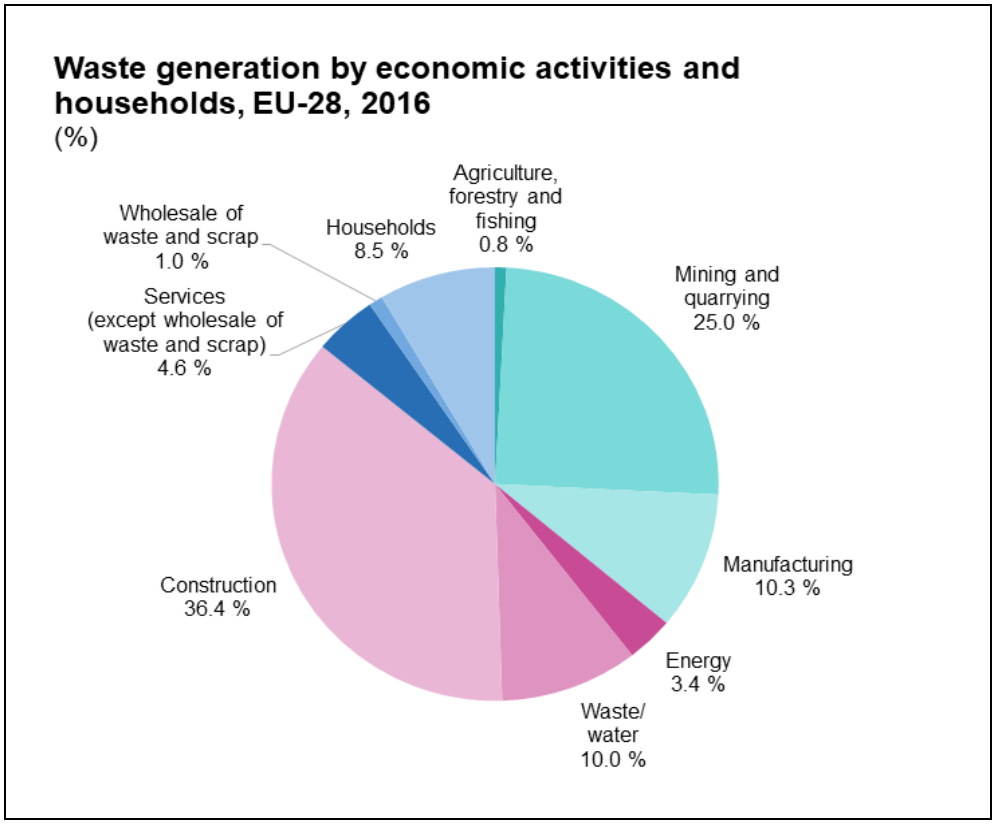


Figure 2.9 An Example of a Pie Chart
(*Source*: Eurostat, 2019)

2.4.4 Scatter Plot

As demonstrated in Figure 2.10, a scatter plot is also referred to as a scatter diagram, or a scatter graph can be defined as a 2-dimensional plot that illustrates the joint variation of two data objects (Samuel et al., 2016). The scatter plot provides data in the Cartesian coordinate as a graphical display showing the relationship between the two variables in which one is depicted as a vertical distance and the other as a horizontal distance. In the scatter plot, the observations are represented by each marker. The location of the marker typically indicates the value of observations. After plotting all the data on a scatter plot, it is possible to determine whether the data points are related.

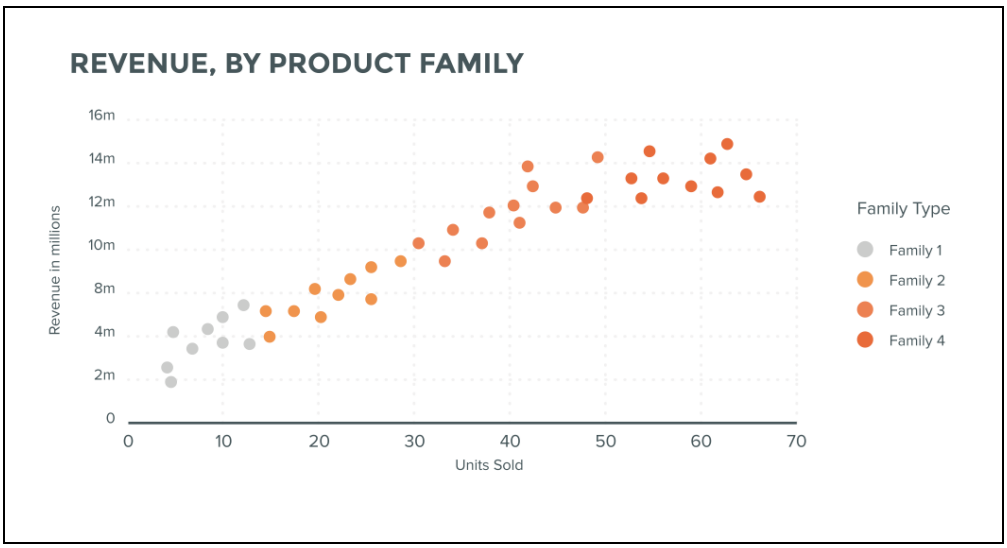


Figure 2.10 An Example of Scatter Plot
(*Source*: Ben, 2015)

2.4.5 Box and Whiskers Plots

The box-and-whisker plot can summarize the variance of a complex dataset (Womack, 2015). It uses median and interquartile ranges of statistical summaries that are reliable in the presence of outliers and skewness. This technique displays a complete range of data from the sample, gives information on the tails, illustrates the data shape, and is appropriate for quick side-by-side comparisons between groups (Nuzzo, 2015). Figure 2.11 shows an example of a box-and-whisker plot.

Figure 2.11 A Box and Whisker Plot Visualization
(*Source*: HBS, 2019)

2.4.6 Word Clouds

As illustrated in Figure 2.12, the word clouds visualization approach uses a set of words as input, with their corresponding correlated frequency value, and demonstrates the information graphically. It provides a straightforward and intuitive way of visually expressing the most common words in text documents. Usually, the text's most common words are shown as a weighted word list in a particular spatial structure layout such as sequential, random, and circular (Lohmann, Heimerl, Bopp, Burch, & Ertl, 2015). The font sizes of the words reflect their significance or frequency of occurrence. Whereas other visual characteristics, including location, color, and orientation, are often differed for artistic purposes or to visually express additional information (Lohmann et al., 2015). Nevertheless, the existing word cloud visualizations offer limited support in comparing the word and word frequency of different text documents (Nuzzo, 2015). To overcome this constraint, ConcentriCloud is implemented, an enhanced word cloud simulation that systematically merges and portrays words from various text documents.

Figure 2.12 A Word Cloud Visualization
(*Source*: Piatetsky, 2014)

2.5 Development Approaches

The complexities of developing interfaces and experiences for different application platforms are not a new issue and have been addressed in various works (Dossey, 2019). Interest has been fuelled by the growing number of different user interaction tools. There are many arguments regarding web applications and native applications, and each approach has its strengths and limitations. Thus, selecting the one that adequately addresses the requirements of a project is an important task. The next subsections will explain the types of development approaches, including web and native development, as well as the comparison between these approaches.

2.5.1 Web Development Approach

A web application is a web-based program designed to operate on devices, including smartphones, tablets, and laptops, through a web browser (Ali & Sari, 2015). Due to standardized technologies, mobile users can access the website on all devices in a similar manner by mobile browsers environment (Heitkötter, Hanschke & Majchrzak, 2012). This approach relies on browser support for mobile platforms and the standardization of the web technology environment. By using this web development approach, developers are applying the framework of a single website optimized for mobile devices.

Developing a web application is cost-saving, as it involves a single development cycle for all browsers. Since the web application resides on a server, it can be easily customized. Hence, developers do not need to force users to install and download any updates like in the native application through the application store (Heitkötter et al., 2012). Accessing a web application requires only a web browser. Hence, it is merely a multiple platforms solution that provides platform independence, which offers an opportunity to reach users in the broadest possible way since users can easily access the application from any device connected to the Internet. Besides, web technologies are considered most relevant to current trends in the software industry, whereby lifecycles are short, developments are rapid, and customer preferences are frequently evolving (Erkkilä, 2013).

Although web-based applications have capabilities in the development and delivery phase, several functional limitations and hardware features exist. The functional limitations of mobile browsers currently hamper web applications (Cardieri & Zaina, 2018). Advancements in a mobile device provide additional features such as Bluetooth, GPS, cameras, and telephony. Native applications are able to exploit these features, whereas browsers are typically incapable of providing these functionalities. Furthermore, it has been observed that there seems to be the main difference between these approaches in terms of performance as the web application relies primarily on the browser and the network connection (Erkkilä, 2013).

2.5.2 Native Development Approach

Ali et al. (2015) claimed that a native application is a device-based application that involves particular programming for a specific operating system (OS) only. For example, the particular application programming for iOS is for Apple devices only, and Android programming is only for Android devices. If several platforms need to be supported by native applications, the application must be developed independently. In developing native applications, developers use the frameworks and software development kit (SDK) to build an application

for one particular target platform (Heitkötter et al., 2012). The application is, therefore, tied to that specific environment.

There are several benefits that a native application can offer as much as a web-based application. Native applications are useful in a way that they can utilize the features of the native operating system, allowing applications to have a more responsive and appealing user interface as well as the ability to function without internet connectivity (Selvarajah, Craven, Massey, Crowe, Vedhara, & Raine-Fenning, 2013). It is because native applications typically have included a persistent information storage mechanism and business logic since data is stored in mobile devices. Therefore, this feature allows users to get instant access to information at their convenience. Moreover, the native application also provides faster performance as fewer data and graphics are being transmitted over the Internet.

Nevertheless, there are many difficulties in this approach. First of all, creating applications for mobile devices entails cross-platform complexities because of the intense rivalry in the mobile device industry, which has led to the proliferation of non-standard and diverse platforms on the market (Ali et al., 2015). The leading mobile operating systems are iOS, and Android requires different programming languages (Cardieri et al., 2018). As users may use their own preferred mobile device brand, it is difficult for organizations to provide an application that is accessible to those main operating systems. Besides, it is challenging for developers to comply with various and unknown hardware requirements, including input mode, screen size, random access memory (RAM), storage capacity, and processing power of different devices (Cardieri et al., 2018).

2.5.3 Comparison between Web and Native Development Approaches

The comparison between the web and native approaches are listed in Table 2.3.

Table 2.3 List of Comparison between Web and Native Approaches

	Web Approach	**Native Approach**
Performance	Varying performance as it relies on browser compatibility, network latency, and code complexity	Faster performance as fewer data and graphics transmitted over the Internet
Availability	Applications are accessible online, a network connection is necessary	Applications are accessible for offline use
User Interface	Responsive design for optimal viewing and interaction, regardless of the size or orientation of the device	Highly responsive and appealing user interface
Functionality	The functionality is restricted to the functionality provided by web browsers	Full access to the Application Programming Interface(API)
Development Time	Faster to develop than the native approach	Medium development time, but time-to-market can vary according to the platform
Application Store Distribution	It does not include the application store feature as it is hosted on the webserver. The applications are always up-to-date	Inclusion of the application store as the applications must be manually installed and updated for each device
Code Reusability	Code is reusable, but the browser compatibility must be considered	Code cannot be changed easily, and it is less likely to be reused

s

After analyzing several criteria and taking into account, all aspects of constraints such as time constraints and the scope for the project to be a fully functioning application, a web application approach is selected for this project. Many frameworks are available and easy to learn for web applications. Thus, it is more prudent to opt for a web application approach. Moreover, development time can be minimized, given that time is a major restriction for this project. The application to be built does not have to include any complex features or access to the devices' hardware, such as geolocation and camera. Considering that websites are more practical for the target audience, rather than worry about the platform's restrictions that constrain native applications such as cross-platform interfaces and capability across various devices, it is more manageable to deal with cross-browser compatibility. The web application only needs to be developed once with cross-browser support, and

it is useable on every device. To summarize, in comparison to a native application, a web-based application is able to outperform in terms of development time and complexity. The next section explores related work on the analysis of insurance Twitter data.

2.6 Related Work

Social media site usage has significantly increased, and businesses, including insurance companies, acknowledge it. Social media generate information that offers insights into businesses. In response, steps were taken to exploit the power of social media sites to make them successful. Insurance companies are investing substantial resources and time to establish and maintain a presence in social media. However, the task may be overwhelming for insurers to access and start making use of this information. One solution would be for a business to employ a group of people to monitor social media sites for content that may interest the business. It would require them to scan through the content to determine if the information is useful and identify the appropriate channels by which the information can be routed. As a result, the budget for social media activities will be expanded, and valuable information is possibly missing as it is impossible to analyze everything.

Research has been carried out to resolve these issues by using Twitter data to analyze insurance companies' performance. It is because SA can identify keywords and phrases expressed, and insurers can use this analysis to help manage their business and communicate more productively with potential and current customers. Previous research in the same area and comparing related works with this project will be discussed in the following subsections.

2.6.1 Social Media Analysis of Allstate Insurance Company Using Twitter Data

This research described the analysis of Twitter posts related to the keyword Allstate corporation. It also discusses the implementation of association

analysis, correlation, and clustering by analyzing insurance Twitter posts. This research aims to extract and analyze customer feedback by identifying keywords related to Allstate to facilitate the use of this information by insurers. Roosevelt et al. (2012) proposed that insurers may apply the analysis findings in the relevant areas to enhance customer satisfaction and improve overall service quality more efficiently.

Allstate corporation was chosen solely based on the accessibility of public historical Twitter data. The language used to extract data from Twitter is in English. Two approaches were implemented in this analysis to distinguish tense and spelling or case differences at the data processing stage. The first approach calculates the difference between two strings by measuring the Levenshtein edit distance (LED). It is described as the sum of single-character insertions, replacements, or deletions needed to convert one string to another. Another approach to make a comparison between strings is to calculate the generalized edit distance (GED). The minimum cost sequence of operations for constructing two strings is calculated between string one and string two. The next step is identifying patterns and variations of words to signify ideas and themes. A simple correlation analysis is done to detect correlations between pairs of words. There are also two additional types of analysis carried out on the data, which are clustering analysis that group tweets depending on their similarities or dissimilarities and association analysis, assigned to analyze the frequency of specific words together.

2.6.2 Using Social Media to Identify Consumers' Sentiments Towards Attributes of Health Insurance

This study aims to identify the sentiments consumers have about health insurance by evaluating what they communicate on Twitter. The primary objective was to use sentiment analysis to classify consumer attitudes towards health insurance and health care providers (Broek-Altenburg & Atherly, 2019). The tweets from Twitter associated with the terms 'health insurance' or 'health

plan' during the United States health insurance enrollment season from 2016 to 2017 were obtained using an application programming interface (API). Apart from that, word association was used to identify words related to 'premium', 'access', 'network', and 'switch'. The Sentiment analysis is used to identify different emotions associated with insurance, and subsequently, the results will help identify the origin of the sentiments that drive consumers. Thus, consumers are able to navigate insurance plan options better, and insurers can better respond to their needs.

2.6.3 Features Comparison of Related Work

This section provides a summary of the techniques and platforms used based on the related works stated. Table 2.4 shows several features and advantages of this project compared to the previous work explained in the previous section.

Table 2.4 Features Comparison of Related Work

Features	**Title**			**Advantage(s) of This Project**
	Social Media Analysis of Allstate insurance company using Twitter data	Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance	Sentiment Analysis of Malaysian Insurance Companies (SAMIC): A Visualization using Support Vector Machine Algorithm	
	Source: Roosevelt & Mosley, 2012	Source: Broek-Altenburg & Atherly, 2019		
Method Used	• Levenshtein Edit Distance (LED) • Generalized Edit Distance (GED) • Data mining methods used, including correlation analysis, clustering analysis, and association analysis, to analyze data	• Sentiment Analysis • Application Programming Interface (API) to gather tweets from the Twitter server • Vector space model to reduce dimensionality • NRC Emotion Lexicons	• Sentiment Analysis • TF-IDF used to mine text data • Visualize analyzed data using data visualization technique, Automatic classification using the Support Vector Machine algorithm	Text classification with a machine learning algorithm is more accurate than human-made rule systems, particularly in complex classification tasks. Additionally, data visualization is capable of making the data more accessible and faster to be read by the insurer
Extracted tweets language	English	English	English and Bahasa Malaysia	This project incorporates bilingual languages capability to extract Twitter data in the sentiment analyzer
Targeted Company	Allstate Corporation	Affordable Care Act marketplaces in the United States	AIA, Great Eastern, Prudential	This project includes an automatic text classification that has not yet been explored in the previous study
Visualization Techniques	Simple graphs and charts	Simple graphs and charts	Plotly's Python Graphing Library	This project offers enhanced interactivity of data
Computing Platform	No platform	No platform	Web-based platform	This project enables quick access to relevant business insights

2.7 Conclusion

In this chapter, previous researches and studies on related topics have been reviewed. This chapter's main objective is to provide a literature review of the relevant studies to establish a sustainable foundation based on previous researchers' critical review. Firstly, to gain insight into the insurance industry in Malaysia, the insurance industry's overview and the types of insurance business provided are discussed. The insurance company ranking that provides both conventional and Takaful insurance policies are disclosed, and these companies will be the main scope of this project. Next, factors that can affect insurance companies' performance, including Service Quality and Customer Satisfaction, are addressed. This conducted literature review also provides information on the approaches and challenges in analyzing social media data. Most of the reviewed papers use Twitter as their social media context due to the availability and accessibility of Twitter content. Different data visualization techniques are explained, and it was identified as the best way to interpret and visualize the data extracted. A comparison of development approaches of both the web and native approach has been made to choose the best approach suited for this project. Lastly, past related work is evaluated.

CHAPTER 3

METHODOLOGY

This chapter defines the research methodology used to design and develop this project. The modeling approach chosen for system development is explained according to the phases involved. It begins with system requirements until the end of the deployment and review phase. The reasons and justification for the research design and data collection techniques will be provided. Besides, the hardware and software specifications required to accomplish the objectives of system development for this project are described in this chapter.

3.1 Introduction

The methodology is a set of principles or guidelines for gathering and validating knowledge of a subject. Alternatively, a software development methodology is a framework used to plan, organize, and manage the process of developing an information system. It is a way of managing a software development project and provides the appropriate guidelines with a set of rules for planning all software development stages (Half, 2019). Various software development models can be used to develop a system, including Waterfall, Spiral, Rational Unified Process (RUP), and Agile. Hence, selecting an appropriate software development approach must be carefully considered because it will make a big difference in achieving a successful result. The model that will be used in the development of this project is the modified methodology. The methodology proposed for this project is discussed in the following section.

3.2 Modified Waterfall Methodology

The waterfall model is the sequential approach adopted in software development in which a continuous flow from one phase into the other occurs. The flow resembles a waterfall and is referred to as the traditional waterfall model (Lynch, 2018). There are various phases that the software passes through in this model. The phases are sequential and rigid, which indicates that the next phase in the software life cycle only began after the previous phase was completed (Casteren, 2017). The traditional waterfall model has the downside, and if something had gone wrong in the initial phase of the project, it could only be revealed at the final phase. Besides, it is impossible to go back to the previous stage because the software has progressed from one phase to the next phase. The modified waterfall methodology came into existence because of this deficiency in the traditional waterfall design (Prashant, 2020). The phases are similar as well, as illustrated in Figure 3.1.

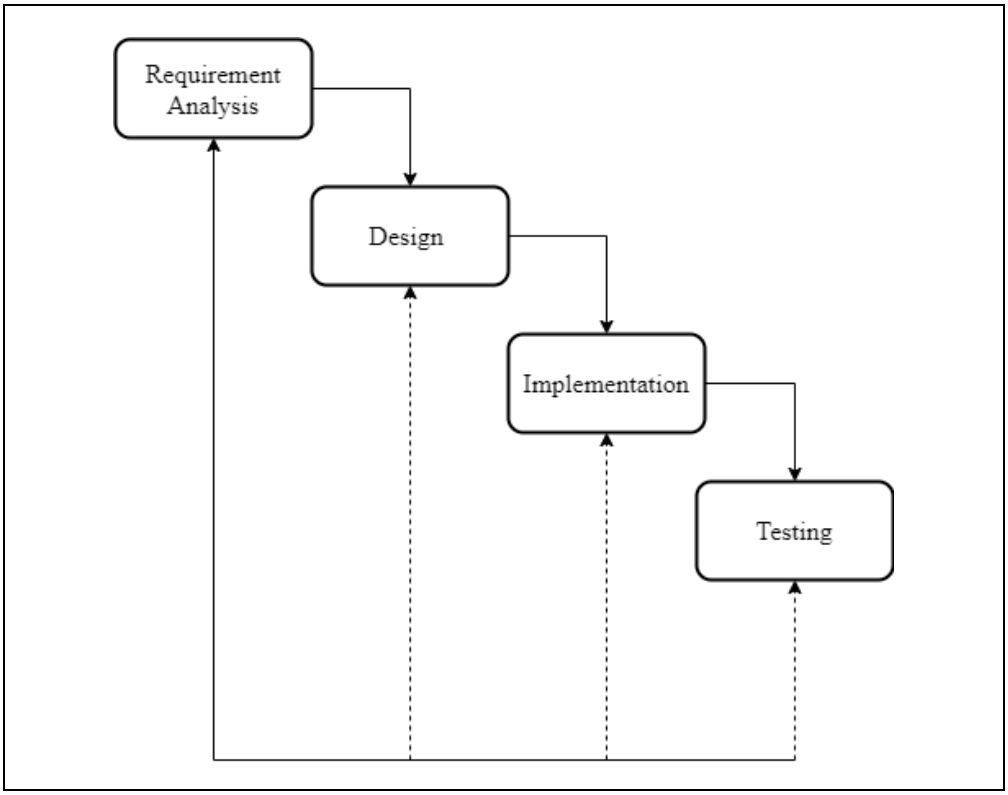


Figure 3.1 Phases in Modified Waterfall SDLC

A traditional waterfall model is not suitable for this project since there may be an overlap between the software lifecycle phases. As it is possible to progress through different phases, modified waterfall models are more appropriate as all stages in this advanced model are allowed to overlap. The main reason behind the adoption of a modified waterfall model is that it allows for overlapping between phases, which can provide flexibility and stability that is highly needed for this project. Because the phases overlap, much flexibility has been introduced (Ratanadewi & Bachtiar, 2020). The model is suitable for the nature of the system to be developed since it is a small project where all the requirements are defined. The overlapping phases will also improve productivity as several phases may be carried out at a time. Thus, the model is ideal for developing this system as the requirements of this project are well-understood and have clear processes, so most of the requirements can be expected in the early phases.

Additionally, making changes to the application's design is also possible since several phases are active at one particular time. In case there are any errors introduced because of the changes made, it is also convenient to rectify them, which helps minimize any oversight errors. In order to obtain a clearer picture of the methodology used for this project, Table 3.1 summarizes the methodology used for the implementation of this proposed project and its deliverable outcomes by phases. The first phase in the modified waterfall methodology will be explained in the next section.

Table 3.1 Summary of Modified Waterfall Methodology

Phase	**Task**	**Deliverables**
Requirement Analysis	• Identify the area and the purpose of the proposed system implementation • Set the project title • Define the problem statements • Identify the objectives, scope, and significance of this project	Chapter 1: • Background of study • Survey • Project Timeline • Problem statements • Project scope • Project significances
	• Collect all relevant information for the project • Review existing journal, article, and book • Review the existing related projects.	Chapter 2: • Literature Review
Design	• Identify project methodology • Define the technique and procedure for the development of the proposed system • Design flowchart, use case, and user interface	Chapter 3: • Methodology • Flowchart • Use Case • User Interface
Implementation	• Data Preparation of training and testing sets • Collection of data for real-world implementation • Develop the Support Vector Machine Classifier Model • Build a system prototype • Apply the classifier model on top of the system prototype	Chapter 4: • Design and Development
Testing	• Test all available functions to ensure all requirements are met and work well together • Conduct system acceptance testing, including the functionality and usability test	Chapter 5: • Results and Findings Chapter 6: • Conclusion and Future Works

3.3 Requirements Analysis Phase

Requirements analysis is the first phase of this project, and it is the most crucial part of SDLC, where a brief set of functionalities that the system requires to meet to be successful was identified. This phase aims to describe in more detail the system inputs, outputs, processes, and interfaces. The problem statements, project objectives, scope, and significance were specified for acquiring the project's requirements. Before setting out the problem statements to determine the policyholders' difficulties during the decision-making process in purchasing insurance policies, a survey was conducted. All of the survey questions are set out in Appendix A.

Next, the project objectives defined were converted into defined system functions for the intended project. Functional and non-functional systems are analyzed and recorded based on the information obtained. Additionally, literature reviews in terms of techniques, suitable development approaches, and reviews of the current system have been stated by referring to the journals, internet, and research papers. The project timeline used to track the project's progress using a Gantt Chart, as attached in Appendix B. Figure 3.2 describes the process's representation in the requirements phase.

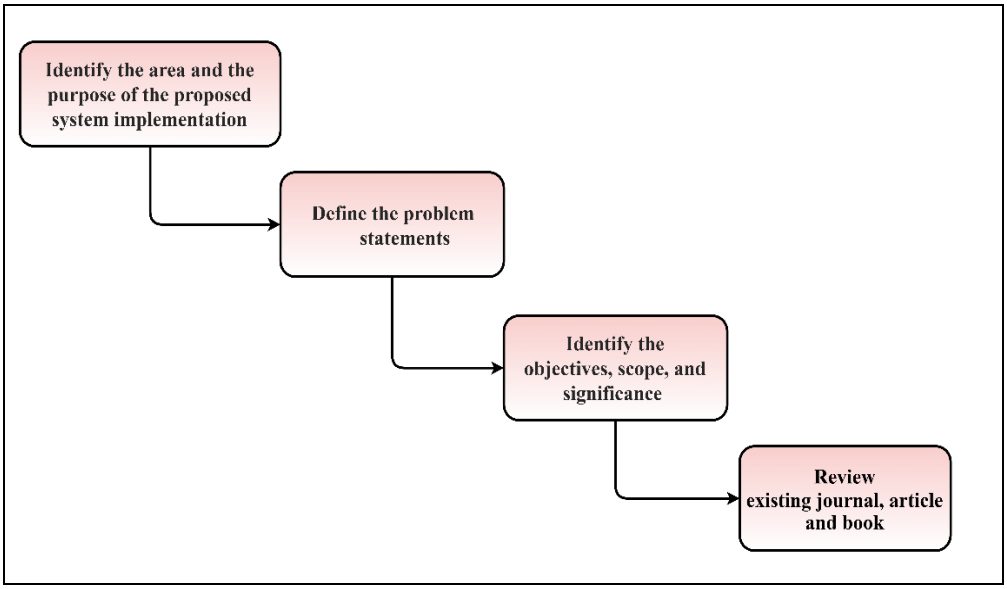


Figure 3.2 Process Flow in the Requirements Phase

3.3.1 Survey

A survey was conducted prior to setting this project's objectives by creating a questionnaire comprising ten questions through Google form, distributed to potential insurance policyholders. The scope of possible responses is limited to insurance policyholders to refine how the questionnaire is answered and prevent incorrect findings. There are 64 respondents recorded in this survey, as in Figure 3.3. However, only the employees' perspectives were assessed since they belong to the more likely groups to purchase an insurance policy. The answers obtained from the questionnaire are summarized and illustrated in the form of charts and graphs.

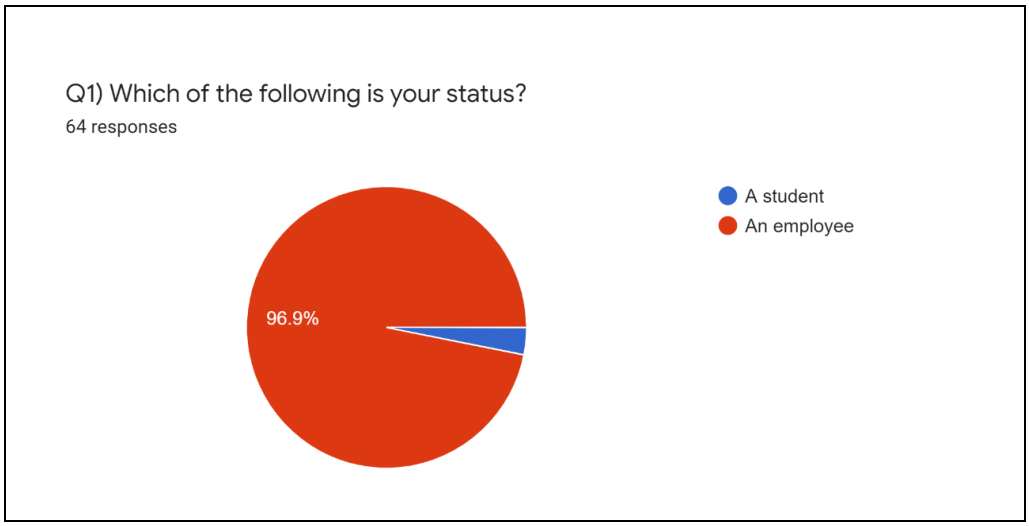


Figure 3.3 Percentage Status of the Survey Respondents

The survey results were further screened so that only employees who have purchased insurance policies are recorded. From the 62 responses submitted, 91.9% of employees had an insurance policy, and 8.1% did not, as illustrated in Figure 3.4.

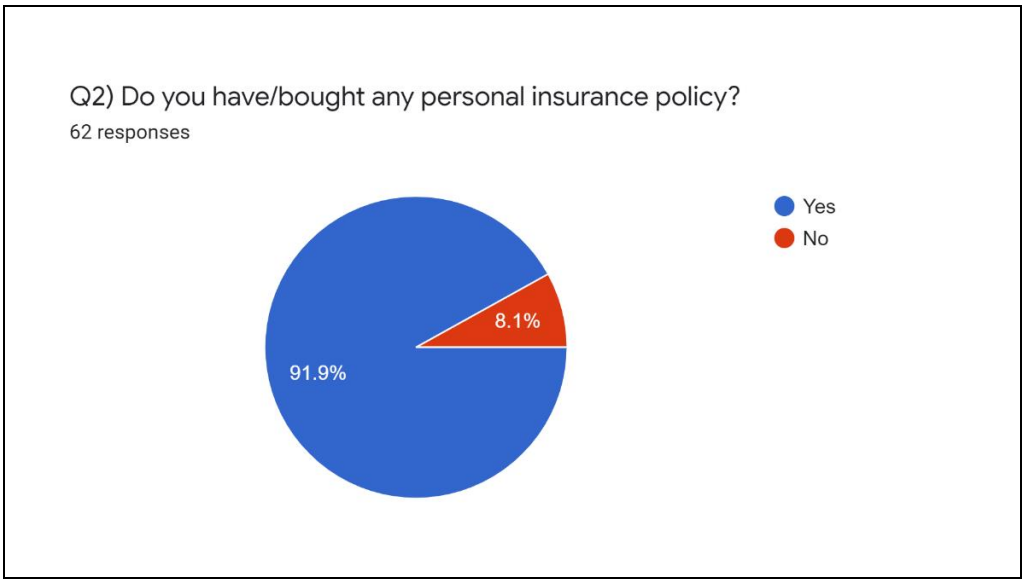


Figure 3.4 Percentage of Employees with Personal Insurance Policy

Following on from the previous question, 96.8% of these 62 employees also claimed that they encountered problems or difficulties in choosing insurance companies in Malaysia to purchase an insurance policy, as in Figure 3.5.

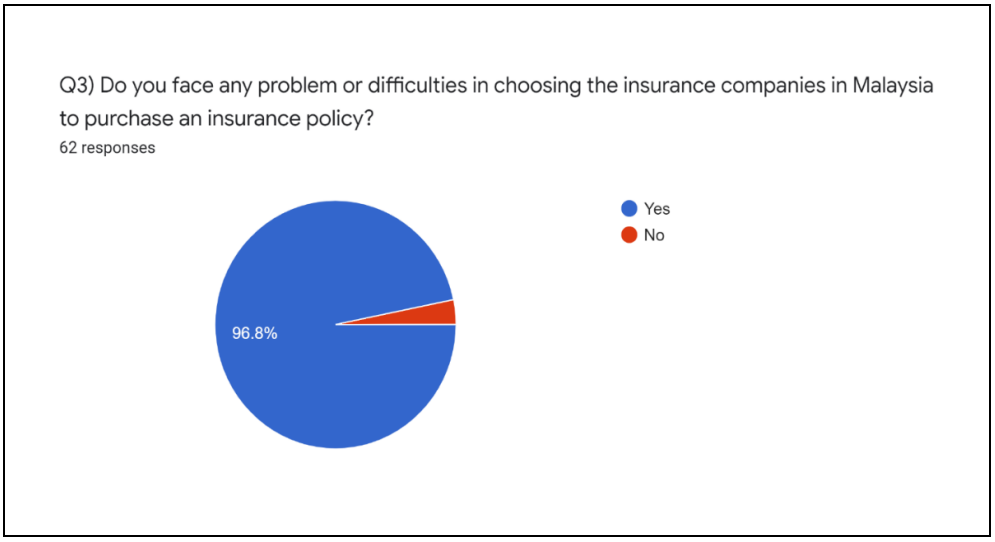


Figure 3.5 Percentage of Respondents' with Difficulties Purchasing Insurance

As demonstrated in Figure 3.6, the chart shows the types of difficulties encountered by insurance policyholders when purchasing insurance. Out of 60 responses, 83.3% found it challenging to look for a trustworthy advisor, 43.3% in finding a proper insurance policy, 45% to choose an insurance provider with financial solidity and, 5.0% for other reasons.

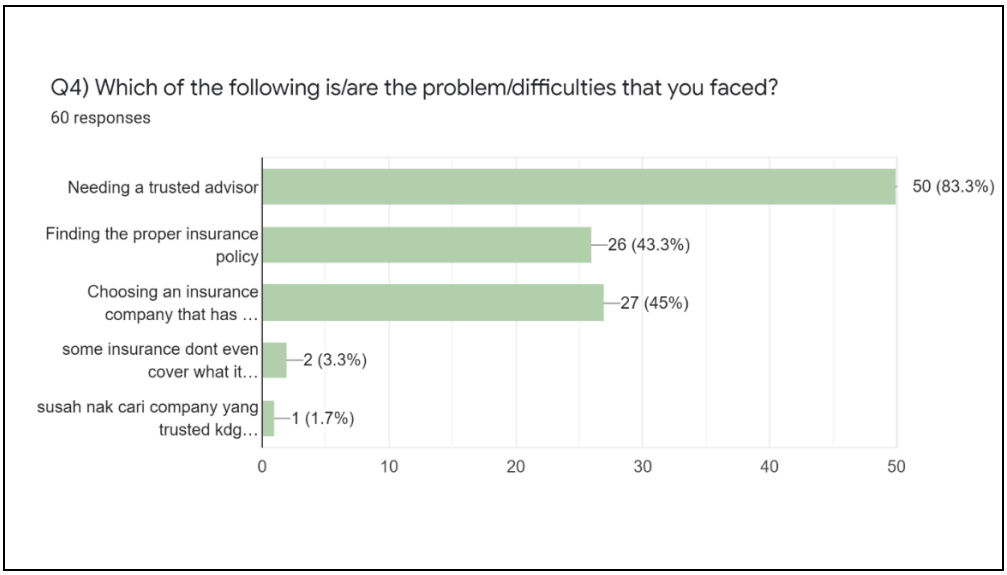


Figure 3.6 Percentage of Difficulties Respondents' Faced

During the purchasing decisions to buy an insurance policy, 94.9% of the 59 respondents referred to online reviews of an insurance company, followed by 27.1% of the brochure distributed by the insurance companies, and 27.1% referred to insurance agents. The other 6.8% of respondents referred to other sources, as in Figure 3.7.

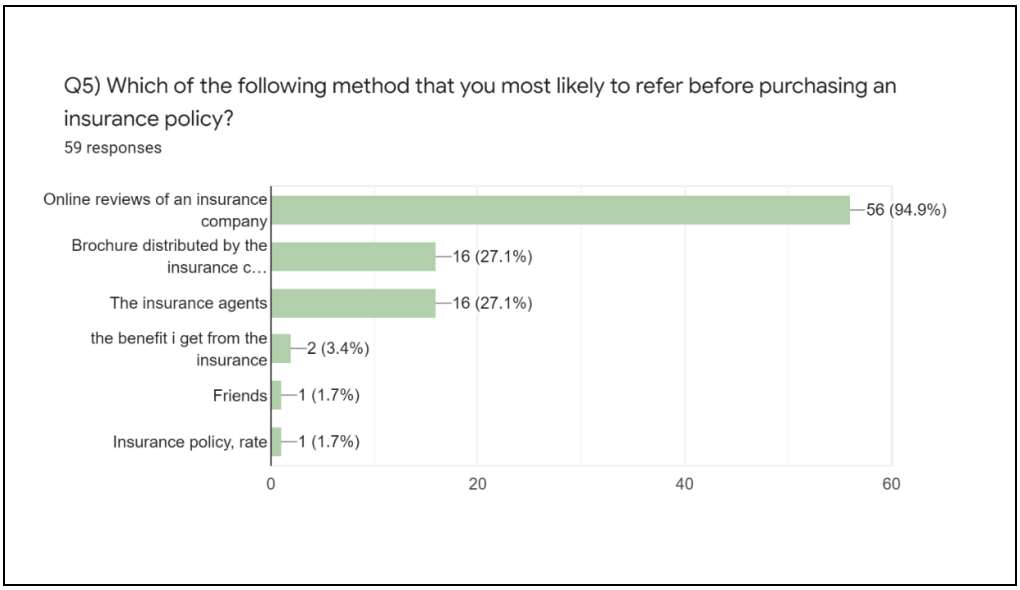


Figure 3.7 Percentage of Respondents' Reference Before Purchasing Insurance

As demonstrated in Figure 3.8, the next question asked whether the respondents are aware that it was not easy to review the company's credibility

and reputation when purchasing insurance. 100% of the 58 respondents agreed that it was time-consuming to review the company's reputation and history during the surveying process to buy the best insurance in accordance with their demands.

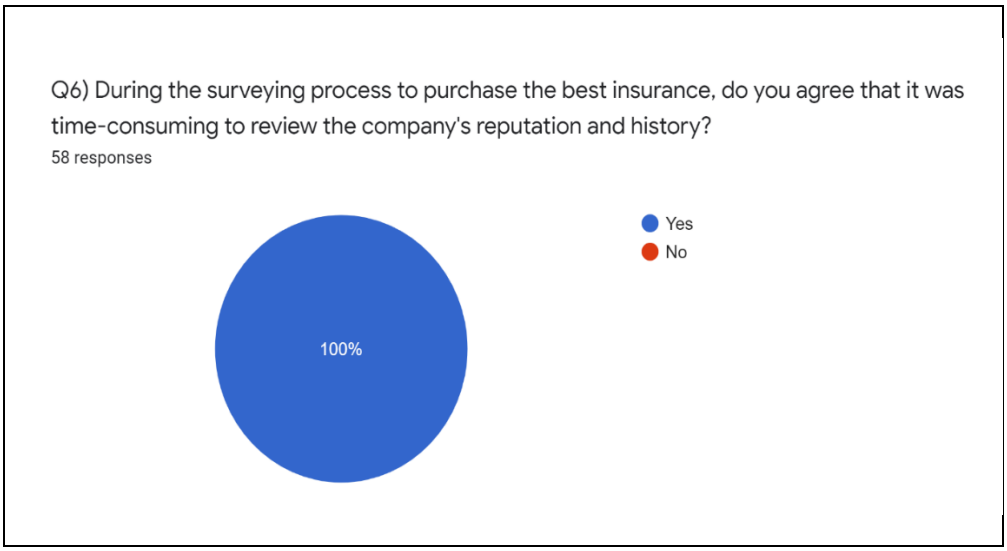


Figure 3.8 Percentage of Respondents' Agreed to the Statement in Question

Figure 3.9 depicts the expectations that need to be met by insurance providers are identified based on the respondents' previous experiences. 86.4% of respondents expected to have good customer service, while another 33.9% expect the insurance providers will have the ability to promptly handling claims. Apart from that, 37.3% wanted to have reliable insurance policy quotes, 20.3% recommendation of sufficient insurance protection, and 1.7% for coverage and benefits.

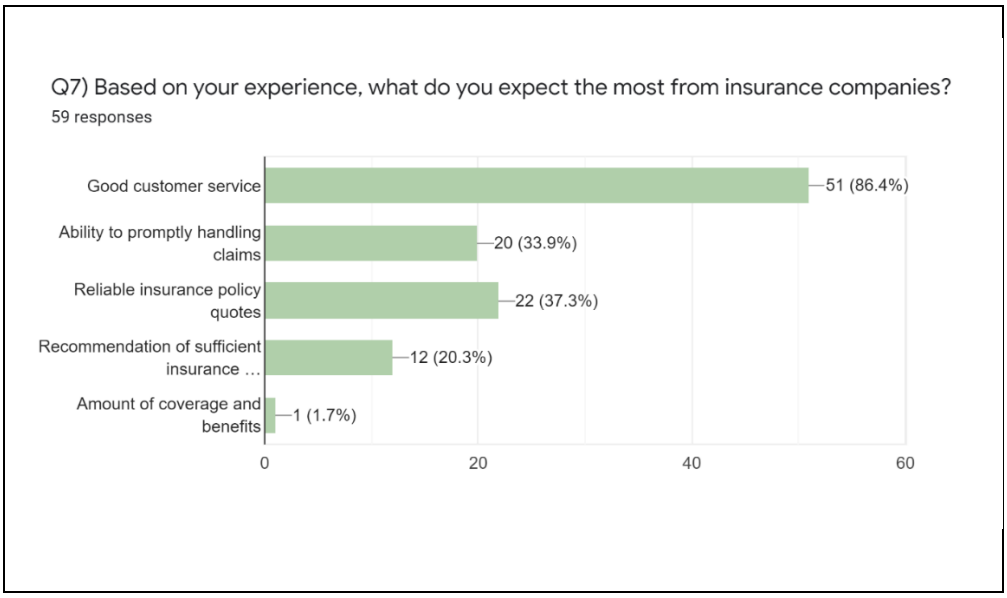


Figure 3.9 Percentage of Respondents' Expectations towards Insurance Companies

In the next question, 40% claimed that great eastern is the best insurance company in Malaysia, another 31.7% prudential, and 28.3% AIA as illustrated in Figure 3.10.

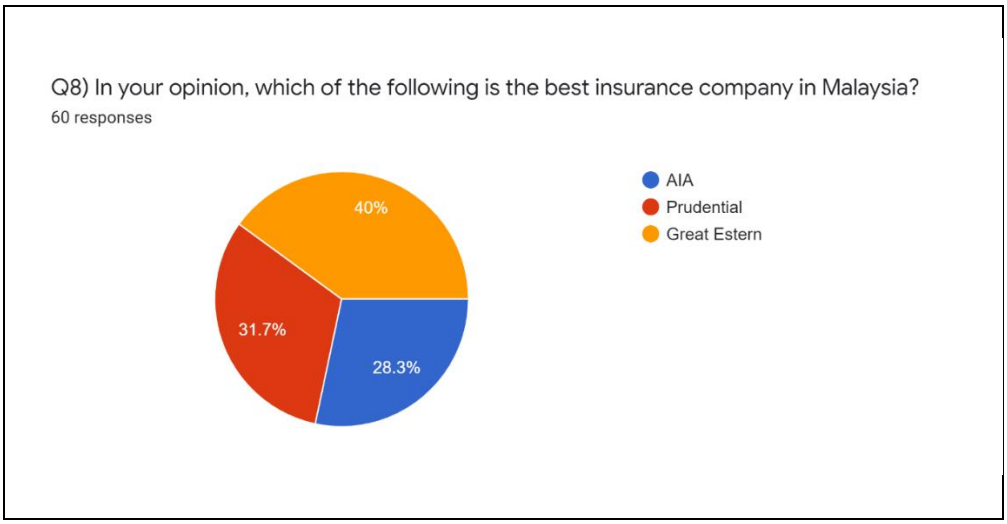


Figure 3.10 Percentage of Best Insurance Companies According to Respondents

As in Figure 3.11, 100% of respondents agree to have a platform that helps them decide on purchasing insurance policies. Therefore, this project's development allows existing and potential insurance policyholders to resolve the current practice during the insurance policy's purchasing decisions.

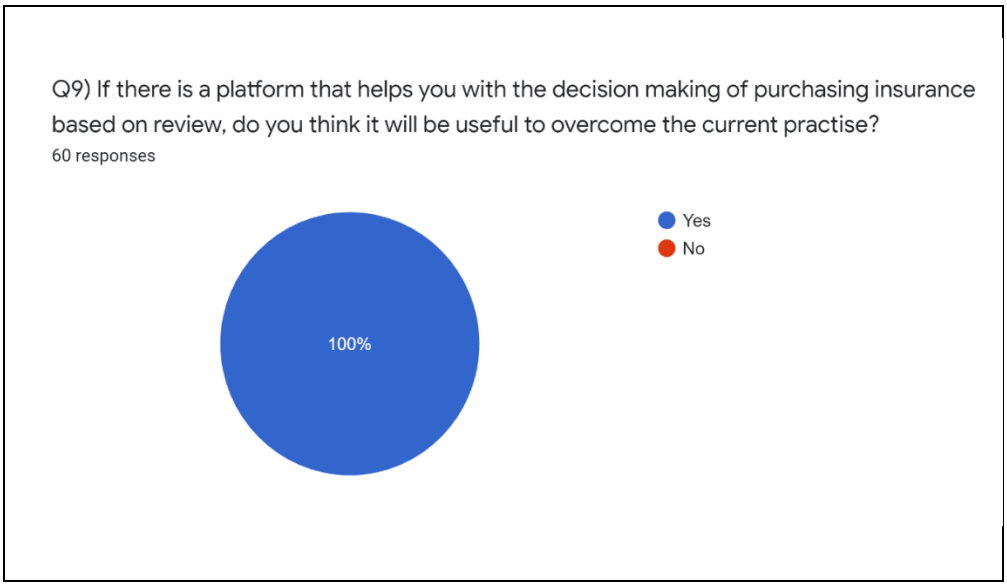


Figure 3.11 Percentage of Respondents' Agreed with the Development of this Project

Figure 3.12 shows that there were 58 responses to the question of why this project could assist policyholders in the decision-making phase of purchasing insurance.

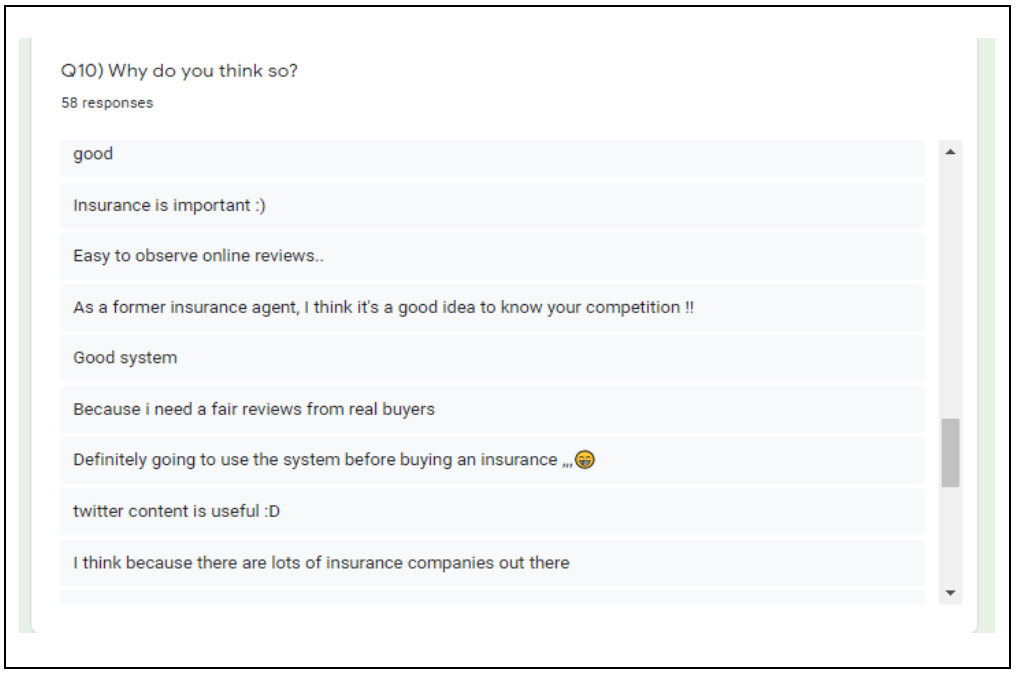


Figure 3.12 The Responses Received from Policyholders

Table 3.2 provides an overview of the findings, and the percentage of each question answered in the survey. In the next section, the project timeline used in this project is defined.

Table 3.2 Results of the Survey

No.	Questions	Answers	Percentage (%)
1.	Which of the following is your status?	A student	3.1
		An employee	96.9
2.	Do you have/bought any personal insurance policy?	Yes	91.9
		No	8.1
3.	Do you face any problems or difficulties in choosing the insurance companies in Malaysia to purchase an insurance policy?	Yes	96.8
		No	3.2
4.	Which of the following is/are the problem/difficulties that you faced?	Needing a trusted advisor	83.3
		Finding the proper insurance policy	43.3
		Choosing an insurance company that has a financial solidity	45
		Others	5.0
5.	Which of the following method that you most likely to refer to before purchasing an insurance policy?	Online reviews of an insurance company	94.9
		Brochure distributed by the insurance companies	27.1
		The insurance agents	3.4
		Other	3.4
6.	During the surveying process to purchase the best insurance, do you agree that it was time-consuming to review the company's reputation and history?	Yes	100
		No	0
7.	Based on your experience, what do you expect the most from insurance companies?	Good customer service	86.4
		Ability to promptly handling claims	33.9
		Reliable insurance policy quotes	37.3
		Recommendation for sufficient insurance protection	20.3
		Other	1.7
8.	In your opinion, which of the following is the best insurance company in Malaysia?	AIA	28.3
		Prudential	31.7
		Great Eastern	40

9.	If there is a platform that helps you with the decision making of purchasing insurance based on a review, do you think it will be useful to overcome the current practise?	Yes	100
		No	0
10.	Why do you think so?	We do need unbiased reviews	
		Because I need to buy good life insurance	
		So that I can know which company is the best	
		Advantageous	
		I do not trust what the insurance providers claimed	
		Twitter has lots of online reviews	
		It is kind of difficult to compare so many companies	
		Nice system	
		Since insurance plays a crucial role in the sustainable growth of an economy	
		No insurance claims denials especially auto insurance claims	
		Good platform and help me make the right decisions	
		Helpful making the right choice	
		Easy to observe online reviews	
		Good system	
		Because I need fair reviews from real buyers	
		Beneficial	

3.3.2 Project Timeline

Project timeline used to track all activities to ensure that this project is completed on schedule, the Gantt charts are used to visually represent all of the activities. It outlines the tasks involved in this project and their order against a timescale. As provided in Appendix B, the project overview and its associated task deadlines are shown as horizontal bars, and the length of the bar implies how long that task is supposed to take. Hence, missing the deadline or finishing a task out of sequence can be prevented. The next section explains the design phase that took place after the requirements phase has ended.

3.4 Design Phase

The next phase of the modified waterfall methodology is the design phase, which began after the requirements phase has been completed. The purpose of conducting the design phase is to plan out a system that meets the requirements phase requirements. As illustrated in Figure 3.13, the means of implementing project solutions are specified in the design phase, such as how the product will be created.

Next, to know the details of the data exchanged between the system and the user, a use case diagram is developed. The subject boundary which separates the system and the user was represented from the use case. Moreover, the activities performed by the user with the system can also be seen. A proposed user interface in terms of appearance and layout will be structured to describe how users interact with the computer system through the system's menus, functions, output, and features. Each of the tools to conduct the key activities in determining the proposed system’s inputs and outputs will be further explained in Chapter 4.

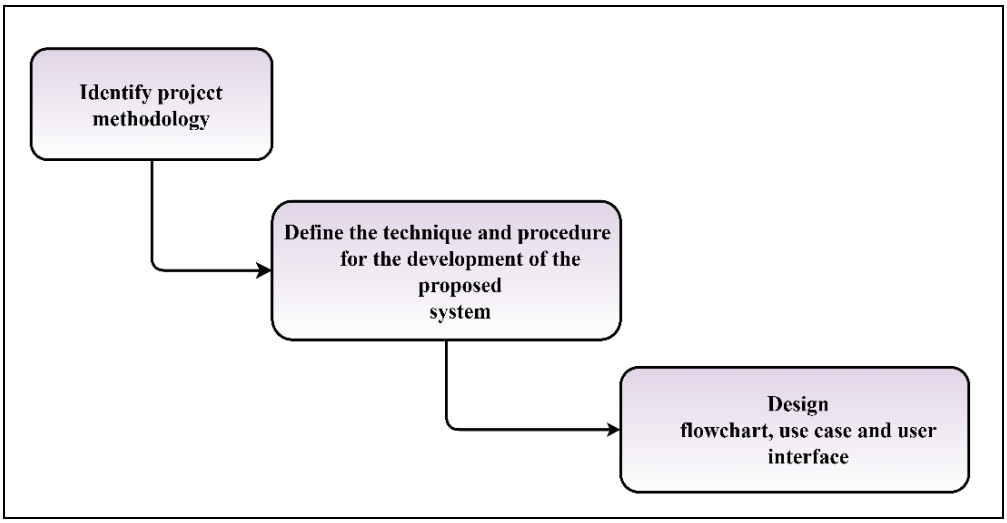


Figure 3.13 Process Flow in the Design Phase

3.5 Implementation Phase

In the implementation phase, all the specifications from the previous phase that laid the foundation for system development are transformed into an actual system. The purpose of the development phase is to convert the system's conceptual design during the design phase into a working system that addresses all documented system requirements. Therefore, the creation of the system and the actual code will begin based on given specifications. At the end of this phase, the working system will enter the testing phase.

In order to implement this project, data related to insurance companies are first collected using dates and keywords to scrap the text data. Twitter was used as a source of data in this project. Tweets in Bahasa Malaysia and English directly or indirectly referring to the insurance companies are extracted using Twint, an advanced Twitter scraping tool written in python that allows tweets to be scrapped from Twitter profiles without the use of the Twitter API. Next, the extracted was pre-processed to obtain better features. Text cleaning is done by removing stop words, punctuation, and stemming were a few of the techniques used to clean the text that does not give information and increase the complexity of the analysis.

Before applying a machine-learning algorithm to extract data features, the term frequency-inverse document frequency (TF-IDF) concept is applied. Irrelevant text is reduced from the corpus of tweets collected, and the document term matrix is then generated. In order to portray the sentiment generation, a support vector classifier is used to classify tweets into positive, neutral, and negative sentiments. The development step representation is defined as in Figure 3.14.

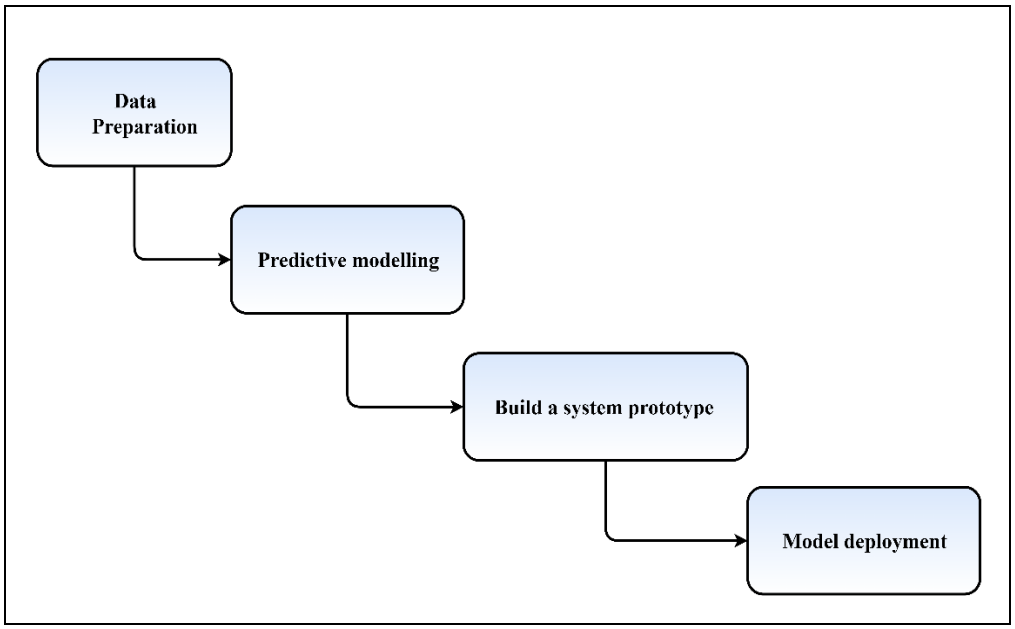


Figure 3.14 Flow of Process in the Implementation Phase

3.6 Software and Hardware Requirements

As listed in Table 3.3 and Table 3.4, the software and hardware specifications are the software programs and hardware devices used for the system development. The proposed system is built using proper hardware devices and software programs to ensure that the functionality of the system will operate effectively. Besides, it is developed in a web environment that allows the system to be accessed on any platform with a web browser connected to the internet. The next section discussed the testing phase, which takes place after the development phase has ended.

Table 3.3 Software Requirements

Software	**Description**
Microsoft Word	To write the project report and Gantt chart to visualize the progress of the project
Microsoft Excel	To prepare training dataset
LucidChart	To create a flowchart and use case diagram
Jupyter Notebook	As the IDE for data pre-processing and model development
PyCharm	As the IDE for application development using Python Flask
Spyder	As the Python development environment for data scraping and data manipulation
Plotly	As the source code editor for website development in JavaScript and CSS

Table 3.4 Hardware Requirements

Hardware	**Specification**
Laptop Model	MSI GF63 Thin 9SCSR
Processor	Intel Core i5-9300H CPU @ 2.40GHz
Operating System	Windows 10
Memory (RAM)	8GB
Solid State Drive (SSD)	458GB
External Hard Drive	1 TB
System Type	64-bit

3.7 Testing Phase

Once the system has been completed, it is deployed in the testing environment. The testing phase focuses on the execution of test cases on the system, and the results reflect the system's reliability. The entire system’s functionality is tested to verify that the whole system complies with the previous phase’s requirements.

The testing phase consists of three stages. The first stage is component testing, where all components are tested to ensure that they are working correctly. The next stage is requirements testing, in which the complete system is checked against the requirements. This stage’s main objective is to ensure that the

system works as specified in the requirements and that the system does not do anything that is not a universal norm. Once the required testing is completed satisfactorily, the acceptance stage will take place in which the complete system will be reviewed from a user perspective, as outlined in Figure 3.15. Functional testing and usability testing conducted during the last stage, which is acceptance testing, will be explained in the next subsection.

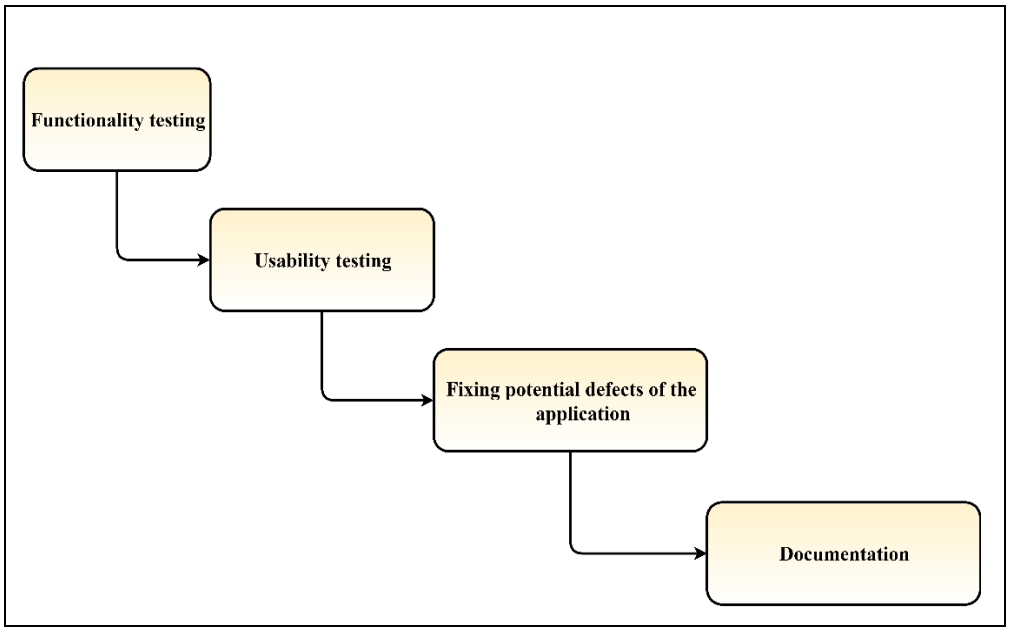


Figure 3.15 Flow of Process in the Testing Phase

3.7.1 Functionality Testing

Functionality testing is a form of software testing that tests the system against its functional requirements. The user must verify all the functionality of the system listed in Table 3.5. Functions included in the test cases are tested by entering the input and examining the output to check whether the functions are successful, thereby ensuring that the system's functionality performs as expected.

Table 3.5 List of Functionality Test Cases

Test Case	**Expected Result**	**Success/Failure**
Login Page	The user credentials must be verified	
View Overview Page	The overview page of the three companies and a line chart to visualize the summary of the analysis will be displayed	
View AIA Page	The results of sentiment analysis are shown, and the data are visualized through data visualization techniques on the AIA page	
View Prudential Page	The results of sentiment analysis are shown, and the data are visualized through data visualization techniques on the Prudential page	
View Great Eastern Page	The results of sentiment analysis are shown, and the data are visualized through data visualization techniques on the Great Eastern page	
Download Excel Files with Labelled Sentiment	The Excel files with labeled sentiment will be saved into the user's computer	
View Twitter Updates	The real-time timeline of Twitter insurance updates is presented to the user	
Analyze Text	Users can insert any text input into the text sentiment analyzer page	
View Results of Sentiment Analysis	The result of sentiment analysis on the text entered by users are displayed	
View Tweet Sentiment Analyzer Page	The result of sentiment analysis from tweets extracted in real-time will be displayed on the page	
Search Tweet Sentiment in Real-Time	The user can search any tweet keywords to analyze the sentiment	
View Agents Analysis Page	The result of extracted Twitter user profile accounts of insurance agents will be displayed	
Compare Performance of Companies	The results of the analysis and the comparison of sentiment between companies are visualized to the user	
Logout Page	The user can logout from the application	

3.7.2 Usability Testing

The usability test is carried out to learn about the system workflow and whether the information is correctly formatted and delivered. Based on the insights gained, improvements to the system can improve user experience and design better system workflows. It is also essential to identify whether the user has found workarounds because the interface is not intuitive. Table 3.6 lists a Likert scale survey questions that are conducted to record user experience. A five-point agreement scale is used to measure respondents' agreement with the statement provided. The response options to measure the overall reaction during the system's use are represented by a scale of 1 is ‘strongly disagree’, 2 is ‘disagree’, 3 is ‘neutral’, 4 is ‘agree’, and 5 is ‘strongly agree’. The last phase of the modified waterfall methodology is explained in the following section.

Table 3.6 The System Usability Survey Questions

Number	**Overall Reaction to the System**	**Scale**				
		1- strongly disagree, 2 - disagree, 3 - neutral, 4 - agree, 5 - strongly agree				
1	I think that I would like to use this feature frequently	1	2	3	4	5
2	I found the feature unnecessarily complex	1	2	3	4	5
3	I found this system was meaningful	1	2	3	4	5
4	I think that I would need the support of a technical person to be able to use the application’s feature	1	2	3	4	5
5	The interface of this system is pleasant	1	2	3	4	5
6	I am not able to navigate to other pages easily	1	2	3	4	5
7	The content on the system was relevant	1	2	3	4	5
8	The content presented was not in the right format	1	2	3	4	5
9	The system has all the expected functions and capabilities	1	2	3	4	5
10	Overall, I am not satisfied with this system	1	2	3	4	5

3.8 Conclusion

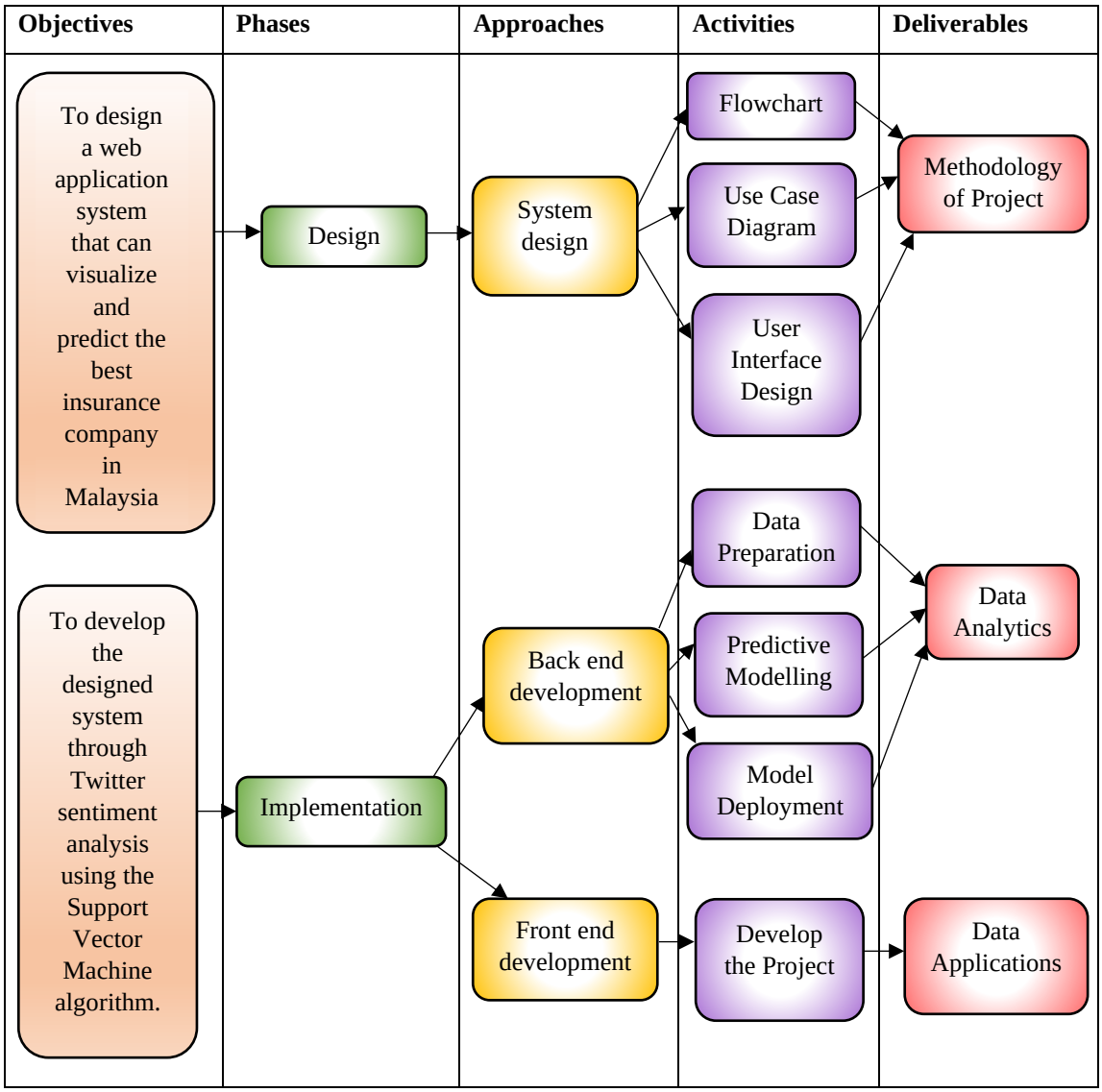
This chapter summarizes the methodology used in this project that guides the completion of this project over a specified time. Modified Waterfall SDLC is selected because the phases are suitable for developing web applications, particularly in terms of developing the system's functionality. Five modified waterfall phases are requirements, design, implementation, testing, deployment, and maintenance, as described throughout this chapter. Each phase has its activities to ensure that the expected output produced is consistent with the input entered.

CHAPTER 4

DESIGN AND DEVELOPMENT

This project involves full-stack web development, which implies that the back end and front end are two parts included in the web application. According to the modified waterfall approach described in the previous chapter of this study, system design and implementation are discussed in this chapter. The details and steps for creating the web application that deploys the machine learning model is illustrated in Table 4.1.

Table 4.1 Design and Implementation Phase of the Application



4.1 System Design

The design of systems may be regarded as the application of systems principles to product development. As provided in the following subsections, the design process is aided by creating design diagrams, such as a use case diagram, flowchart, and user interface (UI).

4.1.1 Flowchart Diagram

A flowchart is a tool for analyzing the process of a system that focuses on the configuration of a system. It is an integral part of the development of the proposed system. Figure 4.1 and Figure 4.2 show the flow of this project's overall system, including the system features. Immediately after the system is running, the user is taken to the login page for admin. They are asked to fill in their username and password to login into the system. Once the system authenticates the details provided, the user is directed to the Overview page. The user is able to access the brand talkable favourability (BTF) for all insurance companies and a summary of the results of the scrapped tweets. If the user clicks the ‘View AIA page’ from the drop-down menu, the AIA page is displayed with all the details, including the BTF and visualized data. Once the user clicks on either ‘Prudential’ or ‘Great Eastern’ from the list, the user is directed to the Prudential or the Great Eastern page. These two pages’ content was similar to the AIA page, except for the respective company domain. Suppose the user chooses the ‘Sentiment Analyzer’ page from the drop-down menu. In that case, the user can choose to perform text or real-time sentiment analysis on the page. On the text sentiment analyzer page, the user is able to submit any input text, and when the ‘Submit’ button is clicked, a sentiment analysis built on the Support Vector Classifier is performed on the text. If the user clicks on the ‘View Twitter Updates’ page, the timeline of Twitter insurance updates is presented. Lastly, suppose the user clicks on the ‘View Competitive Analysis’. In that case, the results of the analysis are visualized, and the user can view the comparison of sentiment between companies.

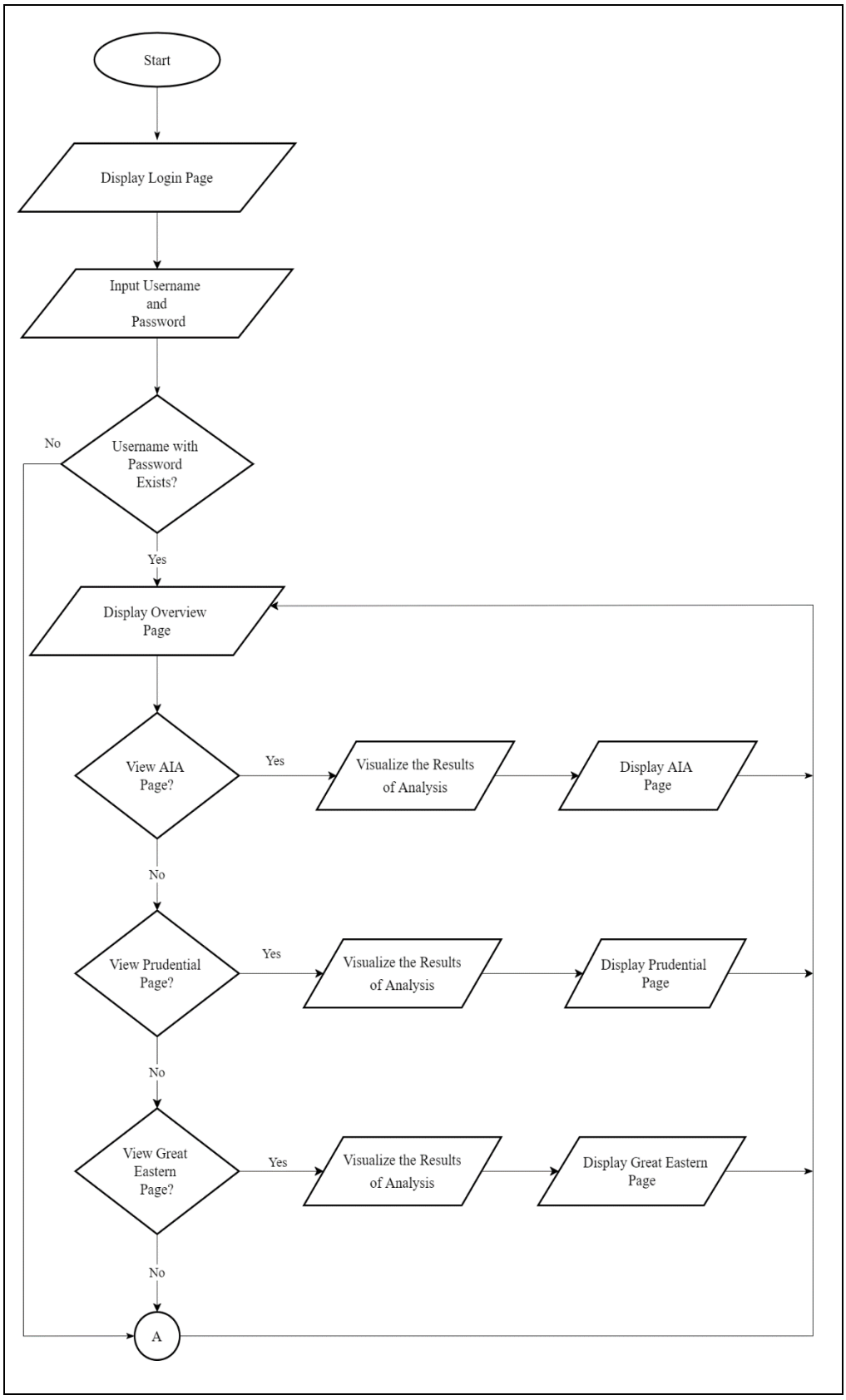


Figure 4.1 Flowchart Diagram of the Overall System Flow

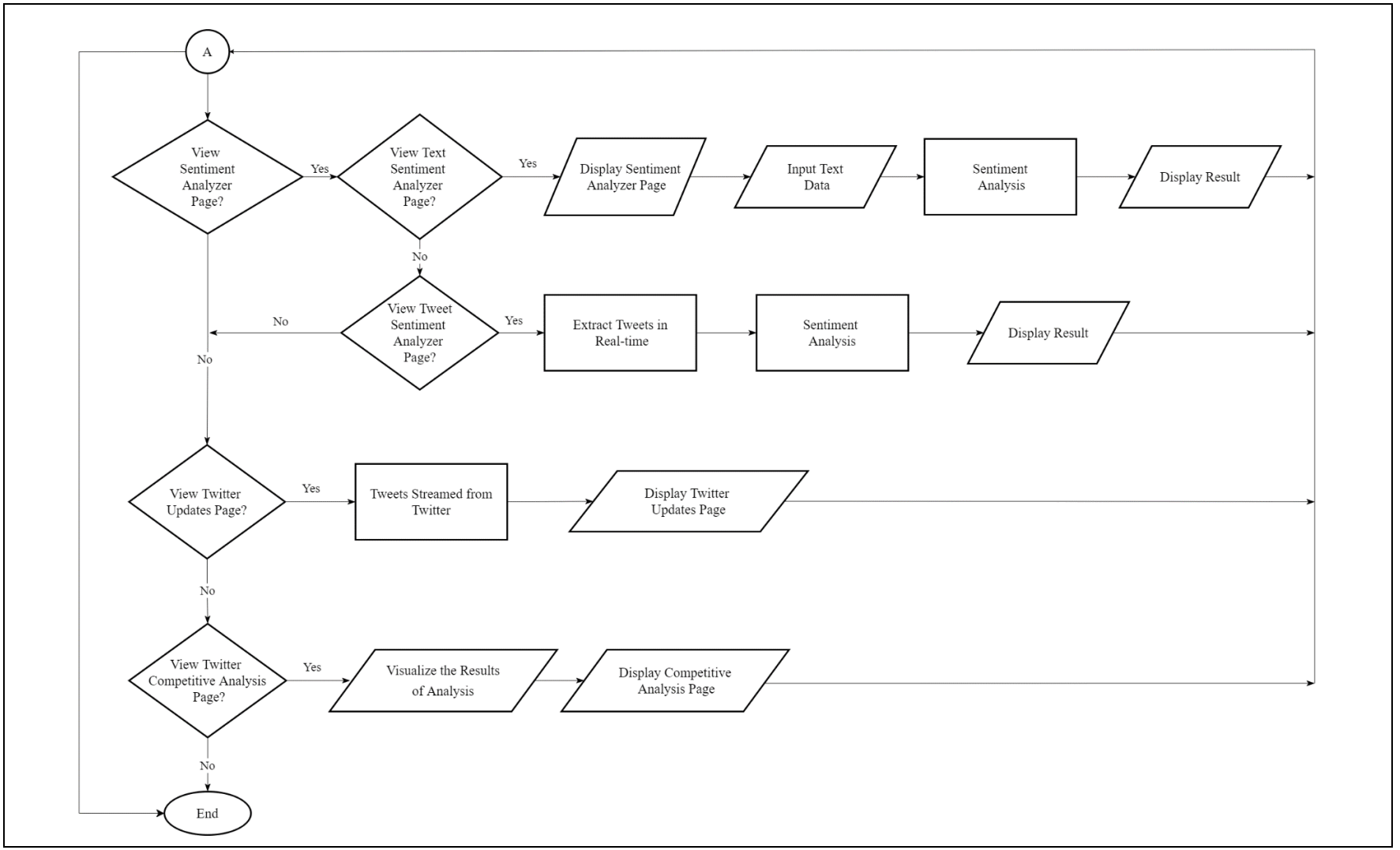


Figure 4.2 Continuation of Flowchart Diagram of the Overall System

4.1.2 Use Case Diagram

The use case diagram is used to gather system requirements, including internal and external factors. Interactions needed between the user and the system to accomplish a particular task can be represented and monitored using a use case diagram, as in Figure 4.3. Creating this diagram aims to make it possible to fully understand the steps involved in accomplishing the user's tasks.

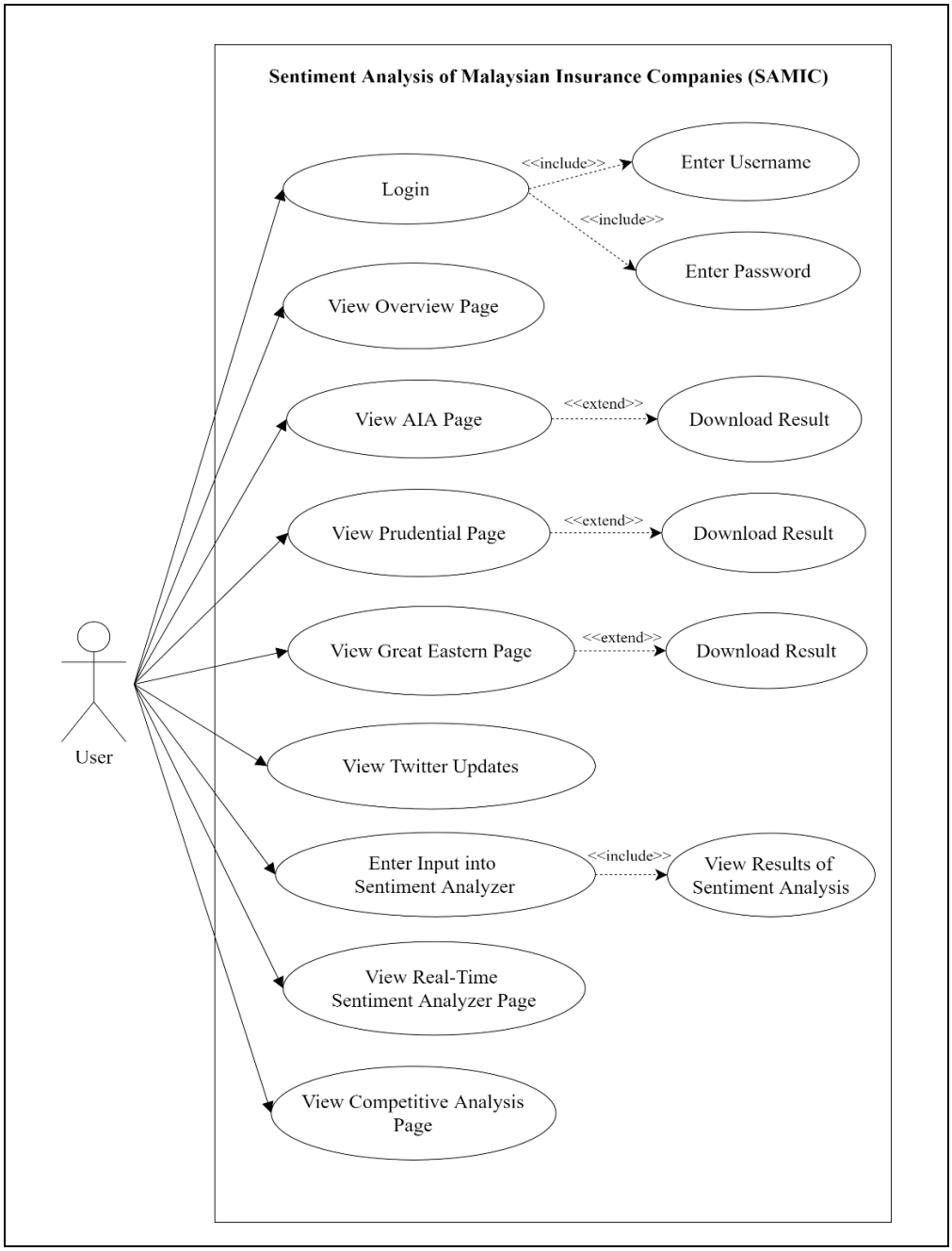


Figure 4.3 Use Case Diagram of the Application

As listed in Table 4.2, a use case description describes how users perform tasks on the system. It outlines the flow of events from the user's perspective and the system's behavior as it responds to a request.

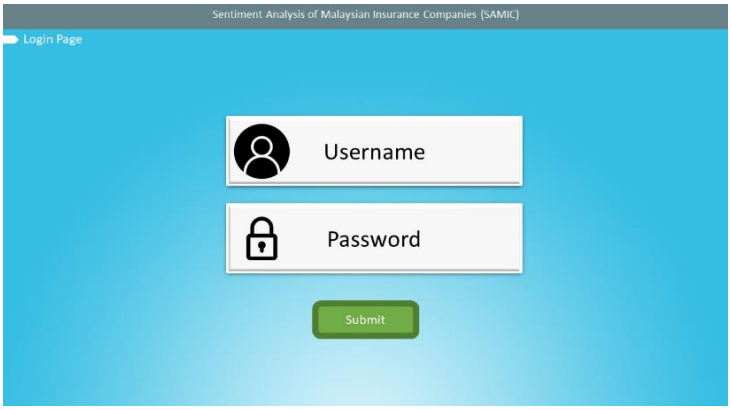
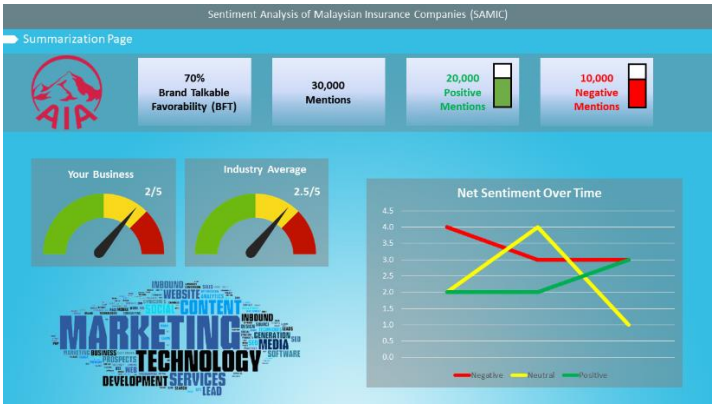
Table 4.2 Overall Use Case Description

Use Case	**Description**
Login Page	Users will be taken to the login page for admin. They will be asked to fill in their username and password to login into the system
View Overview Page	Users can access the overview page through the homepage, which displays the summary of the data analysis carried out in 2019
View AIA Page	Users are able to navigate through the AIA website, where detailed descriptions of the sentiment analysis performed on AIA data in 2019 were provided
View Prudential Page	Users are able to navigate through the Prudential website, where detailed descriptions of the sentiment analysis performed on Prudential data in 2019 were provided
View Great Eastern Page	Users are able to navigate through the Great Eastern website, where detailed descriptions of the sentiment analysis performed on Great Eastern data in 2019 were provided
View Twitter Updates	The real-time timeline of Twitter insurance updates is presented to the user
Enter Input into Sentiment Analyzer	Users can insert any text input into the sentiment analyzer
View Result of Sentiment Analysis	The result of sentiment analysis on the text entered by users are displayed
View Real-time Sentiment Analyzer Page	The result of sentiment analysis on the tweets extracted in real-time are displayed to the users
View Competitive Analysis Page	The results of the analysis are visualized to enable users to view the comparison of sentiment between companies.

4.1.3 User Interface Design

Before proceeding to the implementation phase, the user interface design, which shows the proposed system's visual components layout through which the user interacts, was completed. Table 4.3 shows the proposed user interface and a description of the features of this website.

Table 4.3 Description of the User Interface Diagrams

UI Diagram	Description
 **Figure 4.4** UI Design of the Login Page	Figure 4.4 shows the Login Page in which the users are asked to fill in their username and password to login into the system.
 **Figure 4.5** UI of the Overview Page	Figure 4.5 displays the Overview Page. Users can view the overall BTF, the most common words from negative tweets, and a summary of the sentiment analysis through data visualization.
 **Figure 4.6** UI for the AIA Company Page	Figure 4.6 shows the AIA company page. The results of the sentiment analysis are shown, including the positive and negative mentions on this page.

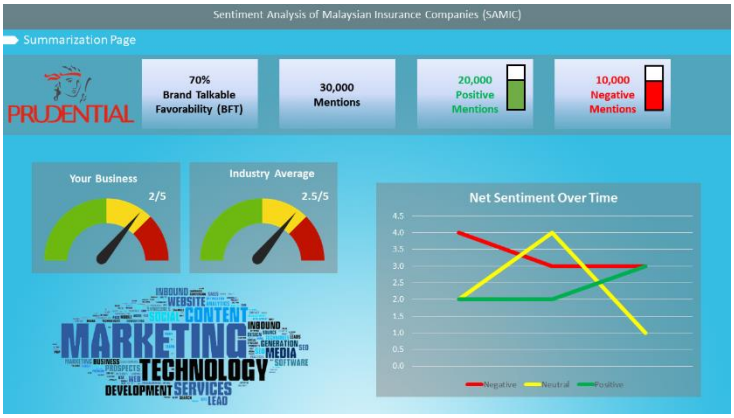


Figure 4.7 UI for the Prudential Company Page

Figure 4.7 shows the Prudential company page. The results of the sentiment analysis are shown, including the positive and negative mentions on this page.

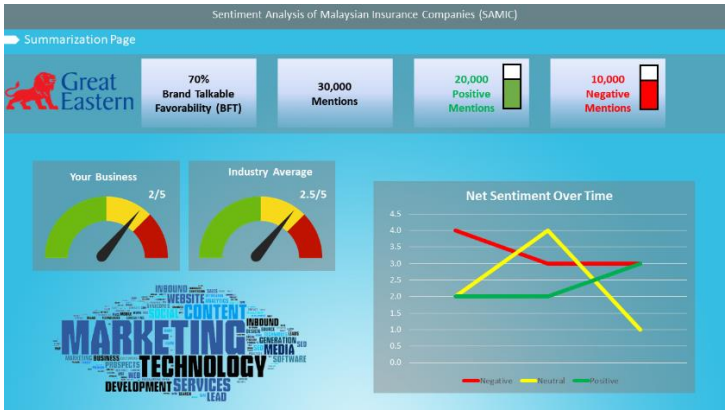


Figure 4.8 UI for the Great Eastern Company Page

Figure 4.8 shows the Great Eastern company page. The results of the sentiment analysis are shown, including the positive and negative mentions on this page.

Figure 4.9 UI of the Twitter Updates Page

Figure 4.9 shows real-time tweets updates from the Twitter page. Tweets related to insurance in Malaysia are streamed and displayed on this page.

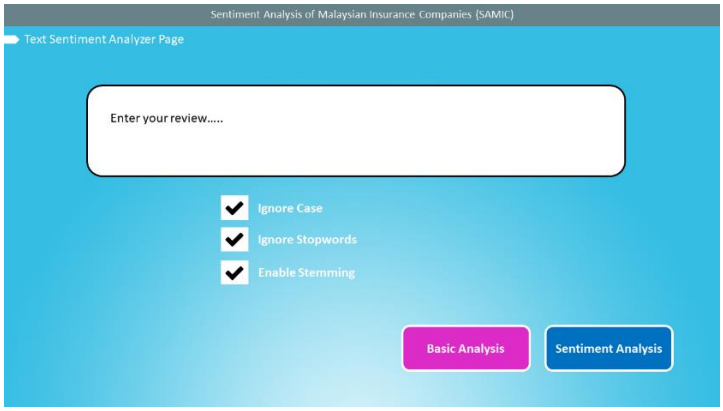


Figure 4.10 UI of the Text Sentiment Analyzer Page

Figure 4.10 displays the Text Sentiment Analyzer page. The user needs to enter the text input, and the output according to the filtered sentiment will be displayed after the user clicks the search button.

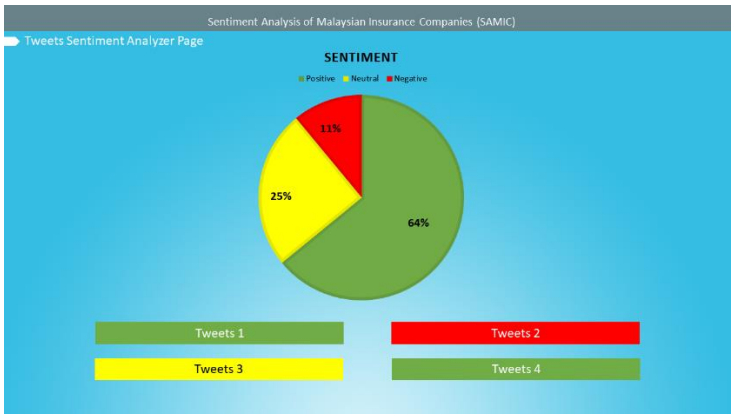


Figure 4.11 UI of the Tweet Sentiment Analyzer Page

Figure 4.11 displays the Tweet Sentiment Analyzer page. The result of sentiment analysis on the tweets extracted in real-time is displayed to users.

Figure 4.12 UI of the Competitive Analysis Page

Figure 4.12 shows the Competitive Analysis page, where the analysis results are visualized to enable users to view the comparison of sentiment between companies.

4.2 Back End Development

The development on the server-side is referred to as the back-end development. The code written in the back-end helps to transmit the underlying information to the front-end site. Python is the back-end programming language that is used for the data manipulation and development of this application. The back-end tasks that have been carried out include data preparation and the Support Vector Machine (SVM) algorithm to build the model and model deployment. The following subsections describe these processes in detail.

4.2.1 Data Preparation

Data preparation involves transforming raw data collected into a form that can be modeled using a machine learning algorithm. The raw data cannot fit into predictive modeling, and it is not possible to evaluate machine learning (ML) algorithms. It is due to the reasons such as the algorithm imposing requirements on the data to be numeric. Besides, statistical noise and errors in the data must also be corrected. As such, prior to being used to satisfy the ML model's requirements, the raw data must be prepared. Data preparation is a multi-step process that involves data collection, cleaning and pre-processing, feature engineering, and labeling. It plays a vital role in the ML model's overall quality because they rely on each other to ensure that the model performs to expectations. In the next subsection, all of the data preparation steps are discussed.

4.2.1.1 Data Collection

Two machine learning models have been built for classification purposes, namely the Malay and the English multi-class text classification model. The difference between these two models is the way the data is pre-processed. The dataset from *https://malaya.readthedocs.io/en/latest/Dataset.html* has been collected from the twitter-sentiment and polarity subfolder for training and testing the Malay model. This dataset is labeled to reflects its sentiment,

whether it is positive or negative, using a semi-supervised model. All data collected in the text corpus is in the format of javascript object notation (JSON). The Malay model's total data is balanced for the positive and negative labeled data in which both of these classes comprised 637,760 sentences each. On the other hand, 1,200000 positive and negative English sentences are collected from *http://help.sentiment140.com/for-students* to built a machine learning model based on English sentences.

Both of these collected datasets are binary classification data, which only annotates negative and positive data. Therefore, in order to add neutral data to the training and testing dataset, the neutral data is gathered from *https://www.kaggle.com/crowdflower/twitter-airline-sentiment*. Next, the neutral dataset has been translated into the Malay language as illustrated in Figure 4.13 according to how Malaya-dataset is prepared wherein Google translator is used to translate the validated English dataset to a Bahasa Melayu dataset.

```
import pandas as pd
import googletrans
from googletrans import Translator
df = pd.read_csv('C:\Users\User\Pre-processed data\neutral.csv')

translator = Translator()
translations = {}
for column in df.columns:
    unique_elements = df[column].unique()
    for element in unique_elements:
        translations[element] = translator.translate(element,  src='en',
                                                     dest='ms').text
translations
df.replace(translations, inplace = True)

df.to_csv(r'C:\Users\User\Pre-processed data\export_neutral.csv'',
          index=False)
```

Figure 4.13 Snippet of Code to Translate the Neutral Dataset

However, only 12,547 neutral sentences were discovered, a small amount compared to positive and negative annotated data because a publicly accessible neutral dataset is difficult to obtain. The total training and testing data that have

been supplemented with neutral data for the Malay model are 1,287,977 and 1,212,547 for the English model. The final data frames of the training and testing data for the Malay and English models are shown in Figure 4.14.

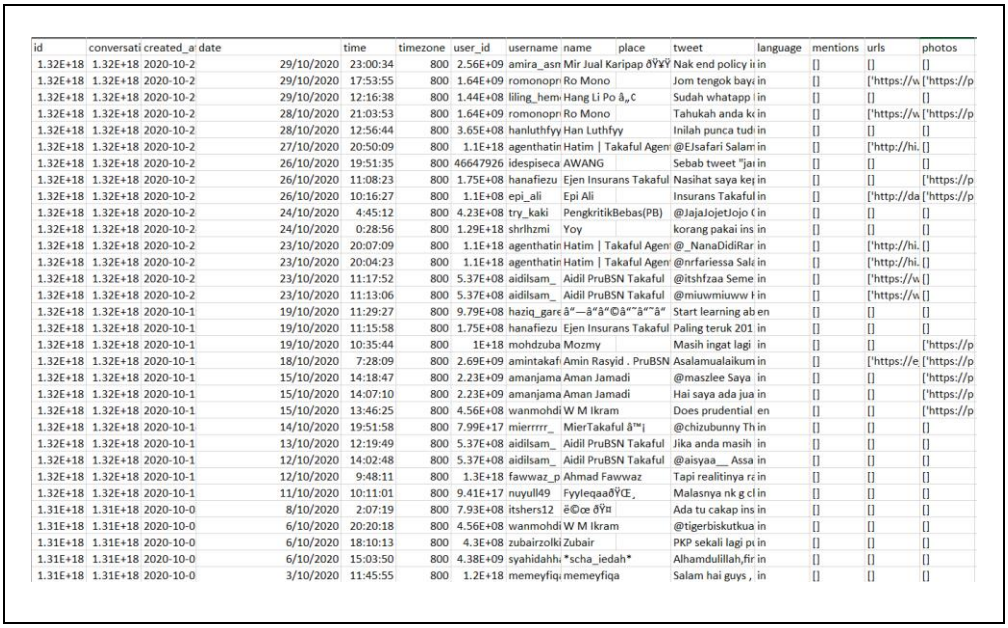
label	tweet
0	im tidak emo mulut saya terlepas!
0	Saya havent telah menonton sakit mungkin esok hari Isnin. kemas kini yang teruk.
0	Saya akan jadi jika anda datang ke vancouver.
0	idk. havent telah tidur dengan baik akhir-akhir ini. Sekiranya anda mengingatkan saya tentang sakit 10ish lakukan kemudian.
0	ya ya yesss! apabila saya mendapat ema, kerana saya miskin
0	Hai semua ! ! ! im baru ini. sesiapa sahaja mahu menunjukkan kepada saya sekitar
0	awww! drop down ke az, baik buat awak makan malam minggu yang indah! tetapi im Belanda, jadi tiada yummies Itali maaf!
0	bahawa unix program kuno lebih muda dari saya namun saya boleh lebih baik keturunannya! keluar dengan yang lama. . .
0	kepalanya terlalu menyakitkan. . . Saya fikir id lebih baik tidur. . mengambil willi untuk kecemasan appt besok berharap mereka boleh memperbaiki giginya
0	anak anjing saya sakit.

label	tweet
0	Body Of Missing Northern Calif. Girl Found: Police have found the remains of a missing Northern California girl .. http://tr.im/imji
0	@mangaaa I hope they will increase the capacity fast, yesterday was such a pain. Got the fail whale +15 times in 2 hours....
0	Behind on my classes for work
0	watching "House"
0	@kpreyes Remember my bum leg? Strikes back this time its serious
0	@paradisej cool, i will. their are all kinds of complaints about this laptop online about overheating, but no recalls
0	Emily will be glad when Mommy is done training at her new job. She misses her. http://apps.facebook.com/dogbook/profile/view/6176014
0	would rather the first party send bad messages than the 3rd party send mixed ones Sophmore year all over again?

Figure 4.14 Data Frames of Training and Testing Data

Twitter has many Malaysian users, most of whom post and write their tweets to share their opinions using the Malay language. The data was scraped from Twitter to obtain real-world data. The historical data for these three insurance companies, which are AIA, Prudential, and Great Eastern, were collected, ranging from January 1st, 2017 until October 31st, 2020. The first approach was to retrieve tweets using general keywords that directly and indirectly, referring to these insurance companies, such as ‘AIA’, ‘Prudential’, and ‘Great Eastern’. However, most tweets extracted using the first approach were not useful because they did not express any opinions, which are required for a sentiment analysis application. Alternatively, it was more practical to use the keywords of the English term ‘AIA Insurance’ and Malay term ‘AIA insurans’ and analyze the hashtags these insurance companies mentioned, which are the most mentioned Twitter topics at a specific time with many users involved. Trending hashtags of these insurance companies, including ‘#aia’, ‘#prudential’, and ‘#greateastern’ have also been scrapped. The total number of collected data is

49,523 for AIA, 17,516 for Prudential, and 38,857 for Great Eastern. All of these tweets are stored as .csv files, as in Figure 4.15.



id	conversati	created_a	date	time	timezone	user_id	username	name	place	tweet	language	mentions	urls	photos
1.32E+18	1.32E+18	2020-10-2	29/10/2020	23:00:34	800	2.56E+09	amira_asn	Mir Jual Karipap ðŸ¥Ÿ		Nak end policy i	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-2	29/10/2020	17:53:55	800	1.64E+09	romonopri	Ro Mono		Jom tengok bayi	in	[]	['https://w	['https://p
1.32E+18	1.32E+18	2020-10-2	29/10/2020	12:16:38	800	1.44E+08	liling_hem	Hang Li Po â¸Ć		Sudah whatapp	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-2	28/10/2020	21:03:53	800	1.64E+09	romonopri	Ro Mono		Tahukah anda k	in	[]	['https://w	['https://p
1.32E+18	1.32E+18	2020-10-2	28/10/2020	12:56:44	800	3.65E+08	hanluthfyy	Han Luthfyy		Inilah punca tud	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-2	27/10/2020	20:50:09	800	1.1E+18	agenthatin	Hatim | Takaful Agen		@EJsafari Salam	in	[]	['http://hi.	[]
1.32E+18	1.32E+18	2020-10-2	26/10/2020	19:51:35	800	46647926	idespiseca	AWANG		Sebab tweet "ja	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-2	26/10/2020	11:08:23	800	1.75E+08	hanafiezu	Ejen Insurans Takaful		Nasihat saya ke	in	[]	[]	['https://p
1.32E+18	1.32E+18	2020-10-2	26/10/2020	10:16:27	800	1.1E+08	epi_ali	Epi Ali		Insurans Takaful	in	[]	['http://da	['https://p
1.32E+18	1.32E+18	2020-10-2	24/10/2020	4:45:12	800	4.23E+08	try_kaki	PengkritikBebas(PB)		@JajaJojetJojo (	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-2	24/10/2020	0:28:56	800	1.29E+18	shrlhzmi	Yoy		korang pakai ins	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-2	23/10/2020	20:07:09	800	1.1E+18	agenthatin	Hatim | Takaful Agen		@_NanaDidiRar	in	[]	['http://hi.	[]
1.32E+18	1.32E+18	2020-10-2	23/10/2020	20:04:23	800	1.1E+18	agenthatin	Hatim | Takaful Agen		@nrfariessa Sal	in	[]	['http://hi.	[]
1.32E+18	1.32E+18	2020-10-2	23/10/2020	11:17:52	800	5.37E+08	aidilsam_	Aidil PruBSN Takaful		@itshfzaa Seme	in	[]	['https://w	[]
1.32E+18	1.32E+18	2020-10-2	23/10/2020	11:13:06	800	5.37E+08	aidilsam_	Aidil PruBSN Takaful		@miuwmiuww I	in	[]	['https://w	[]
1.32E+18	1.32E+18	2020-10-1	19/10/2020	11:29:27	800	9.79E+08	haziq_gare	â"—â"©â"˜â"˜â"		Start learning ab	en	[]	[]	[]
1.32E+18	1.32E+18	2020-10-1	19/10/2020	11:15:58	800	1.75E+08	hanafiezu	Ejen Insurans Takaful		Paling teruk 201	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-1	19/10/2020	10:35:44	800	1E+18	mohdzuba	Mozmy		Masih ingat lagi	in	[]	[]	['https://p
1.32E+18	1.32E+18	2020-10-1	18/10/2020	7:28:09	800	2.69E+09	amintakafi	Amin Rasyid . PruBSN		Asalamualaikum	in	[]	['https://e	['https://p
1.32E+18	1.32E+18	2020-10-1	15/10/2020	14:18:47	800	2.23E+09	amanjama	Aman Jamadi		@maszlee Saya	in	[]	[]	['https://p
1.32E+18	1.32E+18	2020-10-1	15/10/2020	14:07:10	800	2.23E+09	amanjama	Aman Jamadi		Hai saya ada jua	in	[]	[]	['https://p
1.32E+18	1.32E+18	2020-10-1	15/10/2020	13:46:25	800	4.56E+08	wanmohdi	W M Ikram		Does prudential	en	[]	[]	['https://p
1.32E+18	1.32E+18	2020-10-1	14/10/2020	19:51:58	800	7.99E+17	mierrrrr_	MierTakaful â™¡		@chizubunny Th	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-1	13/10/2020	12:19:49	800	5.37E+08	aidilsam_	Aidil PruBSN Takaful		Jika anda masih	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-1	12/10/2020	14:02:48	800	5.37E+08	aidilsam_	Aidil PruBSN Takaful		@aisyaa__ Assa	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-1	12/10/2020	9:48:11	800	1.3E+18	fawwaz_p	Ahmad Fawwaz		Tapi realitinya r	in	[]	[]	[]
1.32E+18	1.32E+18	2020-10-1	11/10/2020	10:11:01	800	9.41E+17	nuyull49	FyyIeqaaðŸŒ¸		Malasnya nk g cl	in	[]	[]	[]
1.31E+18	1.31E+18	2020-10-0	8/10/2020	2:07:19	800	7.93E+08	itshers12	ë©œ ðŸ¤		Ada tu cakap ins	in	[]	[]	[]
1.31E+18	1.31E+18	2020-10-0	6/10/2020	20:20:18	800	4.56E+08	wanmohdi	W M Ikram		@tigerbiskutkua	in	[]	[]	[]
1.31E+18	1.31E+18	2020-10-0	6/10/2020	18:10:13	800	4.3E+08	zubairzolki	Zubair		PKP sekali lagi p	in	[]	[]	[]
1.31E+18	1.31E+18	2020-10-0	6/10/2020	15:03:50	800	4.38E+09	syahidahh	*scha_iedah*		Alhamdulillah,fir	in	[]	[]	[]
1.31E+18	1.31E+18	2020-10-0	3/10/2020	11:45:55	800	1.2E+18	memeyfiq	memeyfiqa		Salam hai guys ,	in	[]	[]	[]

Figure 4.15 Scraped Tweets File

The collection of real-world datasets also involves extracting tweets in real-time and the insurance agents' profiles who persistently promote their respective insurance companies using keywords 'Saya agent AIA', 'Saya agent Prudential' and 'Saya agent Great Eastern' since January 1st, 2019 to October 31st, 2020. Snscrape is used to scrape the link of the Twitter ID from their profile. The Twitter ID is a unique value of a Twitter account, and no two accounts can have the same ID. Although an account can change its username handle, it is not possible to change its Twitter ID. The 329 unique Twitter ID is then passed to Tweepy, a Python library to access Twitter API and scrap the user profile as depicted in Figure 4.16.

```
def fetch_tw(ids):
    list_of_tw_status = api.statuses_lookup(ids, tweet_mode= "extended")
    empty_data = pd.DataFrame()
    for status in list_of_tw_status:
            tweet_elem = {"tweet_id": status.id,
                          "screen_name": status.user.screen_name,
                          "tweet":status.full_text,
                           "date":status.created_at,
                           "location":status.user.location,
                           "followers_count":status.user.followers_count,
                           "following_count":status.user.friends_count,
                           "favourites_count":status.user.favourites_count,
                           "tweet_count":status.user.statuses_count,
                           "retweet_count":status.retweet_count,
                           "created_at":status.user.created_at,
                           "image_url":status.user.profile_image_url_https}
            empty_data = empty_data.append(tweet_elem, ignore_index = True)
    empty_data.to_csv("agents_tweets.csv", mode="a")
```

```
for i in range(chunks):
        batch = ids[i*50:(i+1)*50]
        result = fetch_tw(batch)
```

Figure 4.16 Snippet of Code to Extract User Profile

Tweets have been extracted through the application framework to be analyzed for real-time sentiment analysis using the Twitter API, as in Figure 4.17. Apart from the tweets, the tweets' username and date are extracted, as well as the Twitter ID that enables the user of the application to reply to the tweet and view the tweeted user's profile. This insight helps resolve this project's second problem statement, which is to serve customer service better since this information will be used directly to address future problems or issues relating to the claims process or policy provisions.

```
def get_tweets(self):
    tweets = []
    try:
        recd_tweets = self.api.search(q=self.query,
                                      tweet_mode='extended',
                                      count=self.tweet_count_max,
                                      )

        if not recd_tweets:
            pass
        for tweet in recd_tweets:
            parsed_tweet = {}

            parsed_tweet['text'] = tweet.full_text
            parsed_tweet['user'] = tweet.user.screen_name
            parsed_tweet['id'] = tweet.id_str
            parsed_tweet['date'] = tweet.created_at

            if self.with_sentiment == 1:
                parsed_tweet['sentiment'] = self.get_tweet_sentiment(tweet.full_text)
            else:
                parsed_tweet['sentiment'] = 'unavailable'

            if tweet.retweet_count > 0 and self.retweets_only == 1:
                if parsed_tweet not in tweets:
                    tweets.append(parsed_tweet)
            elif not self.retweets_only:
                if parsed_tweet not in tweets:
                    tweets.append(parsed_tweet)

        return tweets
```

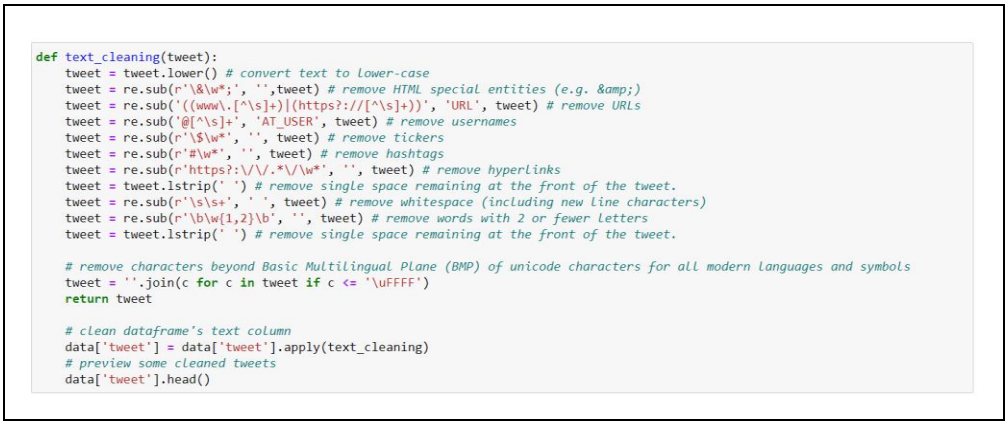
Figure 4.17 Snippet of Code to Extract Real-time Tweets

4.2.1.2 Text Pre-Processing and Normalization

Machine Learning requires text data in a numerical form. Encoding techniques such as BagOfWord, Bi-gram, n-gram, TF-IDF, and Word2Vec have been used to encode the real-world data collected from Twitter into a numeric vector. However, before encoding text data, it needs to be cleaned first, and this process of preparing text data is called text pre-processing. This is the very first step in natural language processing (NLP). Text pre-processing is a method of cleaning and preparing text data for use in a particular context. The NLP sentiment extraction task needs to eliminate any unnecessary features in the data, which would make the final trained model a weak generalizer.

NLTK and re are two of the Python libraries used in this project to perform text pre-processing tasks. Any unimportant columns for the ML model have been

removed before passing the data that needs cleaning to the function of code. The final dataset comprises four columns, which are date, username, language, and tweet. Text cleaning has been done on the dataset, and it involves removing unwanted characters like emojis and properly formatting the text to remove any extra spaces. All characters were converted to a lowercase beforehand to avoid any case-sensitive operation. Terms such as mentions, hashtags, links, and usernames have been removed using patterns that can match the desired terms with a regular expression (Regex). Regex is a special string that contains a pattern that can match words associated with the pattern. A function was created using Regex to perform a bulk pattern formatting, including searching and removing Regex for every tweet in the dataset. Some tweets might also contain a Unicode character that cannot be read in an ASCII format. Mostly, these characters are used for emojis and non-ASCII characters. All of these characters have been removed as well. Duplicate tweets and null values have been dropped as they do not contribute useful information to the dataset and computationally inefficient. Figure 4.18 shows a snippet of code how the text was cleaned, and this code is also used to clean real-time data.



```
def text_cleaning(tweet):
    tweet = tweet.lower() # convert text to lower-case
    tweet = re.sub(r'\&\w*;', '',tweet) # remove HTML special entities (e.g. &amp;)
    tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))', 'URL', tweet) # remove URLs
    tweet = re.sub('@[^\s]+', 'AT_USER', tweet) # remove usernames
    tweet = re.sub(r'\$\w*', '', tweet) # remove tickers
    tweet = re.sub(r'#\w*', '', tweet) # remove hashtags
    tweet = re.sub(r'https?:\/\/.*\/\w*', '', tweet) # remove hyperlinks
    tweet = tweet.lstrip(' ') # remove single space remaining at the front of the tweet.
    tweet = re.sub(r'\s\s+', ' ', tweet) # remove whitespace (including new line characters)
    tweet = re.sub(r'\b\w{1,2}\b', '', tweet) # remove words with 2 or fewer letters
    tweet = tweet.lstrip(' ') # remove single space remaining at the front of the tweet.

    # remove characters beyond Basic Multilingual Plane (BMP) of unicode characters for all modern languages and symbols
    tweet = ''.join(c for c in tweet if c <= '\uFFFF')
    return tweet

    # clean dataframe's text column
    data['tweet'] = data['tweet'].apply(text_cleaning)
    # preview some cleaned tweets
    data['tweet'].head()
```

Figure 4.18 Snippet of Code to Clean the Text Data

Training data has now been converted into an even leaner body of text that will be much easier to extract features. However, some terms in the dataset are typical in human language and are often used in compositions of sentences but were better left out as they do not add any valuable features to the model. These terms are widely known as stop words in NLP, such as ‘the’, ‘a’, ‘me’, ‘is’, ‘to’, and ‘all’. For the English model, the natural language toolkit (NLTK)

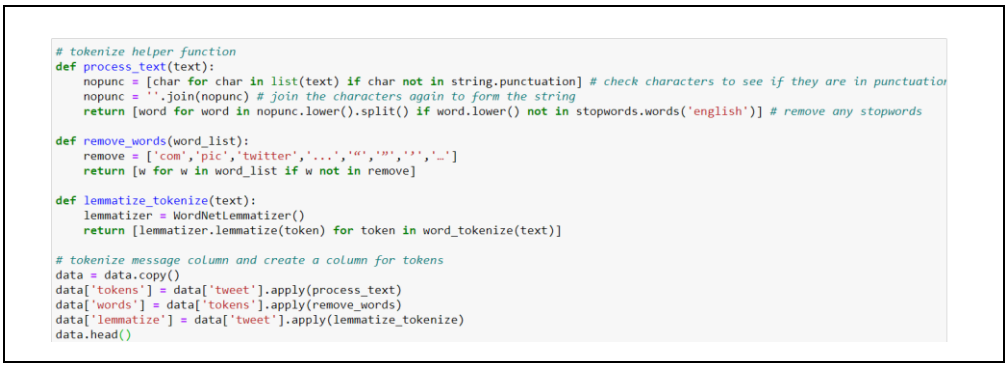
library comes with a pre-built function that can strip these terms from the dataset. On the contrary, the 475 Malay stop-words lists are imported from *https://github.com/stopwords-iso/stopwords-ms/blob/master/stopwords-ms.txt* for the Malay model, as illustrated in Figure 4.19.

```
with open('stopwords-ms.txt','r') as f2:
    b=f2.read().split()
    print(' '.join(x.lower() for x in b))
```

```
abdul abdullah acara ada adalah ahmad air akan akhbar akhir aktiviti alam amat amerika anak anggota antara antarabangsa apa apa
bila april as asas asean asia asing atas atau australia awal awam bagaimanapun bagi bahagian bahan baharu bahawa baik bandar ba
nk banyak barangan baru baru-baru bawah beberapa bekas beliau belum berada berakhir berbanding berdasarkan berharap berikutan b
erjaya berjumlah berkaitan berkata berkenaan berlaku bermula bernama bernilai bersama berubah besar bhd bidang bilion bn boleh
bukan bulan bursa cadangan china dagangan dalam dan dana dapat dari daripada dasar datang datuk demikian dengan depan derivativ
es dewan di diadakan dibuka dicatatkan dijangka diniagakan dis disember ditutup dolar dr dua dunia ekonomi eksekutif eksport em
pat enam faedah feb global hadapan hanya harga hari hasil hingga hubungan ia iaitu ialah indeks india indonesia industri ini is
lam isnin isu itu jabatan jalan jan jawatan jawatankuasa jepun jika jualan juga julai jumaat jumlah jun juta kadar kalangan kal
i kami kata katanya kaunter kawasan ke keadaan kecil kedua kedua-dua kedudukan kekal kementerian kemudahan kenaikan kenyataan k
epada kepentingan keputusan kerajaan kerana kereta kerja kerjasama kes keselamatan keseluruhan kesihatan ketika ketua keuntunga
n kewangan khamis kini kira-kira kita klci klibor komposit kontrak kos kuala kuasa kukuh kumpulan lagi lain langkah laporan leb
ih lepas lima lot luar lumpur mac mahkamah mahu majlis makanan maklumat malam malaysia mana manakala masa masalah masih masing-
masing masyarakat mata media mei melalui melihat memandangkan memastikan membantu membawa memberi memberikan membolehkan membua
t mempunyai menambah menarik menawarkan mencapai mencatatkan mendapat mendapatkan menerima menerusi mengadakan mengambil mengen
ai menggalakkan menggunakan mengikut mengumumkan mengurangkan meningkat meningkatkan menjadi menjelang menokok menteri menunjuk
kan menurut menyaksikan menyediakan mereka merosot merupakan mesyuarat minat minggu minyak modal mohd mudah mungkin naik najib
nasional negara negara-negara negeri niaga nilai nov ogos okt oleh operasi orang pada pagi paling pameran papan para paras parl
imen parti pasaran pasukan pegawai pejabat pekerja pelabur pelaburan pelancongan pelanggan pelbagai peluang pembangunan pemberi
ta pembinaan pemimpin pendapatan pendidikan penduduk penerbangan pengarah pengeluaran pengerusi pengguna pengurusan peniaga pen
ingkatan penting peratus perdagangan perdana peringkat perjanjian perkara perkhidmatan perladangan perlu permintaan perniagaan
persekutuan persidangan pertama pertubuhan pertumbuhan perusahaan peserta petang pihak pilihan pinjaman polis politik presiden
prestasi produk program projek proses proton pukul pula pusat rabu rakan rakyat ramai rantau raya rendah ringgit rumah sabah sa
haja saham sama sarawak satu sawit saya sdn sebagai sebahagian sebanyak sebarang sebelum sebelumnya sebuah secara sedang segi s
ehingga sejak sekarang sektor sekuriti selain selama selasa selatan selepas seluruh semakin semalam semasa sementara semua semu
la sen sendiri seorang sepanjang seperti sept september serantau seri serta sesi setiap setiausaha sidang singapura sini sistem
sokongan sri sudah sukan suku sumber supaya susut syarikat syed tahap tahun tan tanah tanpa tawaran teknologi telah tempat temp
atan tempoh tenaga tengah tentang terbaik terbang terbesar terbuka terdapat terhadap termasuk tersebut terus tetapi thailand ti
ada tidak tiga timbalan timur tindakan tinggi tun tunai turun turut umno unit untuk untung urus usaha utama walaupun wang wanit
```

Figure 4.19 The List of Stop Words of the Malay Model

All the sentences in the dataset were split to get individual tokens, which is a list of words per sentence produced in a previously processed tweet after eliminating stop words. There are two new columns added in the data frame that includes these tokenized tweets. Moreover, an additional normalization technique called lemmatizing that shortened words to their stem words has been applied to the English model as in Figure 4.20.



```
# tokenize helper function
def process_text(text):
    nopunc = [char for char in list(text) if char not in string.punctuation] # check characters to see if they are in punctuatio
    nopunc = ''.join(nopunc) # join the characters again to form the string
    return [word for word in nopunc.lower().split() if word.lower() not in stopwords.words('english')] # remove any stopwords

def remove_words(word_list):
    remove = ['com','pic','twitter','...','“','”','’','…']
    return [w for w in word_list if w not in remove]

def lemmatize_tokenize(text):
    lemmatizer = WordNetLemmatizer()
    return [lemmatizer.lemmatize(token) for token in word_tokenize(text)]

# tokenize message column and create a column for tokens
data = data.copy()
data['tokens'] = data['tweet'].apply(process_text)
data['words'] = data['tokens'].apply(remove_words)
data['lemmatize'] = data['tweet'].apply(lemmatize_tokenize)
data.head()
```

Figure 4.20 Snippet of Code for Text Tokenization and Stopwords Removal

After this pre-processing step, the dataset was saved to the working directory and ready to be fed right to the machine learning (ML) algorithm to extract meaningful information that can be used to classify unlabeled data. Predictive modeling is discussed in the next section.

4.2.2 Predictive Modeling

The model evaluation is conducted using the support vector machine (SVM) classifier, in which the algorithm is used to perform the dataset classification**.** SVM is a learning machine for two-group classification problems developed natively for binary classification problems and cannot be used directly in this project that has multi-class classification problems. Hence, One-vs-Rest (OvR) is incorporated into the classifier. OVR is a multi-class classification method that uses multiple binary SVM classifiers to obtain a multi-class prediction. It involves the splitting of the multi-class dataset into multiple binary classification problems. A binary classifier would then be trained on each binary classification, requiring each model to predict the probability of class membership or a probability-like score. The Argmax of these scores, which is the class index with the highest score, is used to predict the class label, and predictions are made using the most confident model.

The initial step in predictive modeling is to prepare text data that draw features for the machine learning (ML) algorithm to obtain effective predictive modeling. A text representation that represents the occurrence of words inside a document is a bag-of-words. This concept was used as a weighting scheme to transform the whole corpus list of tokens into a matrix-vector that the algorithm can understand and learn. It requires a vocabulary of known words and a measure of known terms to represent all the words in model vocabulary as a list of tokens in this project's context. The SciKit Learn's CountVectorizer function converted a collection of tweets to a token count matrix. As the corpus of text contains millions of tweets, there would be several zero counts for every word in the corpus.

The next step is to construct document vectors, as indicated in Figure 4.21, to score the frequency of words in each document word with TF-IDF, an abbreviation for Term Frequency Inverse Document Frequency. This weight is a statistical measure used in a text or a collection corpus to evaluate the significance of a word, which rises in proportion to the number of times a word appears in the document but is offset by the word's frequency in the corpus. The idea is to weigh down the frequent words while scaling up the uncommon ones. The higher the TF-IDF score, the lesser the term appears and vice-versa.

```
from sklearn.feature_extraction.text import TfidfTransformer
tfidf_transformer = TfidfTransformer().fit(messages_bow)
tfidf_sample = tfidf_transformer.transform(bow_sample)
print(tfidf_sample)
```

```
  (0, 839717)   0.49003299830673713
  (0, 169429)   0.7810424873705181
  (0, 67737)    0.387092099496468
```

```
# to transform the entire bag-of-words corpus
tweet_tfidf = tfidf_transformer.transform(messages_bow)
print(tweet_tfidf.shape)
```

```
(1423924, 984522)
```

Figure 4.21 Snippet of Code to Construct Document Vectors

The TF-IDF weight is composed of two terms, the first of which measures the normalized Term Frequency (TF) to indicate how frequently a term appears in a document. In the context of natural language processing (NLP), terms refer to words or phrases. However, terms could also correspond to any token in the text. Since documents can have different lengths, a term may likely appear more frequently in longer documents than in shorter ones. Thus, it would seem like a term is more significant to a longer document than a shorter one. In order to minimize this effect, the term frequency must be divided by the total number of terms in the document as a means of normalization, indicated by the following formula to calculate TF in Eq. (2)

$$\mathrm{TF}(t) = \frac{(Number\ of\ times\ term\ t\ appears\ in\ a\ document)}{(Total\ number\ of\ terms\ in\ the\ document)} \quad (2)$$

The second term is the inverse document frequency (IDF) that measures how much information the word provides, whether it is common or rare in all documents, represented by the weight of how often a word is used. The weight

is the inverse fraction of the documents containing the term and is logarithmically scaled. It was computed as in Eq. (3) by dividing the total number of documents set by the number of documents containing the term and then using that quotient logarithm. The more frequently it is used across documents, the lower its score. For instance, the word ‘a’ would appear in almost all English texts and have a low inverse document.

$$\mathrm{IDF(t)} = \log\left[\frac{(Total\ number\ of\ document\ set)}{\text{Term Frequency with term } t\ (\mathrm{TF}(t))}\right]+1 \quad (3)$$

TF-IDF is the product of the term TF and IDF scores. By using the Scikit-learn TfidfTransformer module, the formula used to compute the TF-IDF for the term t of document d in a document set is shown in Eq.(4)

$$\text{TF-IDF}(t, d) = \mathrm{TF}(t) \times \mathrm{IDF}(t) \quad (4)$$

The final step before training the model is vector normalization. Scikit-learn uses the ‘l2’ normalization technique for each document to normalize the unit-length vectors and abstract the original document's length. Each vector would have as many dimensions as unique words in the Twitter corpus. The model is built using an SVM classifier by employing Scikit learn LinearSVC since the linear kernel can be scaled to a large number of samples. Besides, a multi-class model is implemented by using the ‘Sklearn.multi-class.OneVsRestClassifier’ wrapper. Scikit Learn provides a pipeline capability that facilitates the concept of a pipeline workflow, which takes all of the above text pre-processing steps and includes them in the classifier and grid search parameters. The pipeline was used to cross-validate the model workflow and, at the same time, choosing the best parameter configuration using GridSearchCV by iterating over ParameterGrid with hyperparameter configuration values.

Besides, using the KFold technique, different hyperparameter configurations have been tested to split the model into random parts to find out whether it generalizes well. There are 16 tested parameter configurations and 10 KFold validations of the model. Hence, the model was trained and tested on the

validation set 160 times, meaning that a combination of 16 parameters are passed to the GridsearchCV object and ten folds for the cross-validation, indicating that for every parameter combination, the grid runs ten different iterations with a different test set every time. The data were separated into two sets beforehand using the holdout method of 80:20 ratio training and testing sets, indicating 1,030382 tweets in the training set and 237,152 tweets in the testing set. After testing the different model parameter combinations, GridsearchCV returns the best performing model to classify real-world data. Figure 4.22 shows the snippet of code to train the classifier model.

```
from imblearn.pipeline import Pipeline
from sklearn.svm import LinearSVC
from sklearn.multiclass import OneVsRestClassifier

# Run Train Data Through Pipeline analyzer=text_process
X_train, X_test, y_train, y_test = train_test_split(data['tweet'], data['label'], test_size=0.2)

# create pipeline
pipeline = Pipeline([
    ('bow', CountVectorizer(strip_accents='ascii',
                            lowercase=True)),
    ('tfidf', TfidfTransformer()),
    ('clf', OneVsRestClassifier(LinearSVC(class_weight='balanced', max_iter=100000)))
])

# this is where we define the values for GridSearchCV to iterate over
parameters = [
            {
                'bow__ngram_range': [(1,1), (1, 2)],
                'tfidf__use_idf': (True, False),
                'clf__estimator__loss' : ["hinge", "squared_hinge"],
                'clf__estimator__tol' : [1e-4, 1e-5],
                'clf__estimator__penalty' : ["l2"]
            },
            ]

# do 10-fold cross validation for each of the 6 possible combinations of the above params
grid = GridSearchCV(pipeline, cv=10, param_grid=parameters, verbose=1, n_jobs= 8)
grid.fit(X_train,y_train)

# summarize results
print("\nBest Model: %f using %s" % (grid.best_score_, grid.best_params_))
print('\n')
means = grid.cv_results_['mean_test_score']
stds = grid.cv_results_['std_test_score']
params = grid.cv_results_['params']
for mean, stdev, param in zip(means, stds, params):
    print("Mean: %f Stdev:(%f) with: %r" % (mean, stdev, param))
```

Figure 4.22 Snippet of Code to Train the Classifier Model

Saving the model is a mandatory step for the implementation of the model in a web application. The model was saved as a pickled file in the working directory using ‘Joblib’, enabling the trained model to be retrieved in the future without retraining it by simply dump the model and load it in the current directory. The next step in this project, model deployment, will be explained in the next section.

4.2.3 Model Deployment

Model deployment is one of the last stages of the machine learning life cycle. Predictive analytics based on trained data is carried out by using the model that has been developed. At the model development stage, the model took the data frames of tweets and returned the predicted classified sentiment label as ‘positive’, ‘neutral’, and ‘negative’ represented as ‘0’, ‘2’, and ‘4’ respectively. These data frames are sorted by date and afterwards integrated into the production environment. In the implementation context, the model results are presented using the data visualization technique via a web service by displaying overall sentiment scores that were produced offline across the batch process. Prior to deployment, the model is tested to ensure that the test input set identified during development produces validated results by setting the production environment in accordance with the development environment. This was done by defining the environment specifics such as specific language versions and library dependencies for the model in Python requirements.txt file.

After sentiment predictions are made using the model classifier on the data obtained, and its performance is evaluated, the data is visualized using Plotly, an open-source interactive graphics library for Python. The data is first imported into the Pandas data frames in Python. Interactive charts are then be generated. Additional information on the results of the analysis will be displayed through hover events. The analysis results are used to create an interactive visualization tool to present real-world data analysis outcomes.

A Wordcloud visual representation of the text data is displayed in each of the proposed system’s insurance company pages**.** A list of words is shown in the Wordcloud, with each word’s significance with font size or color. It indicates how frequently words related to the insurance companies appeared by making each word’s size proportional to its frequency. These words are arranged in a cloud of words since this format is useful to perceive the most prominent terms

for insurance companies quickly. In the next section, the development of the front-end application is explained.

4.3 Front End Development

Front end development refers to the development of the 'client-side', where the emphasis is on what users see in their browser or application visually. In this project, HTML, CSS, Javascript, and Jquery are the front-end languages used to integrate Flask's back-end framework. Flask is used for the back-end but utilizes a template language named Jinja2 to create HTML markup formats that are returned via an HTTP request to the user. In a Python web application environment, the Flask web framework, along with the data visualization tools in Python, are used to construct custom plots and charts after the data manipulation phases to leverage the power of both front-end and back-end development. The workflows in the application that the user will interact with are illustrated for each of the interfaces. For each feature of the completed system's interface design, six subsections will be discussed, including overview page, data dashboard, Twitter updates, sentiment analyzer, competitive analysis, and insurance agents' performance.

4.3.1 Overview Page

Figure 4.23 shows the interface design of the Login page for admin. Admin is asked to fill in the username and password that has been set for the registered admin. Once the user clicks the login button, the username and password are authenticated by the application. If the username and password do not exist in the system, access will be denied, and they will be asked to authenticate their credentials again.

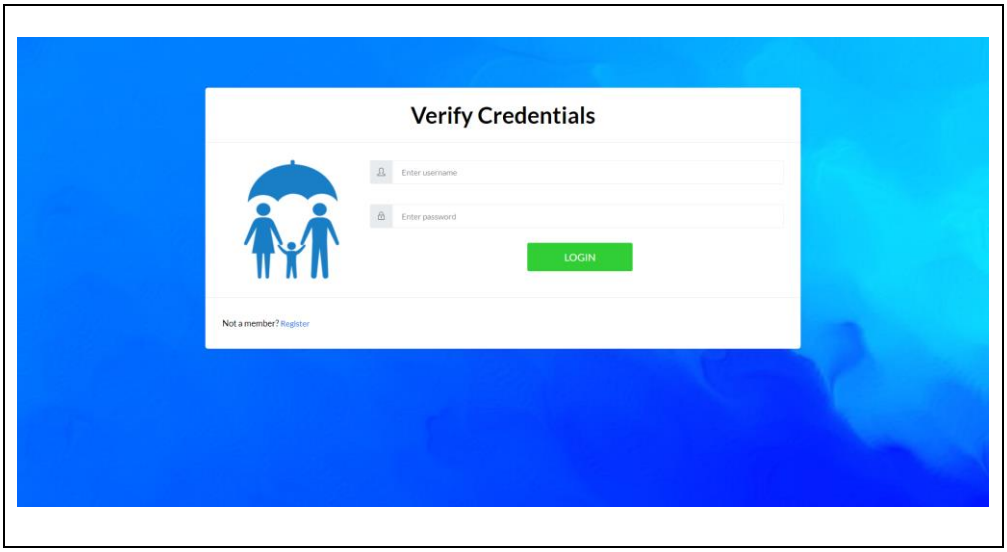


Figure 4.23 Interface Design of the Login Page

The Overview page is presented to the user once they have logged into the application, as in Figure 4.24. The summary analysis includes Brand Talkable Favorability (BFT) and visualization charts, such as the companies' bar chart and word clouds. The admin can roll over or select one of the drop-down menu options that contain links to other sections of the application using the navigation bar. The navigation bar is part of the whole application's design, meaning that it is shown on all the application pages.

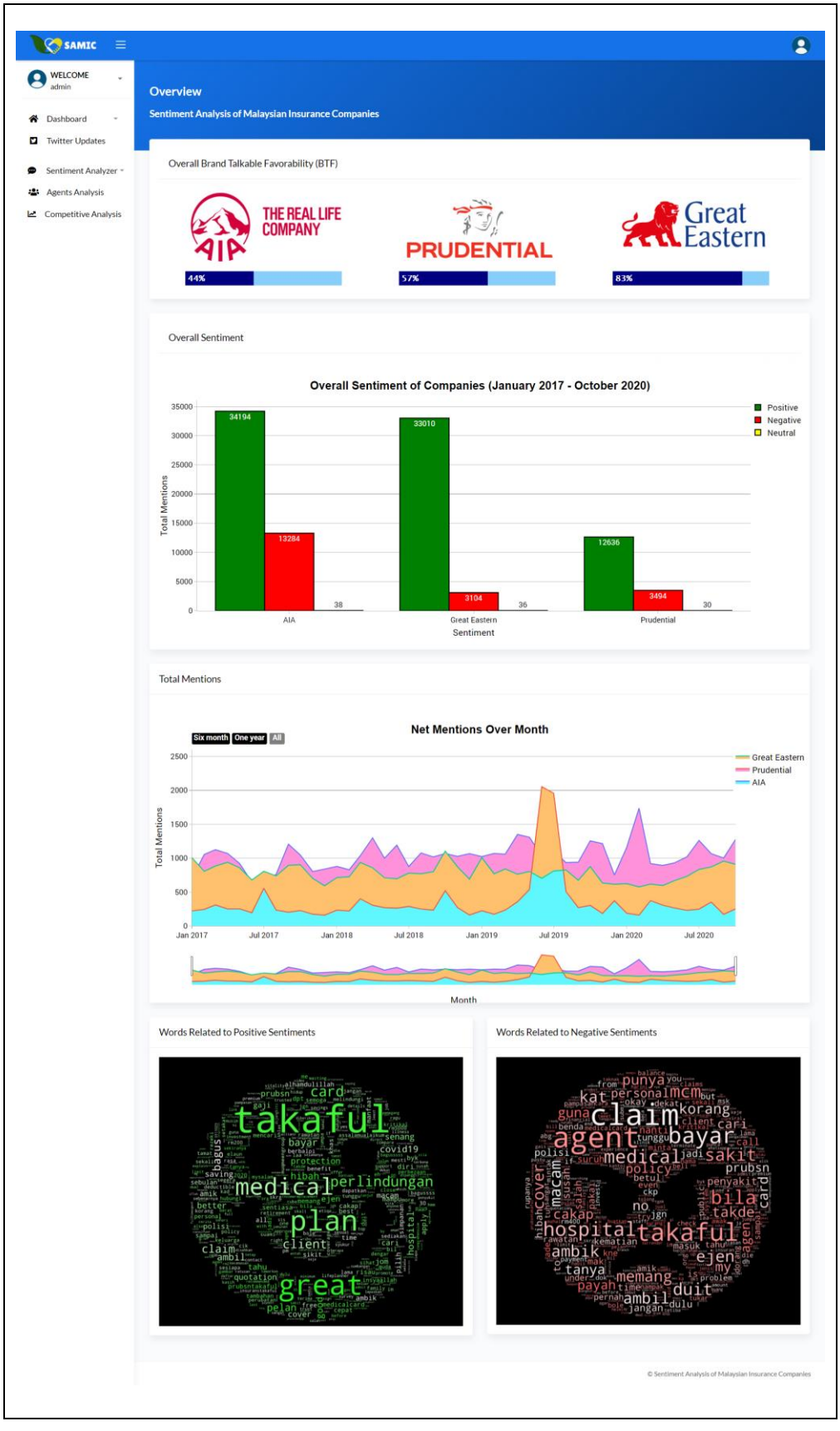


Figure 4.24 Overview Page of the Application

4.3.2 Data Dashboard

The dashboard provides an easily accessible reporting feature so that the user can instantly see and understand trends and patterns. If the admin clicks on the navigation bar's dashboard, they could choose to view one of the three companies' data dashboards. The dashboard page interface for AIA, Prudential, and Great Eastern is shown in Table 4.4. The dashboards visualized Twitter's historical tweets in which the sentiment was classified. The users can download the spreadsheet file of the labeled sentiment for each of the companies on the data dashboard, and they can choose to display the visualized data in time series or bar chart. The overall level of sentiment, whether positive, neutral, or negative, was calculated and shown as a gauge chart. More information on total tweets classified as positive, neutral, and negative was provided in the sentiment breakdown pie chart. Besides, the related positive and negative words from the text corpus were displayed as well.

Table 4.4 User Interface Dashboard of the Companies

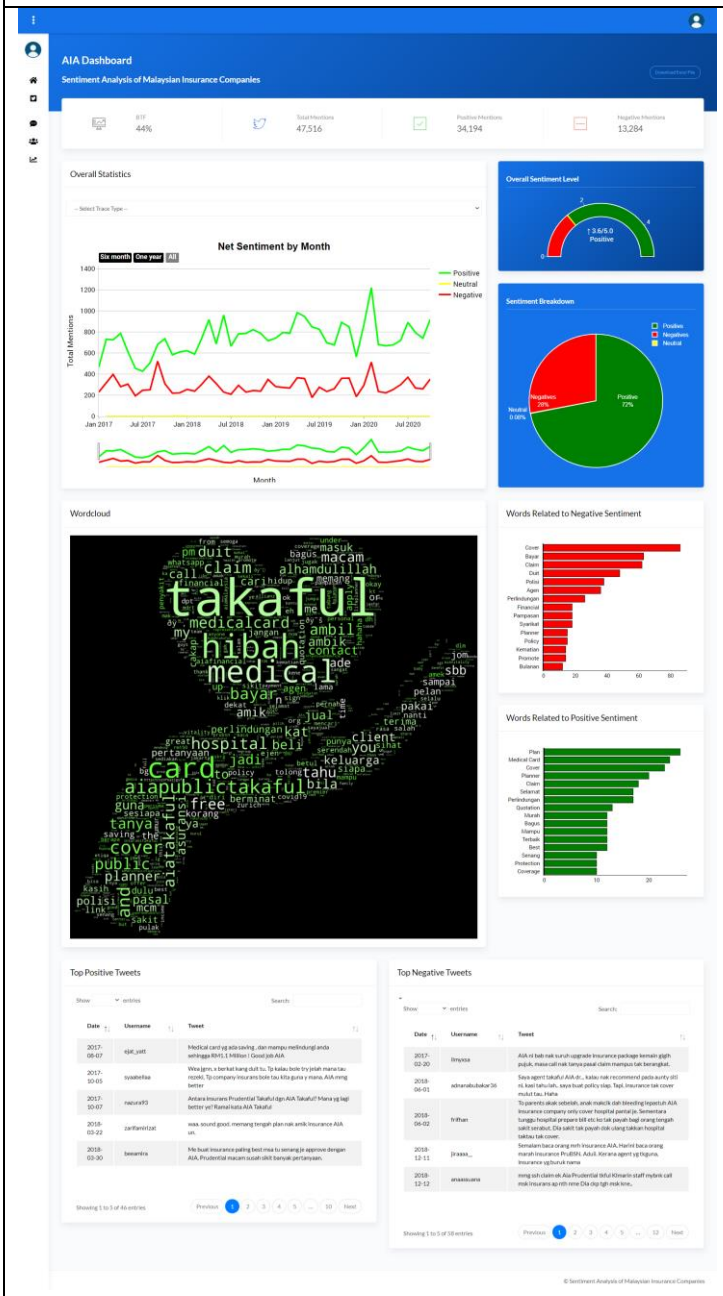
UI Diagram	Description
	Figure 4.25 shows the application interface of the AIA page. The results of the analysis are presented to the users using data visualization techniques.

Figure 4.25 Data Dashboard for AIA Page

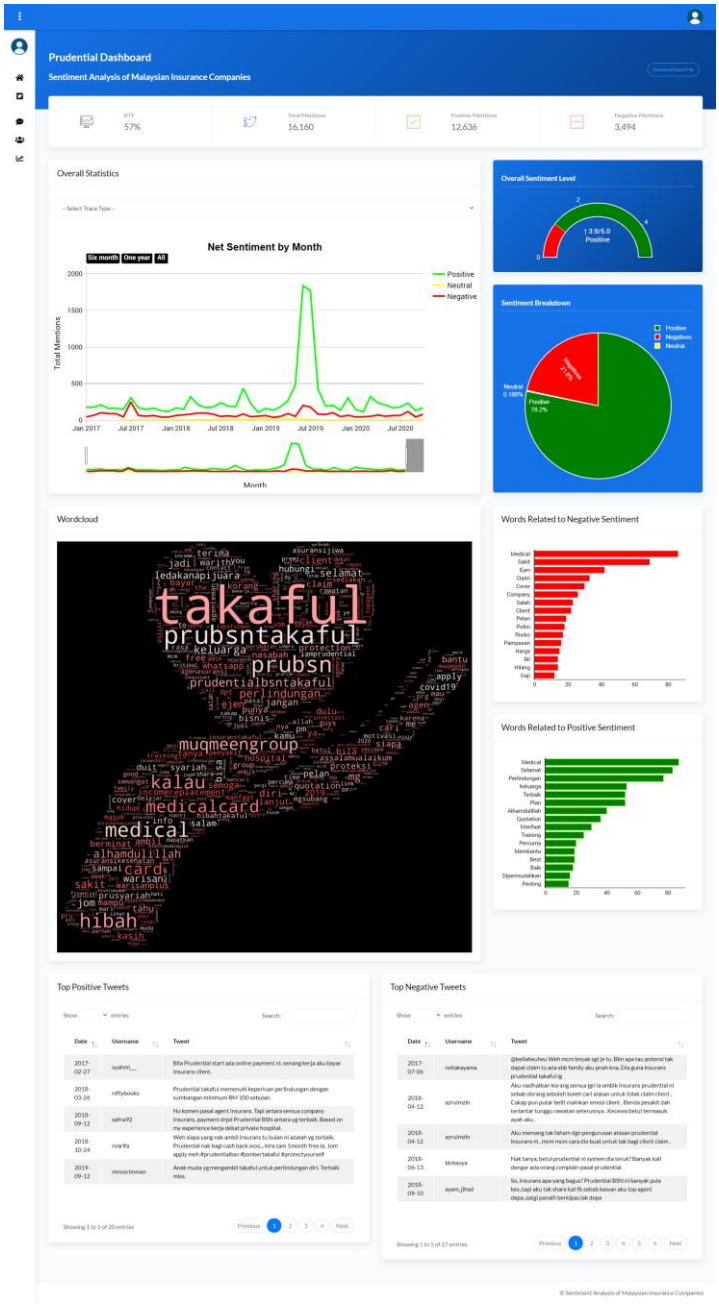


Figure 4.26 Data Dashboard for Prudential Page

Figure 4.26 shows the application interface of the Prudential page. The results of the analysis are presented to the users using data visualization techniques.

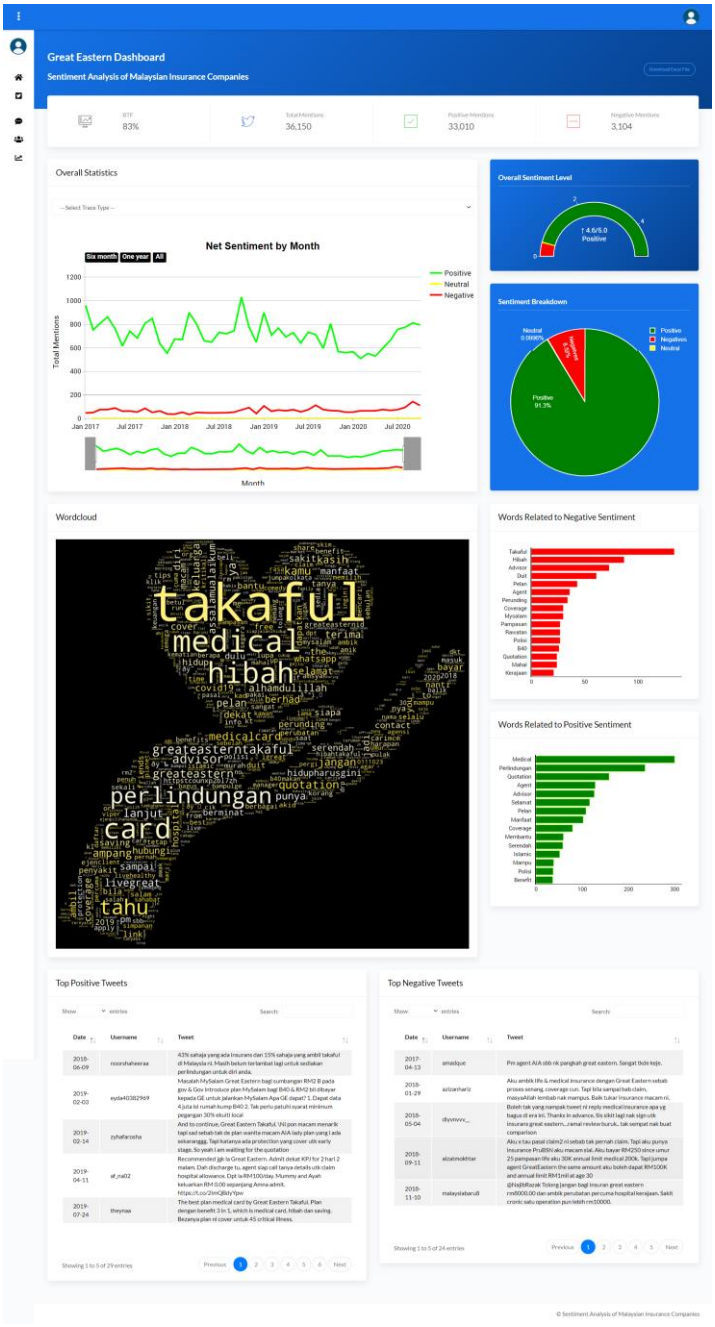


Figure 4.27 Data Dashboard for Great Eastern Page

Figure 4.27 shows the application interface of the Great Eastern page. The results of the analysis are presented to the users using data visualization techniques.

4.3.3 Twitter Updates

The three companies' official Twitter accounts have been embedded on the Twitter updates page, where the user can see real-time updated tweets streaming from these accounts. The embedded tweets were customized and displayed directly on this page from the Twitter developer account. Figure 4.28 indicates the Twitter Updates Page.

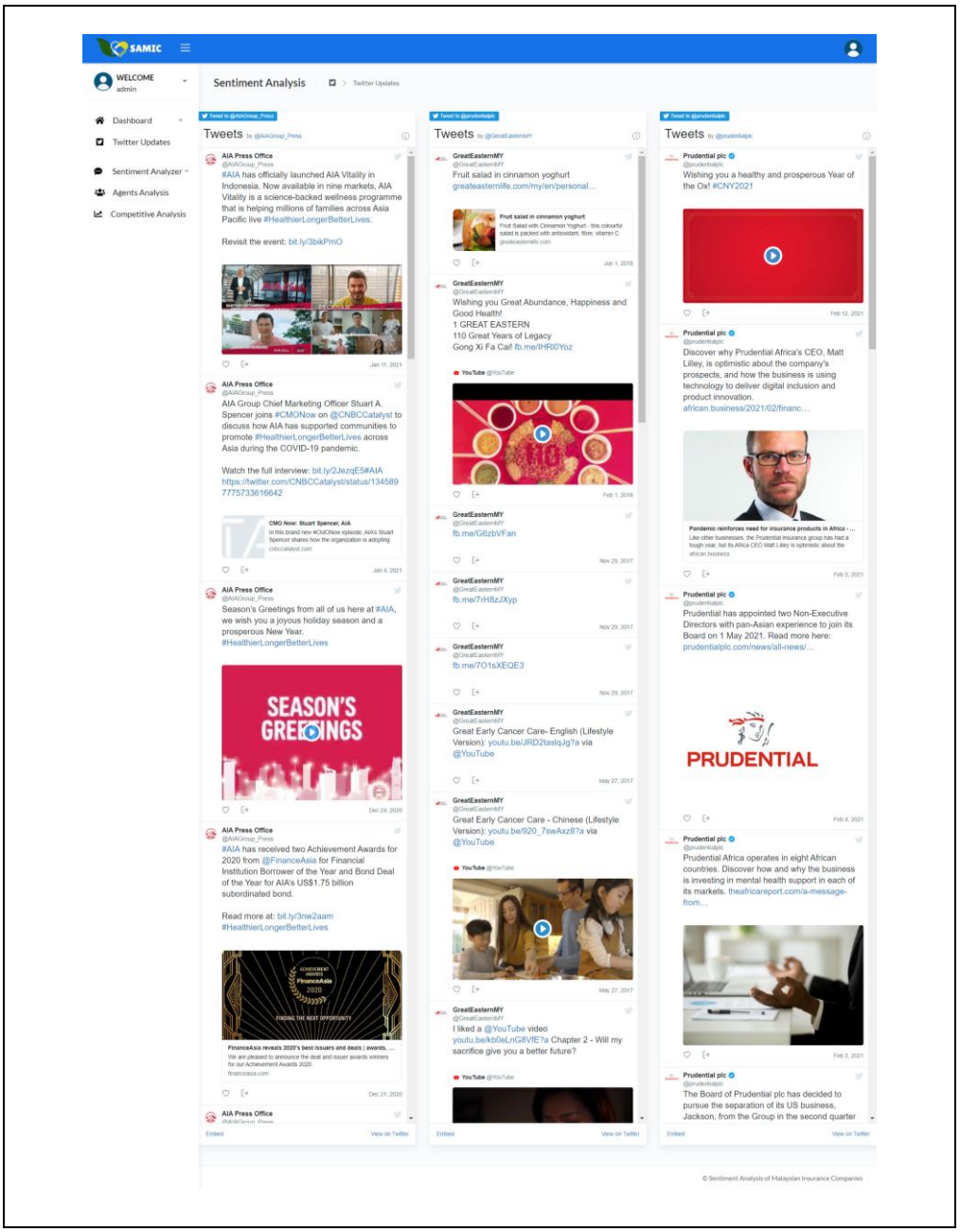


Figure 4.28 Twitter Updates Page of the Companies

4.3.4 Sentiment Analyzer

In this application, there are two sentiment analyzer techniques involved. The first is the Text Sentiment Analyzer. The user can manually input the review in Malay or English to see whether the sentence is positive, negative, or neutral by using the machine learning model built. Additionally, the user may also perform basic analysis in which the sentence structure is analyzed according to natural language processing (NLP). Figure 4.29 displays the page of the Text Sentiment Analyzer.

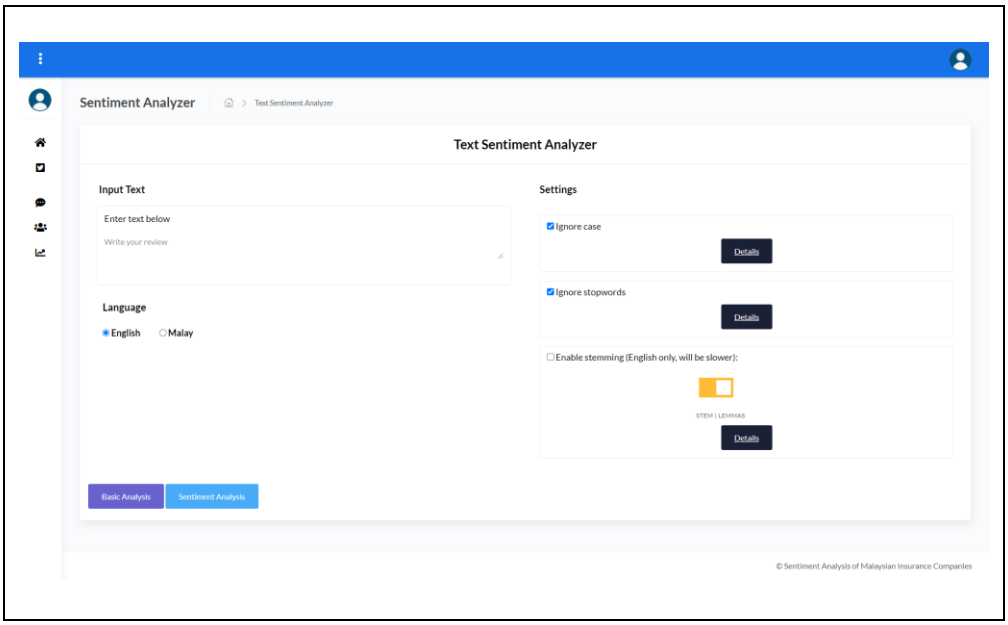


Figure 4.29 Text Sentiment Analyzer Page of the Application

The second technique is the Tweet Sentiment Analyzer, where tweets are retrieved in real-time and sentiment analysis is conducted on the tweets. These tweets colored fields, as in Figure 4.30, represent the sentiment of the tweet, whether it is positive, neutral, and negative. The user may also download CSV files of the extracted real-time tweets with the labeled sentiment. By using this function, the user is able to monitor overall customer sentiment towards their brand in real-time at any given moment and instantly see insights about their brand reputation. However, to avoid the Twitter API rate limit, tweets retrieved are limited to 100 tweets per search. Twitter imposes a limit for how many

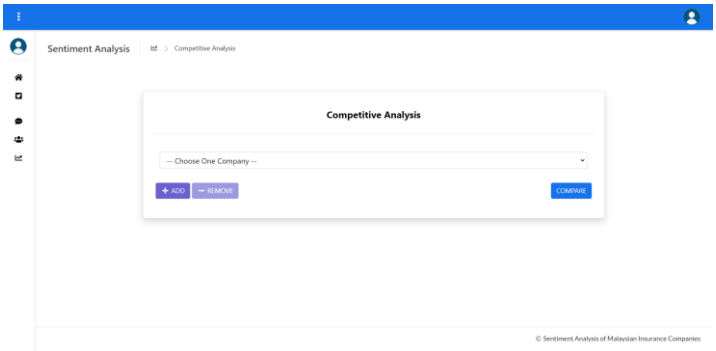
times it may be used in an hour, and the rate limit applies to a basic Twitter developer account that makes a total of 100 API calls per hour.

Figure 4.30 Tweet Sentiment Analyzer Page of the Application

4.3.5 Competitive Analysis

To effectively assess businesses' competitive environment, insurance companies need to monitor and evaluate the analyzed content on their sites and their competitors' sites. The Competitive Analysis page was built to accomplish this objective, in which the user can compare performance between companies of their choice. Graphs and charts comparing the performance of these companies are visualized on this page as in Table 4.5.

Table 4.5 Competitive Analysis Page of the Application

UI Diagram	**Description**
 **Figure 4.31** Interface of the Competitive Analysis Page	The Competitive Analysis page is illustrated in Figure 4.31. The users must insert two or three company names, and a comparison of sentiments between those companies will be displayed.

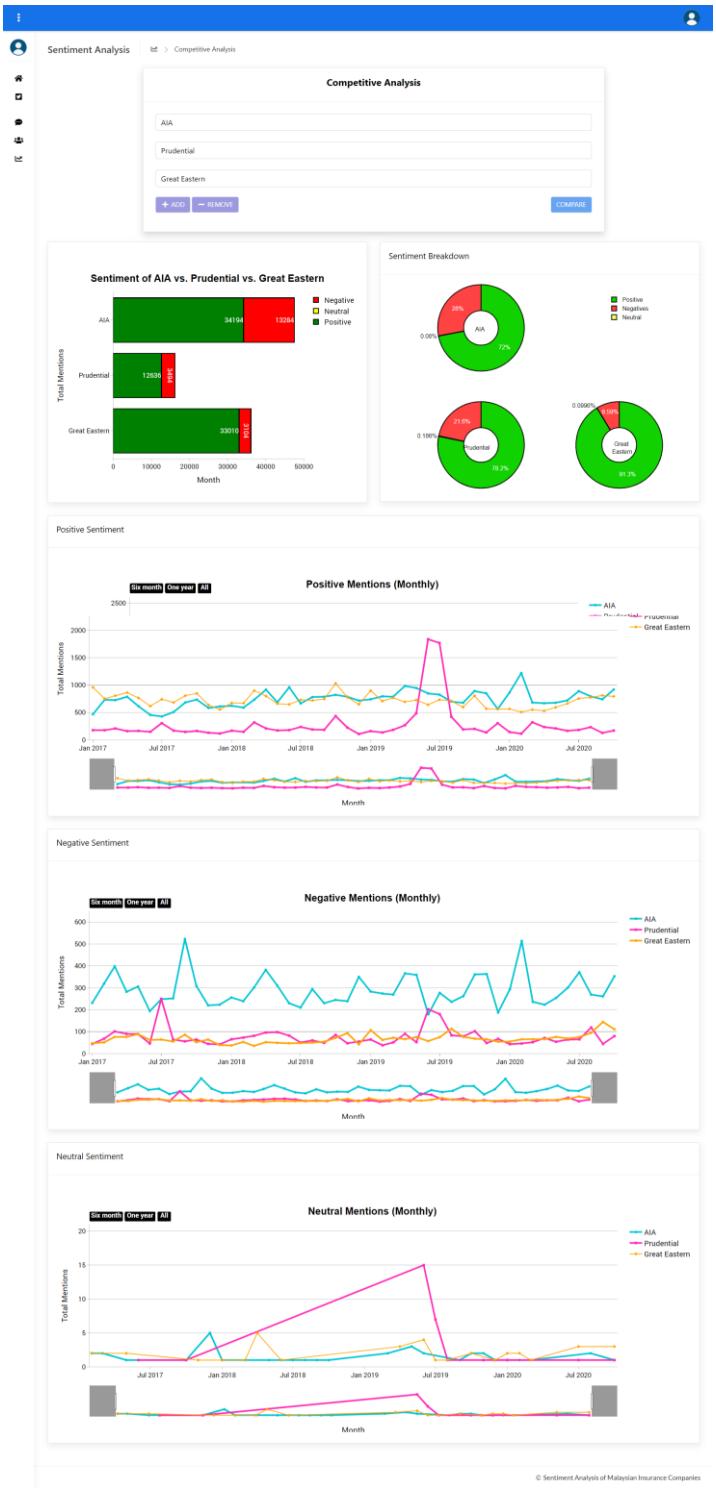


Figure 4.32 Results of Comparison between Companies

Figure 4.32 shows an example of the visualization results between companies.

4.3.6 Additional Functionality

The Twitter user profile accounts of insurance agents that actively promote the company's insurance plan has been extracted for all three companies and displayed on the Insurance Agents' Performance page. Exploratory data analysis (EDA) was carried out using the reach score, which is the number of followers and following count, and popularity score, which is the retweets count and favorites count, to find accounts with the highest-ranking of the overall score. Riquelme and González-Cantergiani (2016) used these metrics to calculate the influence score measure as the Twitter Follower-Followee ratio (TFF) and TRank, ranking users in some dimensions by considering retweets and favorites counts. The user can also visit the insurance agents' profile with the highest overall score from the preference company on this page as in Figure 4.33.

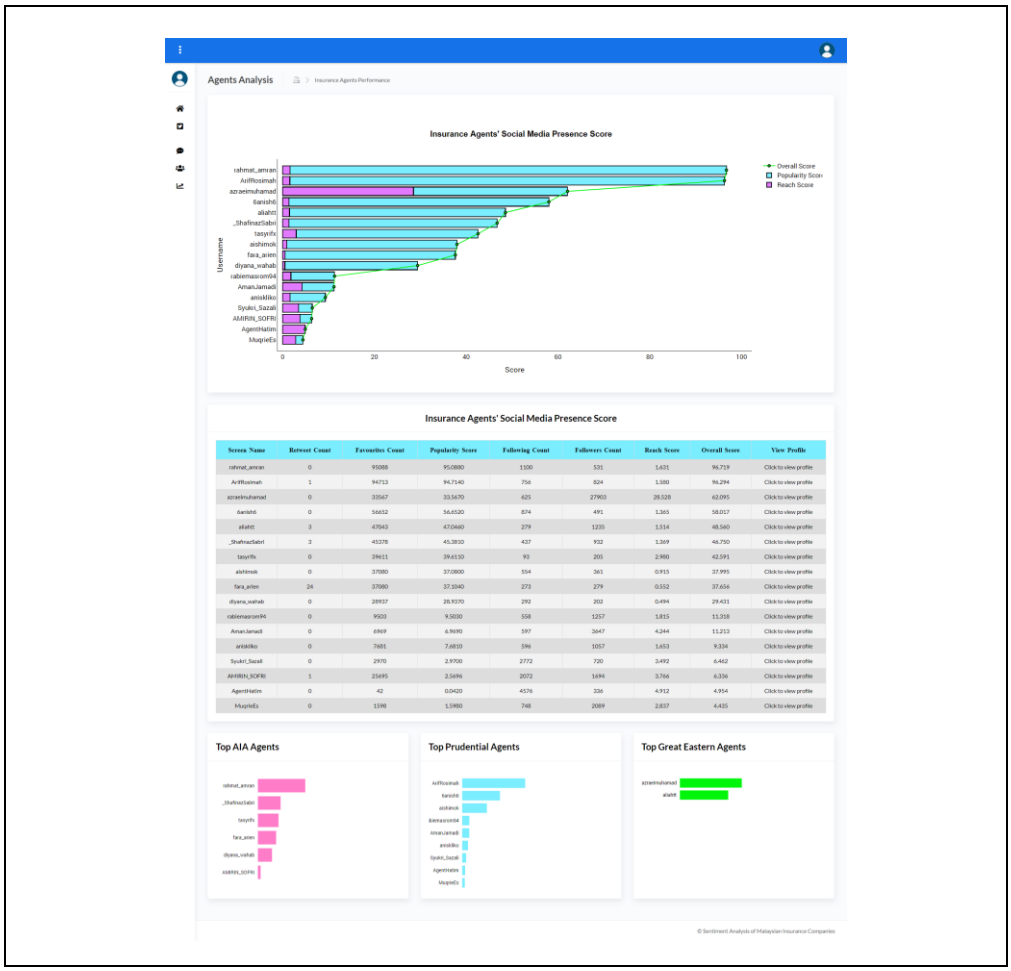


Figure 4.33 Agents' Performance Page of the Application

CHAPTER 5

RESULTS AND DISCUSSION

This chapter presents and discusses the analysis results performed on real-world data and the testing carried out on both functionality and usability testing. Functionality testing is performed to assess the system's functionality and ensure that the system's functionality performs properly, whereas usability testing is carried out to evaluate the user system design and efficiency of use.

5.1 Model Evaluation and Validation Results

Model evaluation is the method by which the model is evaluated to ensure that it can perform acceptably in the real-world data supplied, and it is performed before the model is deployed. The first step in this phase was finding the best algorithm to classify the real-world data is to compare their mean accuracy by tweaking values, trials, and errors. The performance of various machine learning algorithms on the training and test datasets is compared and evaluated to select the best Twitter sentiment classification model. In this context, the model validation method is part of the evaluation metrics where the holdout test set to assess the generalization performance is used. It is the simplest type of cross-validation in which, as stated in the previous chapter, the training dataset collected is separated into two sets, called the training set and the testing data by 80:20 ratio. Additionally, to compute the optimal values of hyperparameters, grid-search hyperparameter tuning has been used. Exhaustive search is performed on the algorithm's specific parameter values.

Pipeline architectures make it possible to perform a sequence of data preparation within cross-validation folds. The cross-validation and grid search for model selection returns the best model parameters and overall accuracy as in Table 5.1. Since the dataset is highly imbalanced, a random oversample

technique is used to balance the dataset. Random oversampling involves randomly selecting, replacing, and adding examples from the minority class to the training dataset to increase the number of instances to the dataset in a balanced manner. It works by generating new instances that are provided as input from existing minority cases. The algorithm with the highest performance will then be applied to the Twitter data collected on the insurance companies to label the sentiment in that dataset.

Table 5.1 Best Accuracy Returned by Different Classifiers

Algorithms	**Accuracy**
Random Forest (RF)	54%
SVC + RandomOverSampler	85%
Multinomial Naïve Bayes (NB)	84%
LinearSVC	90%

LinearSVC, a different implementation of the SVM algorithm from the Scikitlearn package, provides the best accuracy of 90% for classification. Hence, the LinearSVC classifier is used for the final model. The results of grid-search hyperparameter tuning parameters with fine-tuned parameter values are shown in Table 5.2. The LinearSVC classifier outperforms the classifier by a substantial accuracy, with the RF yielding the best accuracy of 54%. It is an increase of 36% over the best accuracy returns by RF classifier.

Table 5.2 Results of Parameters with Fine-Tuned Parameter Values

Parameters	**Value**
'bow__ngram_range'	Bi-gram (1,2)
'clf__estimator__loss'	Hinge
'clf__estimator__penalty'	l2
'clf__estimator__tol'	0.0001
'tfidf__use_idf'	True

The classification report in Figure 5.1 provides the main classification metrics on every class of the data. In comparison, the heatmap represents the proportion

in each class of correctly classified examples. The raw numbers of testing data are also stated within each cell. A simple metric, including classifier error rate and accuracy, has been used to measure the model's performance. In the confusion matrix, the 'negative' class is represented as 0, whereas the 'positive' class is denoted as 2, and 4 represented the 'neutral' class. The classifier has an accuracy of 90%. In simpler terms, out of 10 attempts, the model obtained approximately nine correct results based on the data provided from the testing data set, which were correctly classified as 'positive', 'neutral', and 'negative'. The precision, indicating the percentage of a class predicted labels that rightfully belonged to that class is 92% for the 'negative' class, while the 'positive' class is 88% and 31% for the 'neutral' class.

```
accuracy score:  0.8956365537714208


confusion matrix:
 [[115219  12297    371]
 [  9897  96784    498]
 [   598   1089    399]]

              precision    recall  f1-score   support

           0       0.92      0.90      0.91    127887
           2       0.88      0.90      0.89    107179
           4       0.31      0.19      0.24      2086

    accuracy                           0.90    237152
   macro avg       0.70      0.67      0.68    237152
weighted avg       0.89      0.90      0.89    237152
```

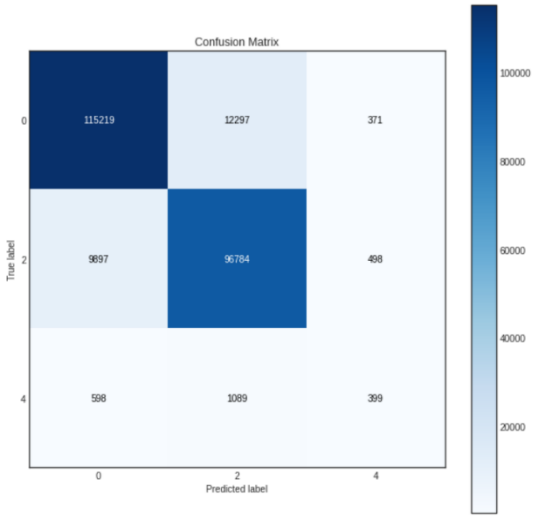


Figure 5.1 Confusion Matrix of Malay Model

The third metric shows in the confusion matrix are recalled how well the model reflects a particular class. The recall shows the percentage of labels that were correctly classified from the total of a class. For both the 'negative' and 'positive' class, the recall is 90%, while for the 'neutral' class, it is 19%. A weighted harmonic means of accuracy and recall is the F1 score. The F1 score for the 'negative' class is 91%, the 'positive' class is 89%, and the 'neutral' class is 24%. As compared to other classes, the 'negative' class is the most correctly classified, with only a few instances being misclassified. However, the 'neutral' class is the hardest to classify and is highly confused with the other classes. The outcomes of precision and recall values for each of the three classes confirm that the 'neutral' class suffers from substantially poorer results than other classes. Therefore, to evaluate the classifier approach, a baseline study has been specified and compared to compare the classifier performances.

Based on the same study conducted by Bouazizi and Ohtsuki in 2018 for multi-class sentiment analysis on Twitter. Neutral tweets are also the hardest to classify, with a true positive rate equal to 28.5%. The overall accuracy of the validation testing data is equal to 76.3%. Additionally, the sentiment analysis model for Twitter data in the Polish language revealed that multi-class classification is still significantly more difficult to be modeled and only achieves an accuracy of 18.1% when used to classify tweets into three classes of positive, neutral, and negative. However, it produced nearly 71% accuracy and high precision of 87.77% when trained with the binary class (Chlasta, Wolk, & Krejtz, 2019). Most studies on sentiment analysis overlooked 'neutral' examples, but the classification is important to distinguish between positive and negative examples. According to Figure 5.2, the Malay model's overall accuracy showed an excellent classification with an accuracy score of 90%. Thus, according to the overall accuracy, the model is acceptable to classify real-world data.

Accuracy	
0.90 – 1.00	Excellent Classification
0.80 – 0.90	Good Classification
0.70 – 0.80	Fair Classification
0.60 – 0.70	Poor Classification
0.50 – 0.60	Failure

Figure 5.2 Classifier Model Accuracy Indicator
(*Source*: Neaeni & Sari, 2017)

The main classification metrics of the English model is provided in the classification report in Figure 5.3. Similarly, in the confusion matrix, the 'negative' class is represented as 0, while the 'positive' class is denoted as 2, and the 'neutral' class is denoted as 4. The classifier's accuracy is 81%, indicating that the model achieved approximately eight correct results out of 10 attempts based on the data provided from the testing data set that were correctly classified as 'positive', 'neutral', and 'negative'. According to research, human raters usually agree 80% of the time due to inter-rater reliability issues. If a program is 'right' 100% of the time, humans would still disagree with it about 20% of the time (Wahome, 2018). Quantification's task being a challenging task is highly subject to the annotators' opinion. As a matter of fact, this is a characteristic that is generally valid for sentiment analysis, including for simple binary classification tasks, where texts can be classified into positive or negative classes.

```
accuracy score:  0.8110693628360655


confusion matrix:
 [[97379 19955   963]
 [21659 97073   399]
 [ 1217  1112    40]]

              precision    recall  f1-score   support

           0       0.81      0.82      0.82    118297
           2       0.82      0.81      0.82    119131
           4       0.03      0.02      0.02      2369

    accuracy                           0.81    239797
   macro avg       0.55      0.55      0.55    239797
weighted avg       0.81      0.81      0.81    239797
```

Figure 5.3 Confusion Matrix of English Model

Over and above that, the relevant findings that can be concluded are that the classification model's performance is independent of the test set, and the quantification part can be carried out without overfitting issues for the classification of a real-world dataset. Since a balanced dataset for 'neutral' data is hard to collect a priori, it is believed that adding 'neutral' instances of training data, and more importantly, a balanced training dataset among all sentiment classes could significantly improve the performance of the model.

5.2 Result of Real-World Data Analysis

Real-world data analysis is plotted and displayed on the application dashboard. The data is visualized as line and bar graphs, pie charts, and word clouds to interpret the data and draw insights. The following subsections described a detailed description of some of the data visualized, including comparing the three insurance providers.

5.2.1 Overview Visualization of the Insurance Companies

The bar graph in Figure 5.4 indicates the insurance companies' overall sentiment between 2017 and October 2020. According to the graph, AIA has the most positive and negative mentions, whereas Prudential has the least amount of mentions on Twitter. In comparison, over the span of 3 years, Great Eastern had a significant amount of positive sentiment. Overall, it is clear that all the insurance companies mentioned on Twitter have more mentions of positive sentiment than negative sentiment. The brand talkable favorability (BTF) of insurance companies is provided as in Table 5.3.

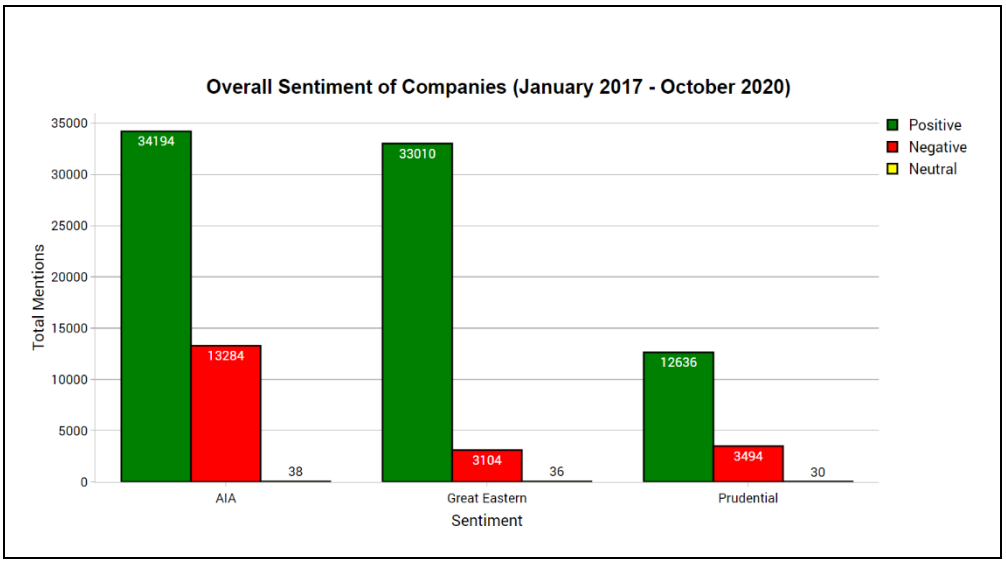


Figure 5.4 Overall Sentiment of Insurance Companies

Table 5.3 Twitter Mentions and Overall BTF

Company	**Total Mentions**	**Positive Mentions**	**Negative Mentions**	**Neutral Mentions**	**BTF**
AIA	47,516	34,194	13,284	38	44%
Prudential	16,160	12,636	3,494	30	57%
Great Eastern	36,150	33,010	3,104	36	83%

The graphical representation in the dataset of positive sentiment text data is displayed as a word cloud as in Figure 5.5. The more frequently a word is used in the dataset, the larger it would appear in the chart. Some of the most common words associated with positive sentiment are related to the insurance providers' plan from the word cloud produced. Besides, words such as 'bagus', 'good', 'better', and 'great' stand out since they have been used frequently.

Figure 5.5 Positive Word Cloud of the Mentions

The word cloud produced in Figure 5.6 is a representation of words related to the negative sentiment. The negative sentiment is primarily attributed to insurance agents, policy, and claims. In the word cloud, terms such as 'problem', 'payah, 'susah' are prominent.

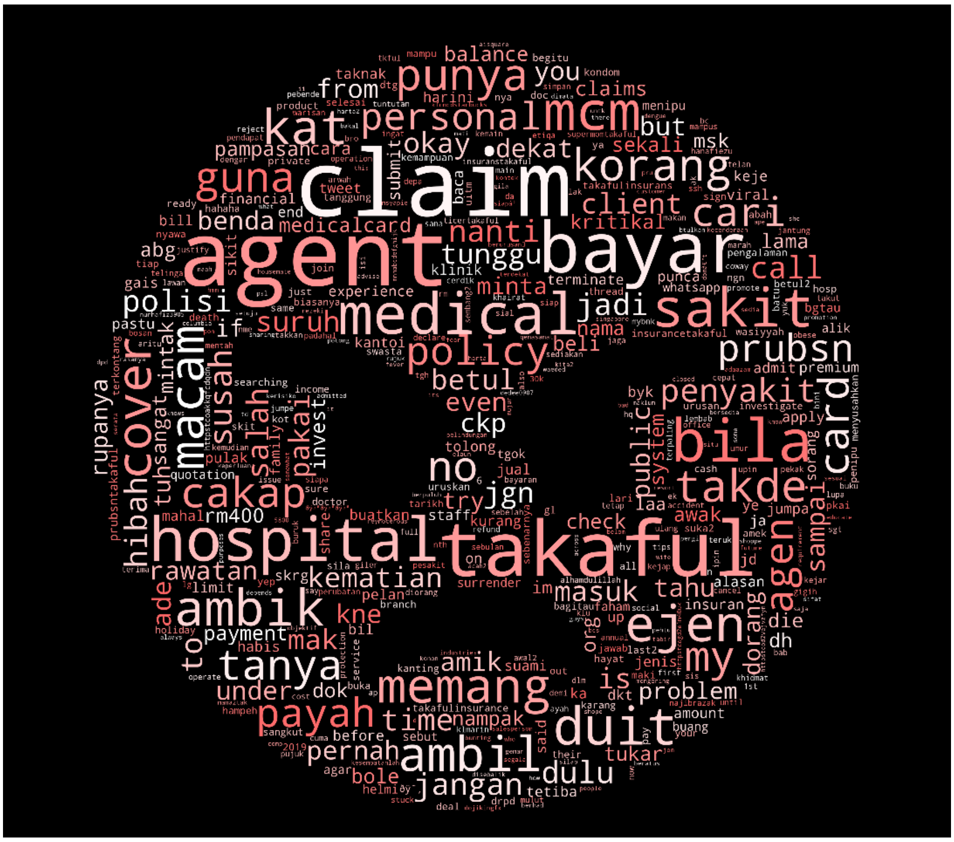


Figure 5.6 Word Cloud of the Negative Mentions

5.2.2 Data Visualization of AIA Company

The pie chart in Figure 5.7 illustrates the sentiment percentage for AIA company. The green color represents positive mentions, the red color represents negative mentions, and neutral mentions are represented by yellow color. The chart shows that 72% positive sentiment constituted a large proportion of overall mentions with 34,194 mentions on Twitter. On the other hand, the negative and neutral mentions made up 28.08% of the chart with 13,322 mentions. The BTF of this company is obtained by adding the percentage of positive mentions to the percentage of neutral mentions and subtracting the percentage of negative mentions. Therefore, the resulting BTF is 44%.

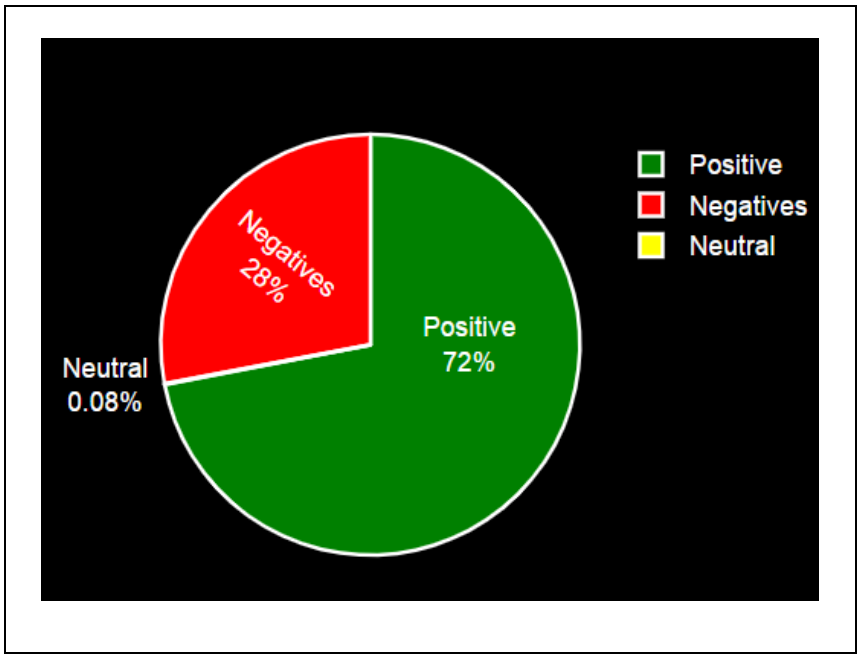


Figure 5.7 Pie Chart of Labelled Sentiment

Table 5.4 displays a time series plot and stacked bar chart of AIA insurance company's total mentions on Twitter by month. From the following plot, it can be observed that there is no consistent trend as the series appears to wander up and down over the entire month slowly. The number of positive mentions is slightly more than the negative mentions, which indicates that Twitter user often gives good feedback about the company. The graph reached its peak in Feb 2020, where the total number of positive mentions is 1,220. After looking closely at the sudden shifts of positive mentions, it was found that there has been a new life insurance plan launched by the company.

Table 5.4 Sentiment of Twitter Mentions by Month

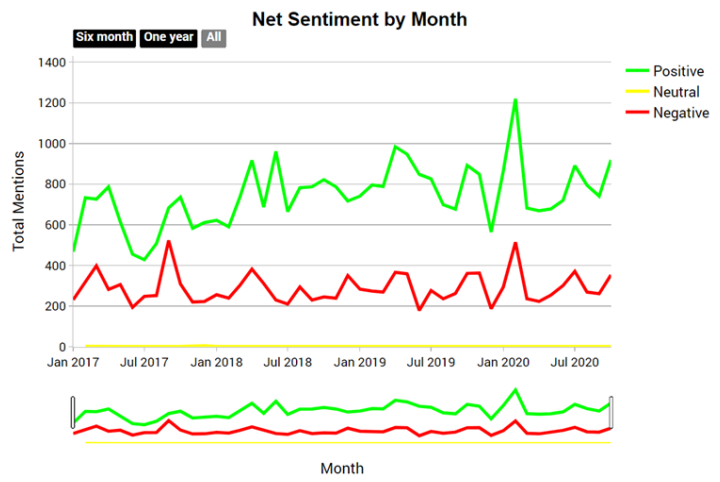


Figure 5.8 Time Series Plot of AIA's Mentions

Figure 5.8 displays a time series plot of AIA insurance company's total mentions on Twitter by month

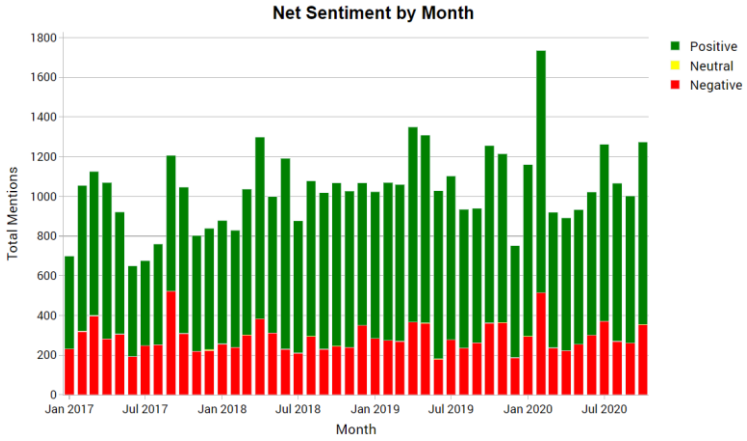


Figure 5.9 Stacked Bar Chart of AIA's Mentions

A stacked bar chart of the AIA insurance company's total Twitter mentions by month is shown in Figure 5.9

The word cloud produced in Figure 5.10 represents words in AIA's company text corpus. From the word cloud, it can be observed that most Twitter users tweet about the insurance plans offered by the company, such as 'takaful', 'medical card', and 'aia public takaful'. Thus, to gain valuable insights, the insurance company will have to take a closer look at what policyholders talk about the plans offered. Besides, terms such as 'claim' and 'policy' are also notable in the word cloud generated. For such a reason, the insurance company must ensure that the customer does not complain about the policy or the claims they have issued.

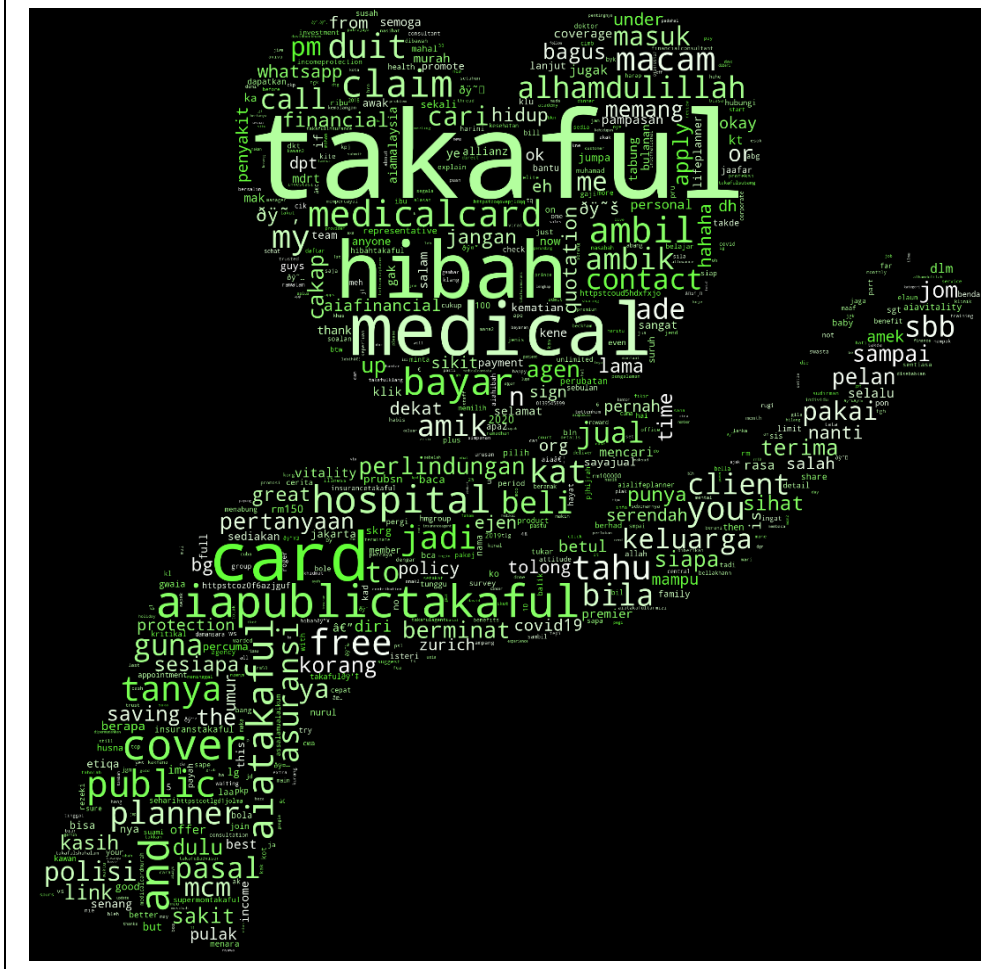


Figure 5.10 Word Cloud of AIA Company Mentions

5.2.3 Data Visualization of Prudential Company

The chart in Figure 5.11 displays the Prudential company's sentiment percentage. From the pie chart, it is clear that most users wrote good reviews about the company on Twitter with 12,636 mentions. The 78.2% positive sentiment formed a major percentage of total mentions. On the contrary, the negative mentions made up about 21.8% of the chart with a total of 3,524 mentions, while a small minority of 0.186% made up the neutral mentions, the resulting BTF is 57%.

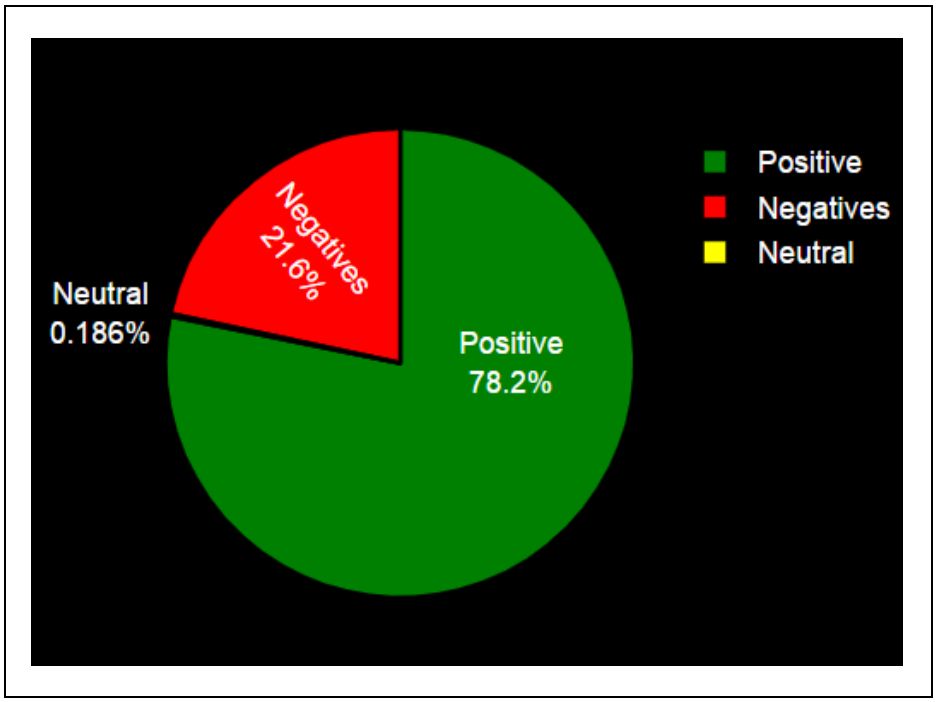


Figure 5.11 Pie Chart of Prudential's Mentions Sentiment

The time-series plotted and stacked bar chart in Table 5.5 illustrates the total mentions by months of Prudential company's on Twitter. In general, the dataset shows a trend of random variation in which there are no patterns or cycles with positive mentions slightly more than negative and neutral mentions. Nevertheless, the plot shows a drastic shift between April to June 2019. The plot rose significantly and hit its peak in Jun 2019 with 1,837 positive mentions. After examining the dataset, the reason for the shift is because of the retirement plan the company released. However, after that month, the graph shows that the positive mentions declined noticeably after a sudden surge. The highest negative mentions per month the company obtained is in July 2017, with 250 mentions because of the campaign 'prudential ride London' that blocks the street for cyclists.

Table 5.5 Sentiment by Month of Prudential's Mentions

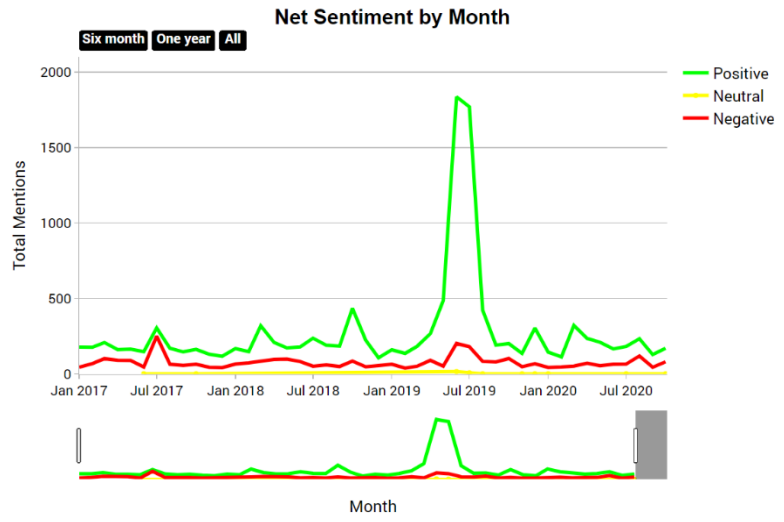


Figure 5.12 Time Series Plot of Prudential's Mentions

Figure 5.12 displays a time series plot of Prudential insurance company's total mentions on Twitter by month

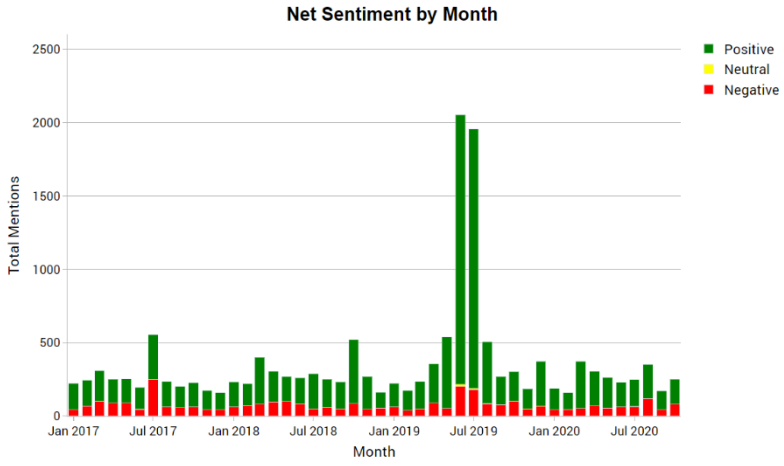


Figure 5.13 Stacked Bar Chart of Prudential's Mentions

A stacked bar chart of the Prudential insurance company's total Twitter mentions by month is displayed in Figure 5.13

From the word cloud produced in Figure 5.14, it is seen that the words often used by Twitter users related to Prudential company are about the plans offered, such as 'takaful' and 'prubsntakaful', similar to the previous AIA company's plotted word cloud. 'Quotation' and 'safe' are among other frequent terms found in the text corpus. A possible explanation for this trend is that Twitter users wrote a positive evaluation of the plan offered.

Figure 5.14 Word Cloud of Prudential's Mentions

5.2.4 Data Visualization of Great Eastern Company

The pie chart in Figure 5.15 displays the percentage of the sentiment of Great Eastern's mentions on Twitter. The positive mentions notably make up the majority percentage of total mentions, which is 91.3% with 33,010 data. Negative and neutral mentions only account for 8.7% of total mentions with 3,140 data. The resulting brand talkable favorability (BTF) is 83%, which is the highest among other companies.

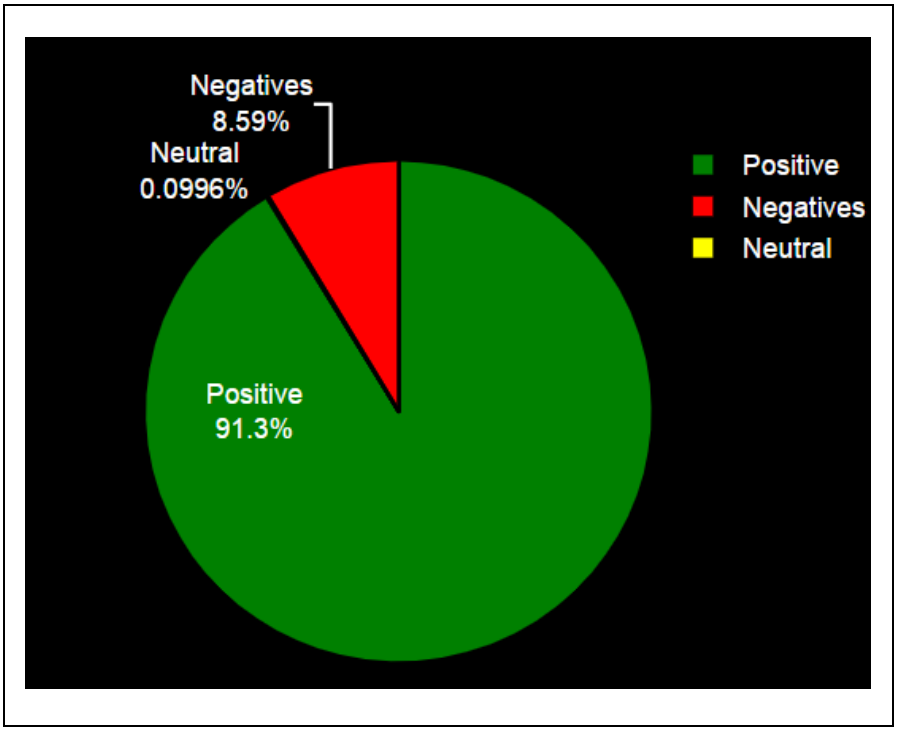


Figure 5.15 Pie Chart of Great Eastern Sentiment of Mentions

The total mentions by months of Great Eastern company on Twitter are depicted as in Table 5.6. It can be observed from the series plotted that the total mentions tend to fluctuate from month to month. The positive mentions are noticeably higher than negative mentions that contributes to the highest overall brand talkable favorability (BTF) among other insurance companies. Besides, over the years, the negative mentions flatten slightly. The highest negative mentions were 144 in September 2020 because of the sudden spike of Covid-19 cases in Malaysia. Individuals began to raise awareness about taking the company's medical card and began to compare the pros and cons of the company's medical card. On the contrary, the highest positive mentions were in October 2018, with 1031 mentions. Great Eastern was awarded the 'Consumer Choice Award 2018' as the best financial performance for life insurance during that month.

Table 5.6 Sentiment of Great Eastern's Mentions by Month

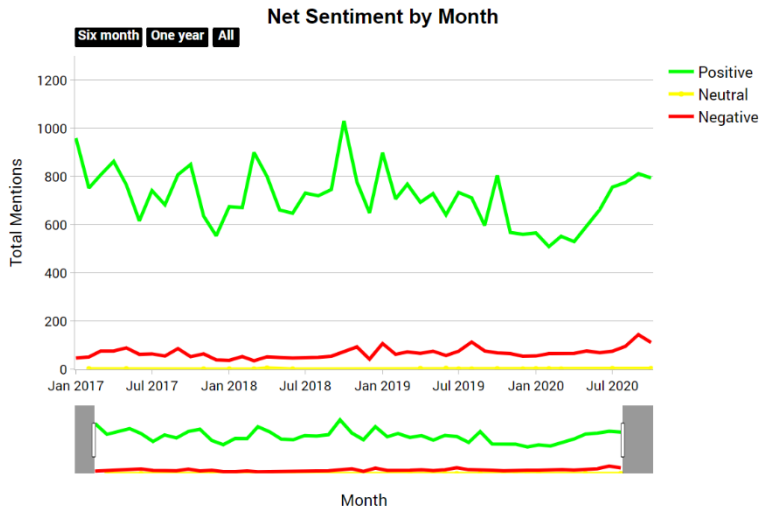


Figure 5.16 Time Series Plot of Great Eastern's Mentions

Figure 5.16 displays a time series plot of Great Eastern insurance company's total mentions on Twitter by month

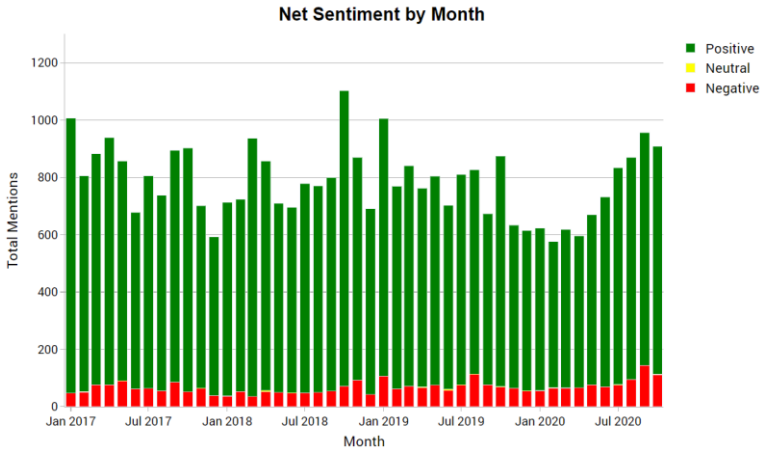


Figure 5.17 Stacked Bar Chart of Great Eastern's Mentions

A stacked bar chart of the Great Eastern insurance company's total Twitter mentions by month is displayed in Figure 5.17

Figure 5.18 shows the word cloud for Great Eastern mentions on Twitter. From the word cloud, it can be concluded that most Twitter users discussed the plans offered by the company, such as 'takaful', 'medical', and 'hibah'. Furthermore, 'quotation' and 'coverage' are among the prominent terms. The company may use this information to improve the services provided related to quotation and coverage of insurance.

Figure 5.18 Word Cloud for Great Eastern Mentions

5.3 Functionality Testing

Testing features is essential to ensure that all application features operate correctly and that any detected error is fixed. The purpose of performing functionality testing is to test each function of the visualization application to determine how closely the specifications are matched by providing appropriate input and checking the output against the functional requirements set out in the previous chapters. The test is conducted based on the development of test case scenarios derived from program specifications, with the list of functionalities tested and the results shown in Table 5.7. In the next subsection, the details of each test case carried out are discussed.

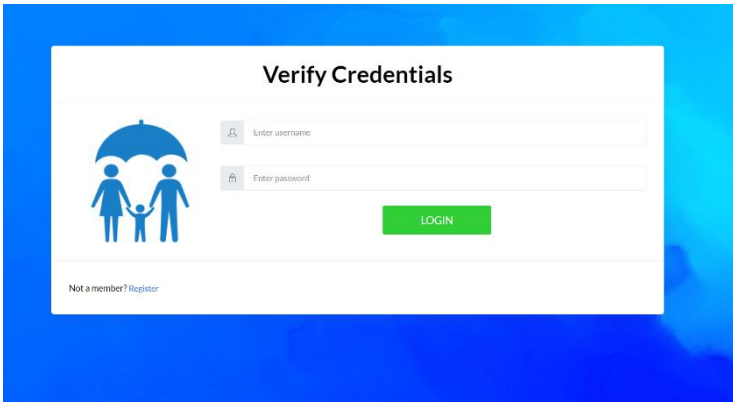
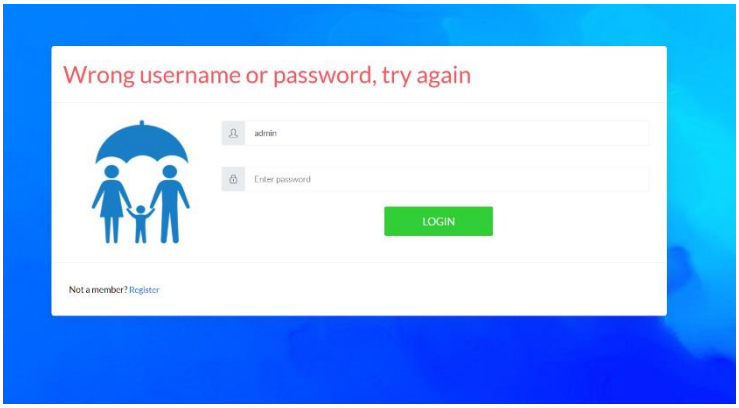
Table 5.7 Results of Overall System Functionality of the Application

Test Case	**Expected Result**	**Success/Failure**
Login Page	The user credentials must be verified	Success
View Overview Page	The overview page of the three companies and a line chart to visualize the summary of the analysis will be displayed	Success
View AIA Page	The results of sentiment analysis are shown, and the data are visualized through data visualization techniques on the AIA page	Success
View Prudential Page	The results of sentiment analysis are shown, and the data are visualized through data visualization techniques on the Prudential page	Success
View Great Eastern Page	The results of sentiment analysis are shown, and the data are visualized through data visualization techniques on the Great Eastern page	Success
Download Excel Files with Labelled Sentiment	The Excel files with labeled sentiment will be saved into the user's computer	Success
View Twitter Updates	The real-time timeline of Twitter insurance updates is presented to the user	Success
Analyze Text	Users can insert any text input into the text sentiment analyzer page	Success
View Results of Sentiment Analysis	The result of sentiment analysis on the text entered by users are displayed	Success
View Tweet Sentiment Analyzer Page	The result of sentiment analysis from tweets extracted in real-time will be displayed on the page	Success
Search Tweet Sentiment in Real-Time	The user can search any tweet keywords to analyze the sentiment	Success
View Agents Analysis Page	The result of extracted Twitter user profile accounts of insurance agents will be displayed	Success
Compare Performance of Companies	The results of the analysis and the comparison of sentiment between companies are visualized to the user	Success
Logout Page	The user can logout from the application	Success

5.3.1 Login Page

The result of the login page test is described in Table 5.8.

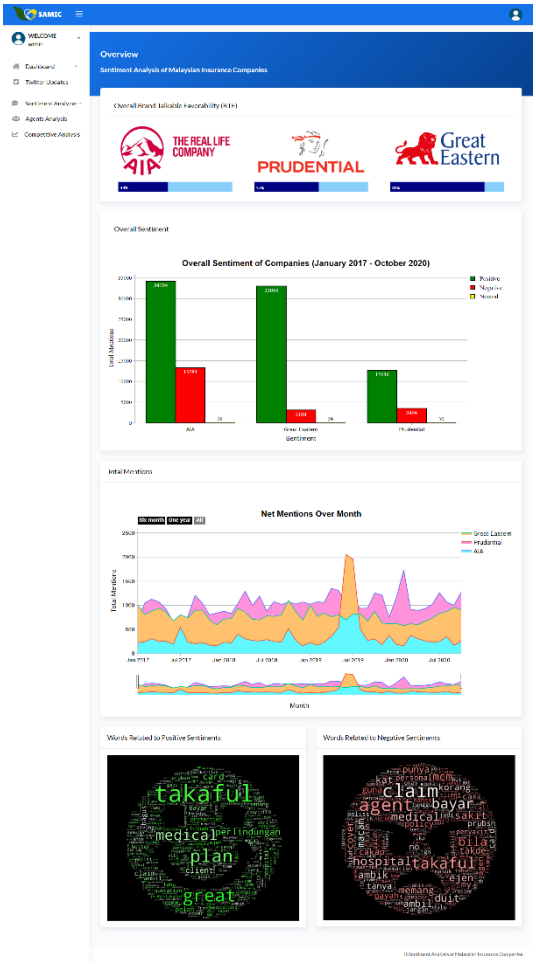
Table 5.8 Functionality Testing for Login Page

Test Objective	To verify if the user can log in to the application using the login page, as in Figure 5.19, the application grants them access and directs them to the application's overview page.
Potential Test Inputs	 **Figure 5.19** Interface for Login Page
Expected Test Outputs	1. The user can log in to the application and will be directed to the application's overview page. 2. The application will show an error message if the password does not match, as in Figure 5.20.
Test Procedures	1. Enter the username and password. 2. Press the 'Login' button to proceed.
Actual Test Results	 **Figure 5.20** Error Page

5.3.2 View Overview Page

The result of the view overview page test is described in Table 5.9.

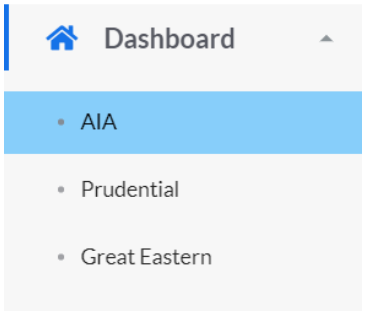
Table 5.9 Functionality Testing for Overview Page

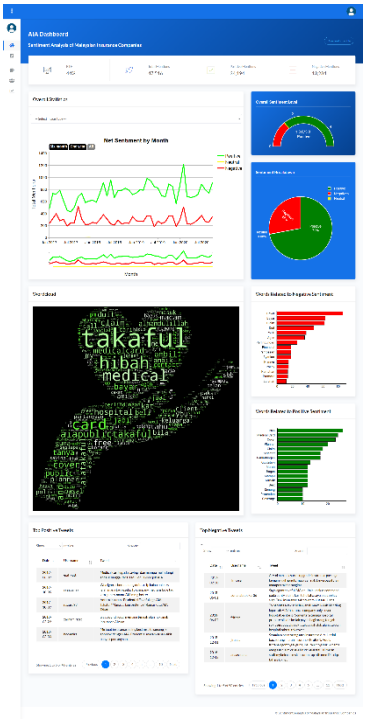
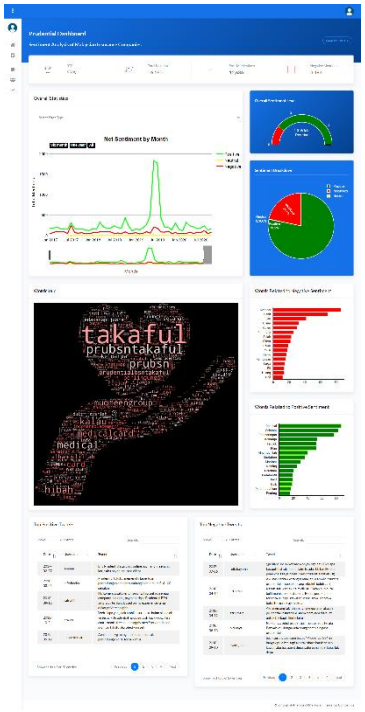
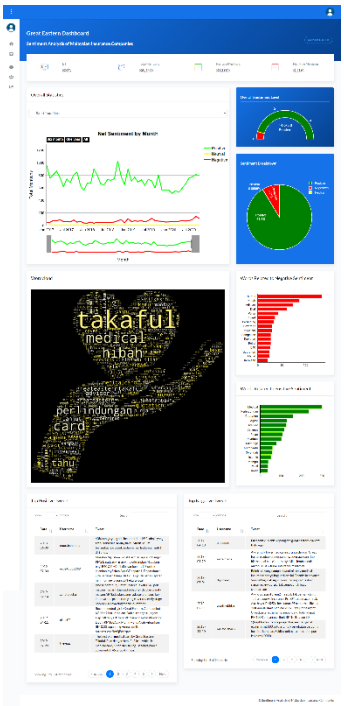
Test Objective	To observe whether the user is directed to the overview page after the user has successfully logged in using the login button as in Figure 5.21.
Potential Test Inputs	 **Figure 5.21** Button Log in to the Application
Expected Test Outputs	The user is able to view the visualized data of all three insurance companies, as illustrated in Figure 5.22.
Test Procedures	The user will be automatically directed to the overview page if the credentials provided exists in the application database.
Actual Test Results	 **Figure 5.22** Interface for Overview Page

5.3.3 View AIA, Prudential, or Great Eastern Page

The result of the view AIA, Prudential, or Great Eastern page test is described in Table 5.10.

Table 5.10 Functionality Testing for AIA Page

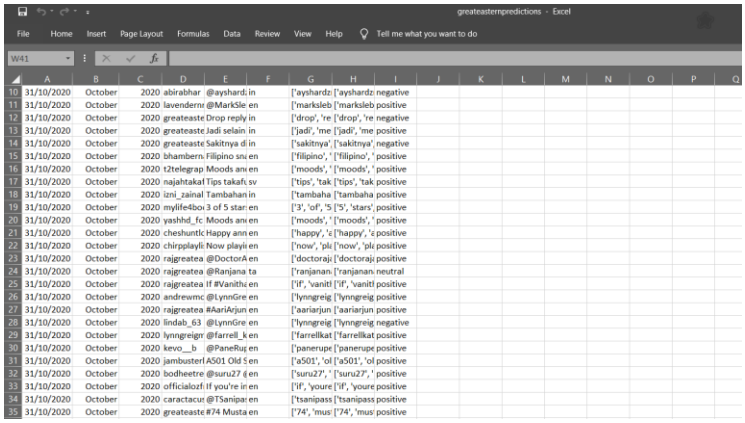
Test Objective	To test if the user is directed to the AIA, Prudential, or Great Eastern page when the user selects the page from drop-down menu options as in Figure 5.23.
Potential Test Inputs	 **Figure 5.23** Drop-down Menu Options
Expected Test Outputs	1. The user can view the visualized data about AIA tweets as in Figure 5.24 if the AIA page is selected from the drop-down menu options. 2. The user can view the visualized data about Prudential tweets as in Figure 5.24 if the Prudential page is selected from the drop-down menu options. 3. The user can view the visualized data about Great Eastern tweets as in Figure 5.24 if the Great Eastern page is selected from the drop-down menu options.
Test Procedures	The user must use the navigation bar to select the AIA, Prudential, or Great Eastern page from the drop-down menu that contains links to other sections of the application.

Actual Test Results	 **Figure 5.24** Interface of AIA Page  **Figure 5.25** Interface of Prudential Page  **Figure 5.26** Interface of Great Eastern Page

5.3.4 Downloading Excel File with Labelled Sentiment

The result of the downloading excel file with the labeled sentiment test is described in Table 5.11.

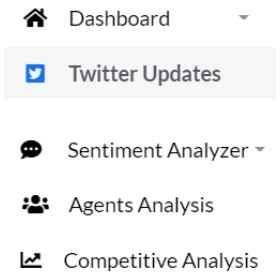
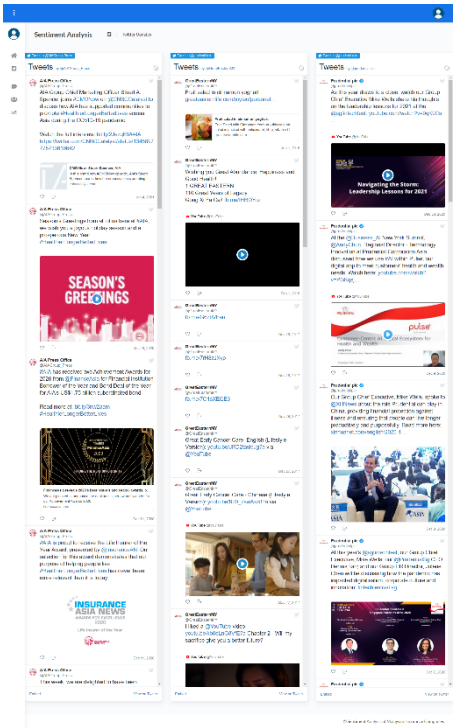
Table 5.11 Functionality Testing for Downloading Excel File

Test Objective	To check whether the user is able to download excel files with a labeled sentiment from each page of the insurance company using the download excel button shown in Figure 5.27.
Potential Test Inputs	 **Figure 5.27** Download Excel Button
Expected Test Outputs	The Excel file of Twitter data with labeled sentiment will be stored on the user's local machine, as in Figure 5.28.
Test Procedures	The user needs to click the 'Download Excel File' button to save the file.
Actual Test Results	 **Figure 5.28** Downloaded Excel file

5.3.5 View Twitter Updates

The result of the view Twitter updates test is described in Table 5.12.

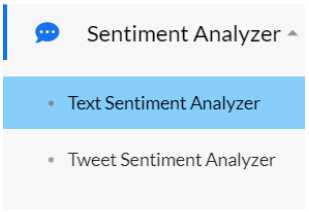
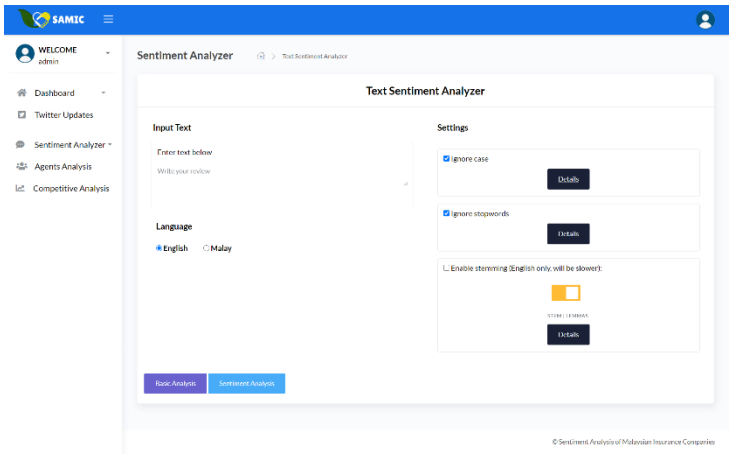
Table 5.12 Functionality Testing for View Twitter Updates Page

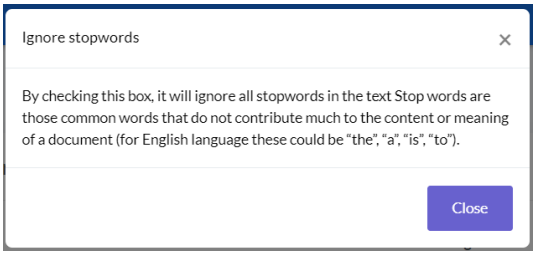
Test Objective	To test if the user is directed to the Twitter updates page when the user selects the page from drop-down menu options shown in Figure 5.29.
Potential Test Inputs	 **Figure 5.29** Drop-down Menu Options
Expected Test Outputs	The user can view the timeline of updates on Twitter by the insurance providers as in Figure 5.30.
Test Procedures	The user must use the navigation bar to select the View Twitter Updates page from the drop-down menu that contains links to other sections of the application.
Actual Test Results	 **Figure 5.30** Interface of View Twitter Updates Page

5.3.6 View Text Sentiment Analyzer Page

The result of the view text sentiment analyzer page test is described in Table 5.13.

Table 5.13 Functionality Testing for View Text Sentiment Analyzer Page

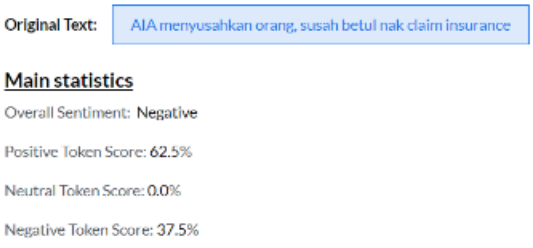
Test Objective	To observe if the user is directed to the text sentiment analyzer page when the user selects the page from drop-down menu options shown in Figure 5.31.
Potential Test Inputs	 **Figure 5.31** Drop-down Menu Options
Expected Test Outputs	The application displays the text sentiment analyzer page to the user and the details if the user clicks on the 'Details' button as in Figure 5.32 and Figure 5.33.
Test Procedures	1. Select the 'Text Sentiment Analyzer' page from the drop-down menu. 2. Click 'Details' to read more details about the text pre-processing task.
Actual Test Results	 **Figure 5.32** Interface of Text Sentiment Analyzer Page

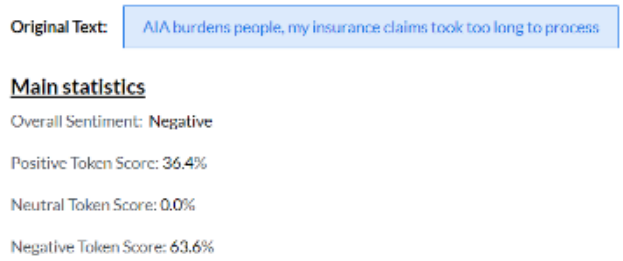
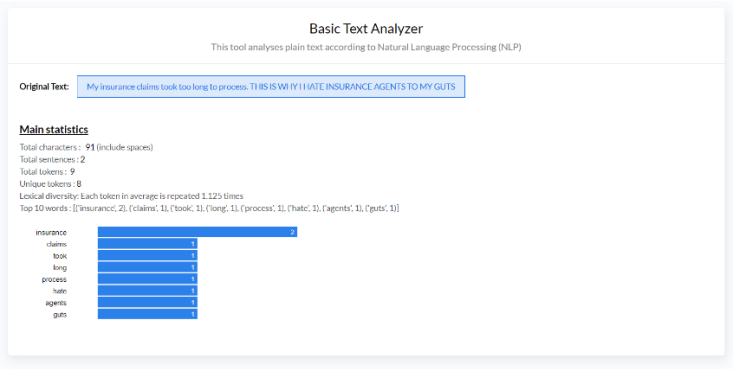
	 **Figure 5.33** View Details Button

5.3.7 Analyze Text

The result to analyze the sentiment of the text page is described in Table 5.14.

Table 5.14 Functionality Testing to Analyze Sentiment of Text

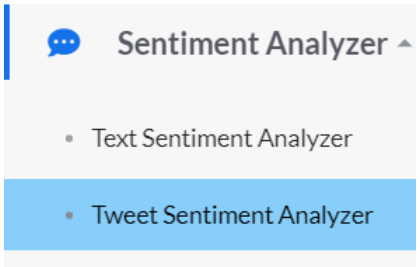
Test Objective	To determine if the user is able to use the machine learning model built to analyze text sentiments or basic analysis according to the processing of natural language using the button shown in Figure 5.34.
Potential Test Inputs	 **Figure 5.34** Button to Analyze Text
Expected Test Outputs	1. The user will be able to view the sentiment of text entered in Malay, as in Figure 5.35. 2. The user will be able to view the sentiment of text entered in English as in Figure 5.36. 3. The user will be able to view the result of natural language processing performed on the text entered as in Figure 5.37.
Test Procedures	1. Enter input text in the text box provided in English or Malay. 2. Choose to analyze the basic text or text sentiment.
Actual Test Results	 **Figure 5.35** Result of Sentiment Analysis Performed on Malay Text

	 **Figure 5.36** Result of Sentiment Analysis Performed on English Text  **Figure 5.37** Result of Basic Analysis Performed on Text

5.3.8 View Tweet Sentiment Analyzer Page

The result to view the tweet sentiment analyzer page test is described in Table 5.15.

Table 5.15 Functionality Testing to View Tweet Sentiment Analyzer Page

Test Objective	To observe if the user is directed to the tweet sentiment analyzer page when the user selects the page from drop-down menu options shown in Figure 5.38.
Potential Test Inputs	 **Figure 5.38** Drop-down Menu Options

Expected Test Outputs	The application displays the tweet sentiment analyzer page to the user where tweets are retrieved in real-time and sentiment analysis is conducted on the tweets in which the tweets' colored fields will represent the sentiment of the tweet. By default, the page will load tweets about 'Insurans' as in Figure 5.39. The user can also reply to the tweet or view the profile of the Twitter account
Test Procedures	The user must use the navigation bar to select the tweet sentiment analyzer page from the drop-down menu that contains links to other sections of the application
Actual Test Results	

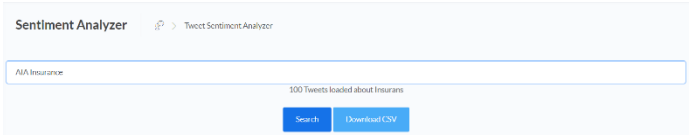
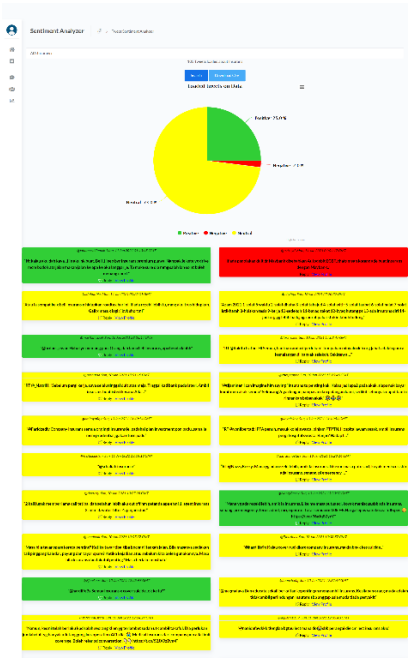
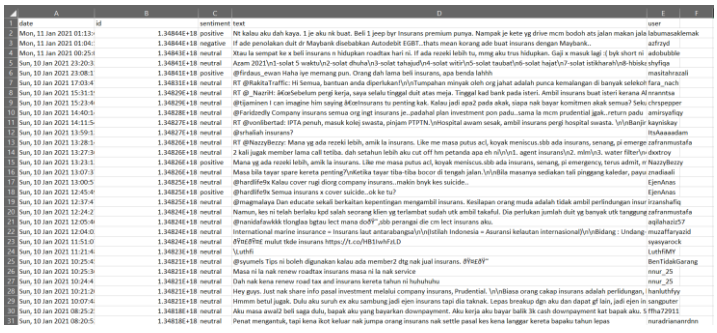
Figure 5.39 Interface of Tweet Sentiment Analyzer Page

5.3.9 Search Tweet Sentiment in Real-Time

The result to search tweet sentiment in real-time is described in Table 5.16.

Table 5.16 Functionality Testing to Search Tweet in Real-Time

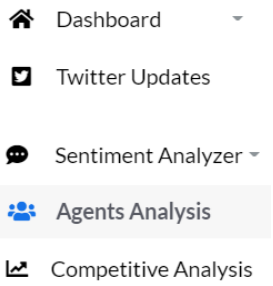
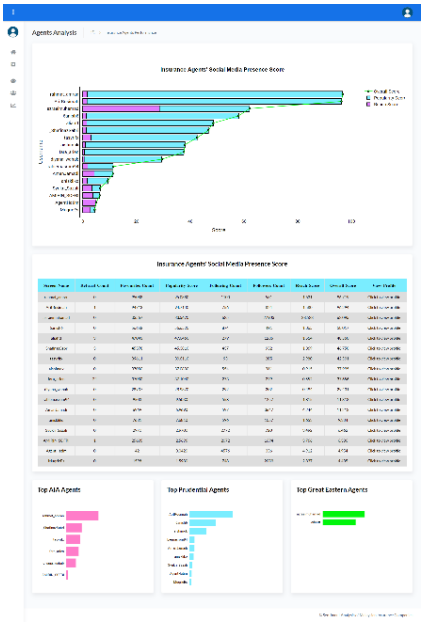
Test Objective	To test if the user can search for tweet keywords using the text box shown in Figure 5.40, get the tweet sentiment analyzed, and download the CSV file with the labeled sentiment.

Potential Test Inputs	 **Figure 5.40** User input Keywords in the Search Box
Expected Test Outputs	1. The user will see the sentiments of tweets represented by colored fields of the text box with 'positive' sentiment represented by green color, 'neutral' color represented by a yellow color, and 'negative' color represented by red color, as in Figure 5.41. 2. The user is able to download CSV files with searched tweets and labeled sentiment as in Figure 5.42.
Test Procedures	1. Enter keywords in the search box provided. 2. Click the 'Search' button symbol to proceed. 3. Click the 'Download CSV' button to download an excel file consist of searched tweets with the labeled sentiment.
Actual Test Results	 **Figure 5.41** Results of 'AIA Insurance' Search  **Figure 5.42** Downloaded CSV file

5.3.10 View Agents Analysis Page

The result to view the agents' analysis page test is described in Table 5.17.

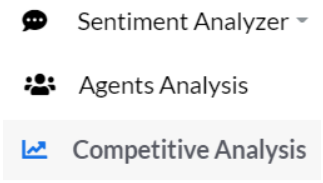
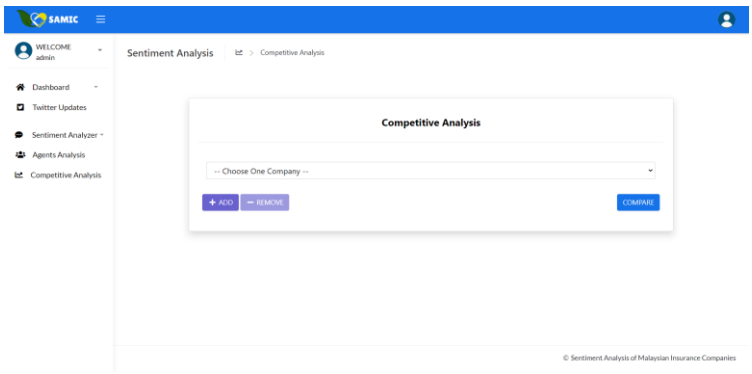
Table 5.17 Functionality Testing to View Agents Analysis Page

Test Objective	To observe if the user is directed to the agents' analysis page when the user selects the page from drop-down menu options shown in Figure 5.43.
Potential Test Inputs	 **Figure 5.43** Drop-down Menu Options
Expected Test Outputs	The application displays the view agents analysis page in Figure 5.44 to the user where the Twitter user profile accounts of insurance agents that actively promote the company's insurance plan have been extracted and visualized for all three companies.
Test Procedures	The user must use the navigation bar to select the view agents analysis page from the drop-down menu that contains links to other sections of the application.
Actual Test Results	 **Figure 5.44** Interface of View Agents Analysis Page

5.3.11 Compare Performance of Companies

The result to compare the performance of the companies page test is described in Table 5.18.

Table 5.18 Functionality Testing to Compare Performance of Companies

Test Objective	To observe if the user is directed to the competitive analysis page when the user selects the page from drop-down menu options shown in Figure 5.45.
Potential Test Inputs	 **Figure 5.45** Drop-down Menu Options
Expected Test Outputs	The application displays the competitive analysis page to the user in which the user can compare performance between companies of their choice using the drop-down list as in Figure 5.46. The results will then displayed as in Figure 5.47, Figure 5.48, Figure 5.49 and Figure 5.50.
Test Procedures	1. The user must use the navigation bar to select the competitive analysis page from the drop-down menu that contains links to other sections of the application. 2. Users must select companies to compare from the drop-down button. 3. Select the 'Compare' button to proceed.
Actual Test Results	 **Figure 5.46** Interface of Competitive Analysis Page

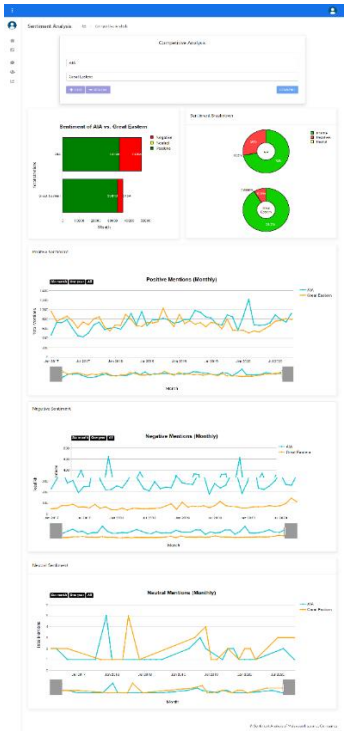


Figure 5.47 Results of Comparison between AIA and Great Eastern

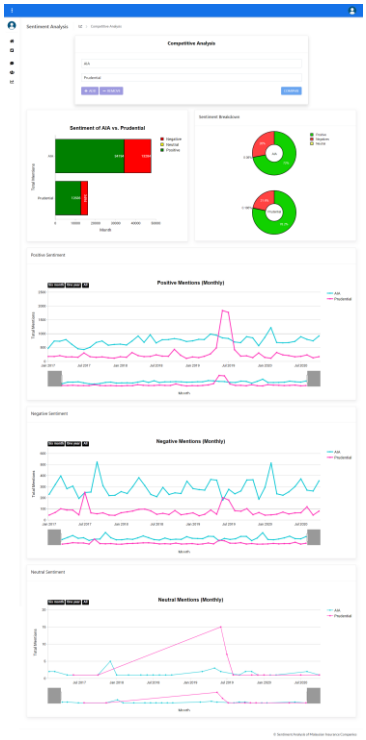


Figure 5.48 Results of Comparison between AIA and Prudential

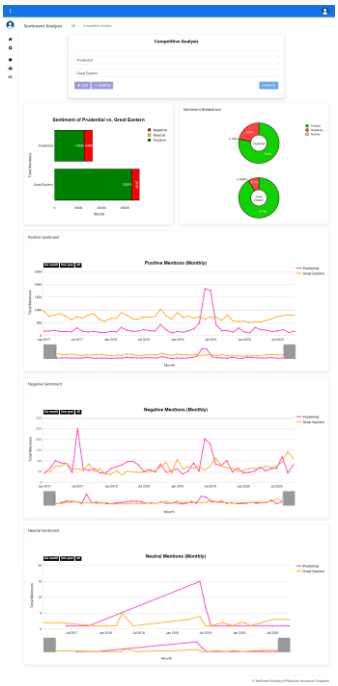


Figure 5.49 Results of Comparison between Prudential and Great Eastern

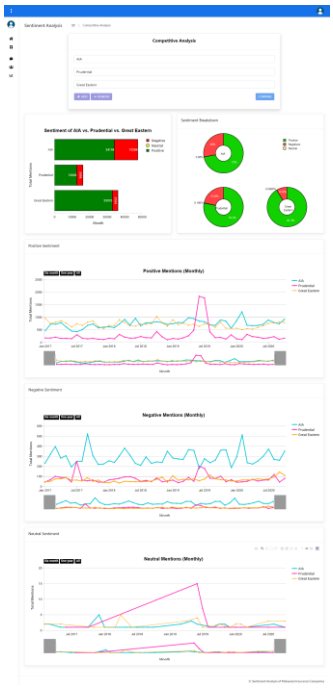
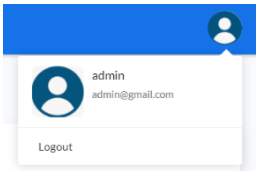
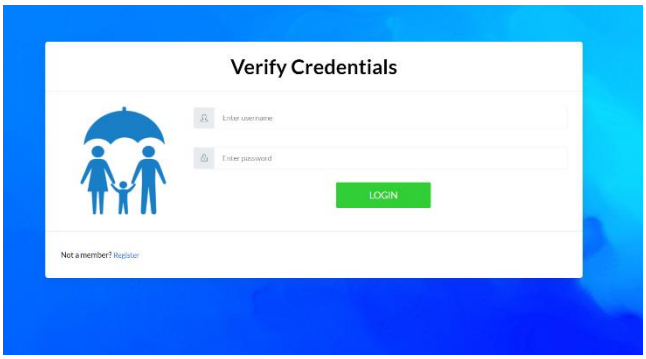


Figure 5.50 Results of Comparison between AIA, Prudential and Great Eastern

5.3.12 Logout Page

The result of the logout page test is described in Table 5.19.

Table 5.19 Functionality Testing for User Logout Page

Test Objective	To test whether the user can logout from the application using the logout button shown in Figure 5.51.
Potential Test Inputs	 **Figure 5.51** Logout Button
Expected Test Outputs	The user will log out of the application and redirect the application to the login page, as in Figure 5.52.
Test Procedures	Click the 'Log Out' button to proceed.
Actual Test Results	 **Figure 5.52** Interface of Log in Page

5.4 Usability Testing

Usability testing is a means of testing the application for a group of representative users to see how convenient and easy it is to use the application. Users are asked to test the application features while being observed to see where the users encounter problems and experience confusion. Recommendations would be made to solve these usability issues if users encounter similar problems. Approaches used in usability testing, including

system usability assessment and system usability result, will be discussed in the next subsection. The System Usability Scale is a Likert Scale consisting of ten short questions answered by the application users.

5.4.1 System Usability Assessment

The usability tests were conducted individually with the user using TeamViewer, where users gained remote access to the applications because of the Covid-19 situation, which prevented in-person usability testing from being conducted. Thirty users who have prior experience in purchasing insurance were selected to participate in the usability test and asked to complete some common tasks, such as evaluating each system's feature. The questionnaire included requests from users for information, including their name, status in the context of insurance, and the user's experience using the application. The steps taken to conduct the system usability test are illustrated as in Figure 5.53,

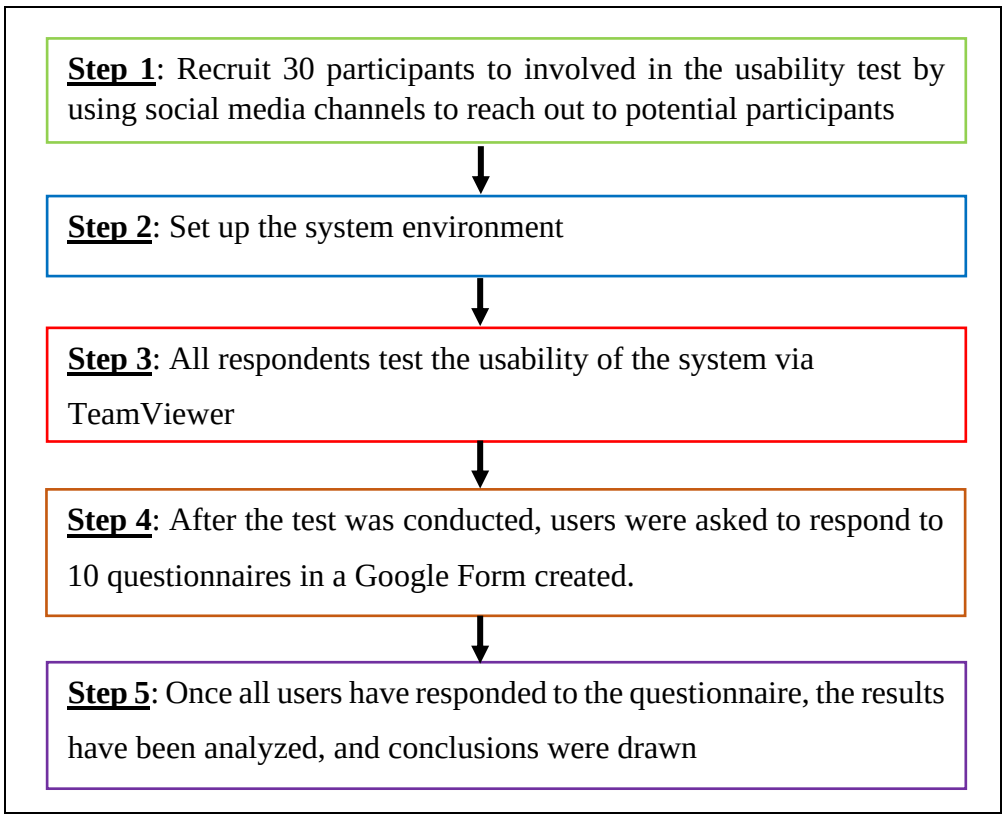


Figure 5.53 Steps in Conducting System Usability Test

The users responded to the ten questionnaires at the end of the evaluation. The questionnaire provided the following ten standard statements with five response options, indicating five Likert scales with anchors for strongly agree and strongly disagree, as in Table 5.20. Users have ranked each question from 1 to 5 based on how much they agree with the statement they read, and it is compulsory to fill in all the questions. Additionally, users are asked to be upfront if any aspect of the survey tools was unclear or confusing. The individual user scores are examined for any possible problems arising from confusion or inadvertent mistakes after the test. These responses are then used to generate a reliable overall usability score of the application.

Table 5.20 Usability Testing Questionnaire

Number	**Overall Reaction to the System**	**Scale**				
		1- strongly disagree, 2 - disagree, 3 - neutral, 4 - agree, 5 - strongly agree				
1	I think that I would like to use this feature frequently	1	2	3	4	5
2	I found the feature unnecessarily complex	1	2	3	4	5
3	I found this system was meaningful	1	2	3	4	5
4	I think that I would need the support of a technical person to be able to use the application’s feature	1	2	3	4	5
5	The interface of this system is pleasant	1	2	3	4	5
6	I am not able to navigate to other pages easily	1	2	3	4	5
7	The content on the system was relevant	1	2	3	4	5
8	The content presented was not in the right format	1	2	3	4	5
9	The system has all the expected functions and capabilities	1	2	3	4	5
10	Overall, I am not satisfied with this system	1	2	3	4	5

5.4.2 System Usability Scale Results

The system usability score (SUS) result is computed by summing the score contributions from each item. The users have ranked each of the ten system usability questions with items ranging from 1 to 5 based on their agreement level. All of the SUS results are illustrated in Table 5.21. The statements in odd-numbered questions, including question 1, question 3, question 5, question 7, and question 9, are expressed positively. The score contribution of the scale from 1 to 5 is subtracted by one from the score. On the other hand, the even-numbered questions such as question 2, question 4, question 6, question 8, and question 10 are phrased negatively. The contribution score will be subtracted by five from the score. These new values are added up as a total score contribution and multiplied by 2.5 to produce an average SUS score ranging from 0 to 100.

Table 5.21 SUS Results

Participants	1	2	3	4	5	6	7	8	9	10	SUS RAW Score	SUS Final Score
RESPONDENT 1	5	2	5	1	5	2	5	2	4	1	36	90
RESPONDENT 2	5	1	5	2	5	2	5	1	5	2	37	92.5
RESPONDENT 3	5	2	5	1	5	2	5	2	5	1	37	92.5
RESPONDENT 4	5	2	5	2	5	2	5	1	5	1	37	92.5
RESPONDENT 5	4	1	5	2	5	1	5	2	5	1	37	92.5
RESPONDENT 6	5	1	5	2	5	1	5	2	5	1	38	95
RESPONDENT 7	5	1	5	2	5	1	5	1	5	2	38	95
RESPONDENT 8	5	2	5	1	5	1	5	2	5	1	38	95
RESPONDENT 9	5	1	4	1	4	1	5	1	5	1	38	95
RESPONDENT 10	4	1	4	2	4	1	5	2	5	1	35	87.5
RESPONDENT 11	4	1	4	2	5	1	5	1	5	2	36	90
RESPONDENT 12	5	1	5	2	5	2	5	1	4	1	37	92.5
RESPONDENT 13	5	1	5	2	5	2	4	2	5	2	35	87.5
RESPONDENT 14	5	1	4	1	5	2	5	1	5	1	38	95
RESPONDENT 15	4	1	5	1	5	1	5	2	5	1	38	95
RESPONDENT 16	5	1	5	2	5	1	5	1	5	1	39	97.5
RESPONDENT 17	5	1	5	1	5	1	4	1	5	1	39	97.5
RESPONDENT 18	4	1	5	1	5	1	5	1	4	1	38	95
RESPONDENT 19	5	1	4	2	5	2	5	3	5	1	35	87.5
RESPONDENT 20	5	3	5	2	5	1	5	3	5	1	35	87.5
RESPONDENT 21	5	3	5	2	5	1	5	3	5	1	35	87.5
RESPONDENT 22	5	2	4	1	5	2	5	2	5	1	36	90
RESPONDENT 23	4	2	5	2	5	2	5	1	5	1	36	90
RESPONDENT 24	5	2	5	3	5	1	5	2	5	2	35	87.5
RESPONDENT 25	5	2	5	1	5	2	5	2	5	1	37	92.5
RESPONDENT 26	5	3	5	2	5	2	5	1	5	2	35	87.5
RESPONDENT 27	5	2	5	1	5	1	5	1	5	1	39	97.5
RESPONDENT 28	4	3	5	3	5	1	5	1	5	1	35	87.5
RESPONDENT 29	5	1	5	1	5	2	5	1	5	2	38	95
RESPONDENT 30	5	1	5	2	5	2	5	2	5	1	37	92.5
											AVERAGE	92.5

The bar chart diagram of the ten SUS statements scores is plotted, as in Figure 5.54. It displays the scale of SUS statements answered by the users in the questionnaire. The plotted graph shows that most users agreed with the odd-numbered question that reflects positive statements. It implies that the application has been evaluated as meaningful by the users. The users also believed that the application's interface is pleasant and has all the expected

functions and capabilities. Most users disagreed and had a neutral response for the even-numbered questions that indicate negative statements. It is an indication that users felt that they did not require technical assistance to be able to use the features of the application and easily navigate to other pages. Overall, most users were satisfied with the application.

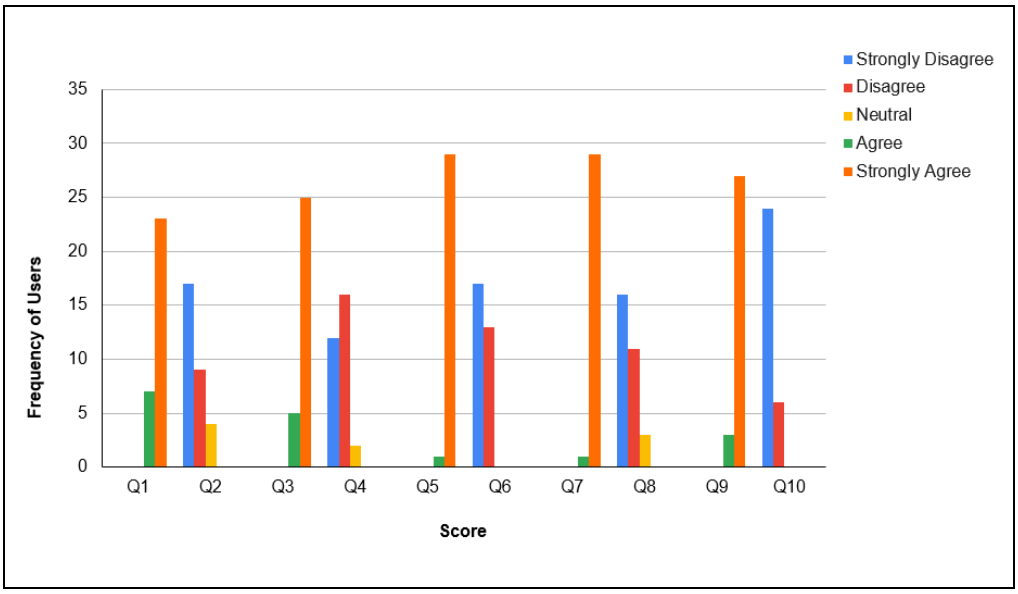


Figure 5.54 Bar Chart of SUS Result

Figure 5.55 shows the histogram of the SUS scores. The y-axis or the plotted histogram's vertical axis illustrates the frequency of users that answered the SUS, while the x-axis or the horizontal axis shows the percentage of the SUS score range. Based on the histogram, the data is spread between 88% to 98%. The plotted graph has a normal distribution with a range starting at 88%, and followed by a new range for every increase of 2%. The highest frequency is 92% to 94%, of which nine respondents fall into that range. The histogram is centered on the same value with the highest frequency of 92% to 94%. Ten respondents fall in the category below the central value, and eleven respondents lie in the above central value.

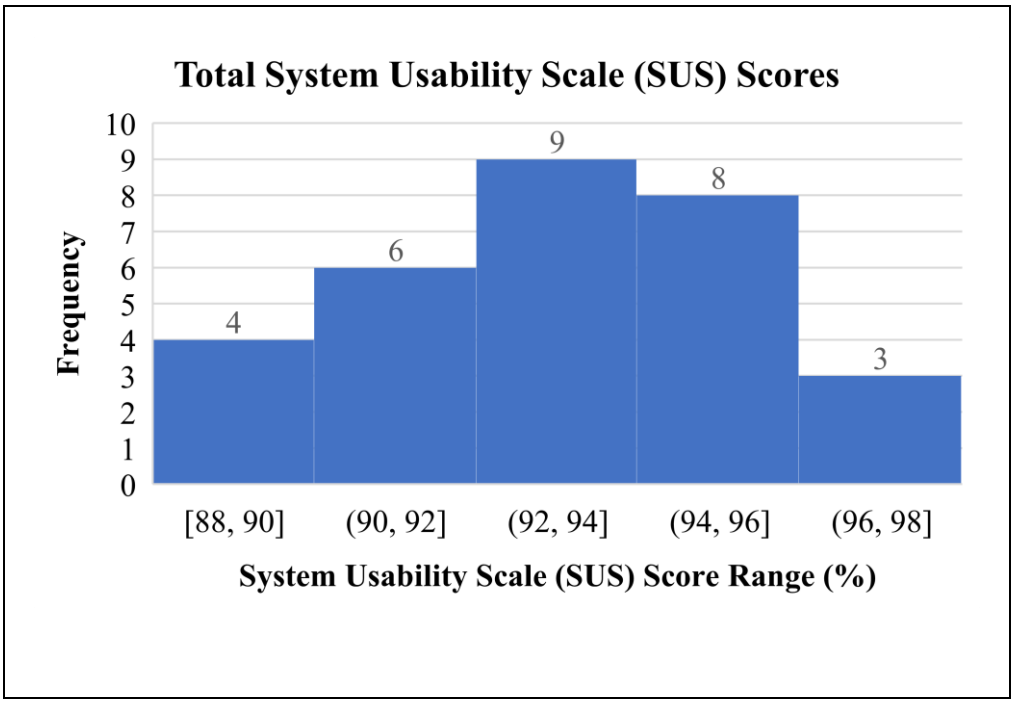


Figure 5.55 Histogram of SUS Scores

Overall, the SUS score of 92.5% was obtained from 30 users in the SUS questionnaire. The SUS average score's baseline is 68%, indicating that the system has average system usability. If the score is under 68, then there are probably significant issues with the application's usability that should be resolved. Whenever SUS exceeds 80%, the system has excellent usability (Alathas, 2018). To measure how the SUS scores are evaluated, 80.3 scores or higher is graded as an 'A' (Wang, 2021). Hence, by receiving an excellent score of more than 80%, this application is proven to have good usability. It is also noted that most of the testers gave good reviews about the application and would recommend it to the potential insurance buyers they knew.

5.5 Conclusion

To summarize, the results of the machine learning model were evaluated and compared with the baseline study. Besides, the analysis of real-world data is illustrated as graphs, and the insights gained have been discussed in this chapter. Two tests in the acceptance test have been carried out, including functionality testing and usability testing. This application's feature is tested in the functionality test to observe whether it operates as required by recording the expected test result and actual test result. It is observed that every feature in this application functions properly. Finally, thirty potential insurance buyers and policyholders were selected as testers to respond to the SUS scale questionnaire in usability testing.

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

This chapter discussed the conclusion, limitations of the project, and future recommendations that can be implemented to improve the application's quality. In this chapter, the overview of this project and its objectives are addressed, followed by the limitations and future recommendations for this project.

6.1 Conclusion

The visualization application developed is meant to assist the policyholders to understand the public views of the policy offered by the top insurance companies in Malaysia. In the analysis of public sentiments, a prototype application has been developed that uses Twitter data, in which the application acts as a medium to visualize the results of the sentiment analysis carried out on tweets that mentioned insurance companies. Hence, policyholders could make subtle and informed decisions to better understand outside information from real-world data. The support vector model used for the classification process has been incorporated into the application, allowing the user to use the model on any English and Malay textual data reviews.

The results are visualized in a dashboard to make the analysis outcomes readable and understandable by the policyholders. Interactive visualization provided in the application improved the way users interact with data by focusing on graphic representations of data. The dashboard also summarized most keywords that people discussed. Thus, the company can anticipate or mitigate the possible impact on their brand. This application's architecture and method are generic and can be easily adapted and extended to other domains. For instance, the application for gauging sentiments about various industrial can be visualized using the extracted real-world data. Additionally, this

application also includes real-time monitoring of tweets sentiment that uses the data processing infrastructure and sentiment analyzer to evaluate public sentiment changes in response to user search keywords

6.1.1 Objective I

This project's first objective is to design a web application system that can visualize Malaysia's best insurance company. The project's specifications were acquired by defining the problem statements, project objectives, scope, and significance of the project before proceeding to the design process. Besides, a survey was conducted before setting out the problem statements to identify the policyholders' difficulties during the decision-making process in purchasing insurance policies to establish a solid proof of this application's relevance. The design process was also assisted by the review of articles and journals from a variety of sources. The project requirements gathered have been converted into defined system functions for the intended project to ensure the completed application satisfies the criteria required for solving the specified problem statements. The details of the application's functionality and data exchanged between the system and the user are illustrated using a case diagram. As the designing process of both the Support Vector Classifier Model and the web application is completed, thereby the first objective was achieved.

6.1.2 Objective II

The second objective is to develop the designed system through Twitter sentiment analysis using the Support Vector Machine algorithm. During the design phase, the system's conceptual design has been developed into a working system that addresses all documented system requirements. In the development phase, text pre-processing was applied to the training and testing datasets. Irrelevant text is reduced from the corpus of tweets collected, and the document term matrix is generated. Additionally, before applying a machine-learning algorithm to extract data features, the term frequency-inverse document frequency (TF-IDF) concept was applied. Data are collected from

Twitter to obtain real-world datasets to be classified. In order to portray the sentiment generation, a support vector classifier is used to classify tweets into positive, neutral, and negative sentiments. The classification results are visualized and presented through a web application. Thus, the second objective was accomplished.

6.1.3 Objective III

The third objective is to test the functionality and usability of the system. Once the system has been completed, it is deployed in the testing environment. During functionality testing, the entire system's functionality has been tested and verified to ensure that it operates according to the requirements obtained in the previous phase. By entering the input and examining the output, the functions included in the test cases are tested. It is observed that the functions are successful, and all functionality of the system performs as expected.

On the other hand, the usability test was carried out to learn about the system workflow and whether the information is correctly formatted and delivered. The system produces an excellent result with a cumulative average of 92 for usability testing, which implies that the user is satisfied with the system. Thus, it is concluded that the third objective was fulfilled.

6.2 Project Limitations

This project has several shortcomings due to the limitation of knowledge and time constraints, which have impeded the application quality delivery. This project's most significant limitation is limited access to the publicly accessible neutral dataset in Malay and English language. The uneven distribution of neutral tweets compared to the negative and positive tweets results in an imbalanced dataset that reduced the classifier's overall performance. Although an approach has been used to increase the number of neutral training instances to even out the three classes, the overall accuracy is still not sufficient to be deployed for real-world data. In the context of the Twitter environment, data

analysis is limited to the extent that platform owners are willing to share their data or that data can be legally mined. The tweets which are in a private setting are not retrievable through the Twitter API.

Additionally, it is observed in the real-time sentiment analysis feature that the Twitter API provides a limited search capacity concerning the volume and time frame of returned data. In the context of limitations on sentiment analysis, the issue is that the automated application does not add information in context and often fails to grasp nuances of the human language, notably sarcasm. Memes or graphics that appear with the post are not analyzed by text-based sentiment analysis. Besides, social media slang will evolve quickly, and the syntax of sentences varies from the conventional style of conversation.

6.3 Project Recommendations

For future recommendation, it is suggested that the quantity of neutral dataset be increased in multiclass classification, hoping that there will be a publicly available neutral dataset in Malay and English language so that more accurate results can be obtained to optimize the performance of the classifier. Besides, this application's expansion could explore more features of PySpark and take advantage of its vast large-scale data processing modules. Doing so makes it possible to optimize the classifier's overall performance by adopting different techniques to build the classifiers and tackle large pre-processing and modeling classifiers for a more computationally efficient solution. It also provides parallel execution on all computer cores.

Furthermore, it is also recommended to purchase Twitter premium APIs as it offers higher rate limits and more complex queries that are great for developing more robust solutions with the additional webhook connections and scaling subscriptions to access more accounts activities and enhance the quality and functionality of the real-time tweets analyzer of the application. Finally, the research study is not limited to insurance companies. It is also possible to collect and use data from social media campaigns of different industries for

buzz tracking to track mentions of a business to help sustain a positive image. Social media monitoring is a way that companies should use to categorize emergency problems instantly and deal with them quickly by digging into all of the consumer's social media views about their brand.

REFERENCES

Ahmad, M., Aftab, S., Muhammad, S., & Ahmad, S. (2017). Machine learning techniques for sentiment analysis: A review. *International Journal of Multidisciplinary Sciences and Engineering*, *8*(3), 27–32.

Alaoui, I. El., & Gahi, Y. (2019). The impact of big data quality on sentiment analysis approaches. *Procedia Computer Science*, *160*, 803–810.

Alathas, H. (2018). How to measure product usability with the system usability scale (SUS) score. Retrieved Jan 25, 2021, from UX Planet website: https://uxplanet.org/how-to-measure-product-usability-with-the-system-usability-scale-sus-score-69f3875b858f

Ali, N. A. B. M. A., & Sari, M. N. S. (2015). The suitability of native application for university e-learning compared to web-based application. *International Journal of Science and Research (IJSR)*, *4*(1), 2045–2049.

Alsolamy, A., Siddiqui, A. A., & Khan, M. H. (2019). A corpus based approach to build arabic sentiment lexicon. *International Journal of Information Engineering and Electronic Business*, *11*(6), 16–23.

Amrani, Y., Lazaar, M., & Kadirp, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, *127*, 511–520.

Angelova, B., & Zekiri, J. (2011). Measuring customer satisfaction with service quality using American customer satisfaction model (ACSI Model). *International Journal of Academic Research in Business and Social Sciences*, *1*(3), 27.

Bank Negara Malaysia (BNM). (2016). *Bank negara annual report 2016*, 109-110.

Bao, N. J., Ramlan, R., Mohamad, F., & Yassin, A. M. (2018). Performance of Malaysian insurance companies using data envelopment analysis. *Indonesian Journal of Electrical Engineering and Computer Science*, *11*(3), 1147–1151.

Ben S.B (2015). How to design scatter plots. Retrieved May 25, 2020, from Visage website: https://visage.co/data-visualization-101-scatter-plots/

Bouazizi, M., & Ohtsuki, T. (2017). A pattern-based approach for multi-class sentiment analysis in Twitter. *IEEE Access*, *5*(May 2020), 20617–20639.

Bovbjerg, R. R., & Hadley, J. (2007). Why health insurance is important. *Health Policy Briefs. The Urban Institute. Washington, DC*, (11), 3.

Broek-Altenburg, E. M., & Atherly, A. J. (2019). Using social media to identify consumers' sentiments towards attributes of health insurance during enrollment season. *Journal of Applied Sciences (Switzerland)*, *9*(10).

Cardieri, G. de A., & Zaina, L. M. (2018). Analyzing user experience in mobile web, native and progressive web applications: A user and HCI specialist perspectives. *ACM International Conference Proceeding Series*.

Casteren, W. Van. (2017). The Waterfall Model And The Agile Methodologies : A comparison by project characteristics-short the waterfall model and agile methodologies. *Academic Competences in the Bachelor*, (February), 10–13.

Chlasta, K., Wołk, K., & Krejtz, I. (2019). Automated speech-based screening of depression using deep convolutional neural networks. *Procedia Computer Science*, *164*, 618–628.

Deng, Z., & Wang, Z. (2016). Early-mover advantages at cross-border business-tobusiness e-commerce portals. *Journal of Business Research*, *69*(12), 6002–6011.

Dey, A., & Bandyopadhyay, S. (2016). Automated glaucoma detection using support vector machine classification method. *British Journal of Medicine and Medical Research*, *11*(12), 1–12.

Dossey, A. (2019). A guide to mobile app development: Web vs. Native vs. Hybrid. Retrieved June 11, 2020, from Clear Bridge Mobile website: https://clearbridgemobile.com/mobile-app-development-native-vs-web-vshybrid/

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, *161*, 707–714.

Erkkilä, J.-P. (2013). *Web and native technologies in mobile application development*. Aalto University.

Ghaleb, O. A. M., & Vijendran, A. S. (2017). The challenges of sentiment analysis on social web communities. *International Journal of Advance Research in Science and Engineering (IJARSE)*, 117–125.

Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., & Badhani, P. (2017). Study of Twitter sentiment analysis using machine learning algorithms on python. *International Journal of Computer Applications*, *165*(9), 29–34.

Half, R. (2018). 6 basic SDLC methodologies : Which one is best? Retrieved May 25, 2020, from 2018 website: https://www.roberthalf.com/blog/salaries-and-skills/6basic-sdlc-methodologies-which-one-is-best

Hamid, M. A., Osman, J., & Nordin, B. A. A. (2009). Determinants of corporate demand for islamic insurance in Malaysia. *International Journal of Economics and Management*, *3*(2), 278–296.

HBS Online. (2019). 9 data visualization techniques all professionals should know. Retrieved May 25, 2020, from https://online.hbs.edu/blog/post/datavisualization-techniques

Hegde, Y., & Padma, S. K. (2017). Sentiment analysis using random forest ensemble for mobile product reviews in Kannada. *International Advance Computing Conference, 777-782*.

Heitkotter, Sebastian, H., Hanschke, S., & Majchrzak, T. A. (2013). Evaluating crossplatform development. *Springer Berlin Heidelberg*, 120–121.

Highcharts. (2011). Line chart. Retrieved May 25, 2020, from Springer Reference website: https://www.highcharts.com/docs/chart-and-series-types/line-chart

Hourrane, O.H., Benlahmar, E. H., & Zellou, A. (2018). Comparative study of deep learning models for sentiment analysis. *International Journal of Engineering & Technology*, *7*(2.14), 5726.

JavaTPoint (2020). K-Nearest Neighbor (KNN) algorithm for machine learning. Retrieved May 25, 2020, from https://www.javatpoint.com/k-nearest-neighboralgorithm-for-machine-learning

Karthika, P., Murugeswari, R., & Manoranjithem, R. (2019). Sentiment analysis of social media network using random forest algorithm. *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2019*, 1–5.

Kumar, D., Shah, F., & Manab, N. (2018). Service quality of social insurance industry and customer satisfaction from Malaysian perspective: Customer knowledge as a moderator . International Academic Research Journal of Business and Technology, 4(1), 1–7.

Lohmann, S., Heimerl, F., Bopp, F., Burch, M., & Ertl, T. (2015). Concentri cloud: Word cloud visualization for multiple text documents. *Proceedings of the International Conference on Information Visualisation*, (November 2015), 114–120.

Lynch, W. (2018). What is the role of waterfall model in sdlc? Retrieved Jan 31, 2021, from Medium website: https://warren2lynch.medium.com/what-is-the-problems-of-waterfall-model-38de858f1058

Masud, M. M., Rana, M. S., Mia, M. A., & Saifullah, M. K. (2019). How productive are life insurance institutions in Malaysia: A malmquist approach. *Journal of Asian Finance, Economics and Business*, *6*(1), 241–248.

McDougall, G. H. G., & Levesque, T. (2000). Customer satisfaction with services: putting perceived value into the equation. *Journal of Services Marketing*, *14*(5), 392–410.

McMaken, L. (2019). Types of insurance everyone needs. Retrieved May 25, 2020, from https://www.investopedia.com/financial-edge/0212/4-types-of-insuranceeveryone-needs.aspx

Muhamat, A. A., Karim, N. A., Mainal, S. A., Alwi, S. F. S., & Jaafar, M. N. (2018).

Determinants of agents performance: A case study of AmMetLife Malaysia Berhad. *International Journal of Academic Research in Business and Social Sciences*, *8*(11), 768–778.

Muhammad, F., Karim, A., & Jamal, J. (2012). Takaful and conventional insurance in Malaysia: An overview. *Journal of Contemporary Issues and Thought*, *2*(1), 82–92.

Naeni, O., & Sari, R. A. (2017). Comparison of data mining methods for recipient prediction poor student assistance (BSM) in MAN 2 North Lampung. *3rd International Conferences on Information Technology and Business (ICITB)*, (7th Dec 2017), 207–213.

Navlani, A.N. (2019). KNN classifcation using scikit-learn curse of dimensionality. Retrieved May 25, 2020, from https://www.datacamp.com/community/tutorials/k-nearest-neighborclassification-scikit-learn

Neves-Silva, R., Gamito, M., Pina, P., & Campos, A. R. (2016). Modelling influence and reach in sentiment analysis. *Procedia Cooperative Institutional Research Program*, *47*, 48–53.

Nikhil Bodapati. (2017). Visualizing text analysis results with word clouds. Retrieved May 25, 2020, from Dundas BI website: https://www.dundas.com/support/blog/visualizing-text-analysis-results-withword-clouds

Nuruzzaman, M., & Hussain, O. K. (2018). A survey on chatbot implementation in customer service industry through deep neural networks. *International Conference on E-Business Engineering, ICEBE 2018*, 54–61.

Nuzzo, R. (2016). The box plots alternative for visualizing quantitative data. *The American Academy of Physical Medicine and Rehabilitation*, *8*(3), 268–272.

Piatetsky, G. (2020). Top trends in analytics and big data. Retrieved May 25, 2020, from https://www.kdnuggets.com/2014/01/top-trends-analytics-big-data-strata2014-santa-clara.html

Prashant (2020). What is waterfall model in software engineering? Retrieved Jan 31,

2021, from The Study Genius website: https://www.thestudygenius.com/what-is-waterfall-model/

Ratanadewi, S., & Bachtiar, M. (2020). Inventory and order system development at PT.X. *IOP Conference Series: Materials Science and Engineering*, *847*(1).

Reis, I., Baron, D., & Shahaf, S. (2018). Probabilistic random forest: A machine learning algorithm for noisy data sets. *The Astronomical Journal*, *157*(1), 16.

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. *Information Processing and Management*, *52*(5), 949–975.

Roosevelt, Jr. Mosley, C. (2012). Social media analytics: Data mining applied to insurance twitter posts. *Causality Actuarial Society E-Forum*, *2*, 1–27.

Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from twitter text. *Journal of Computational Science*, *36*, 18-19.

Salman, S. S., Khan, M. U. K., & Iqbal, M. Z. I. (2019). *A systematic mapping study on testing of machine learning programs.*

Samah, K. A. F. A., Badarudin, I. M., Odzaly, E. E., Ismail, K. N., Nasarudin, N. I. S., Tahar, N. F., & Khairuddin, M. H. (2019). Optimization of house purchase recommendation system (HPRS) using genetic algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, *16*(3), 1530–1538.

Samuel, S. A., & Anthonia, A. (2016). An overview of big data visualization techniques in data mining. *International Journal of Computer Science and Information Technology Research*, *4*(3), 105–113.

Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Wai, P.S., Chung, S.Y., Piprani, M. A. (2018). Sentiment analysis of big data: Methods, applications, and open challenges. *Institute of Electrical and Electronics Engineers (IEEE) Access*, *6*, 21-23.

Smith, M. S., & Hayhoe, C. R. (2005). Life insurance: Whole-life insurance. *Virginia Cooperative Extension*, 354-145, 1-3.

Sum, R., & Nordin, N. (2018). Decision-making biases in insurance purchasing. *Journal of Advanced Research in Social and Behavioural Sciences*, *10*(1), 165–179.

Tang, J. (2017). The theory and method of sentiment analysis approaches for application in the big data frameworks. *DEStech Transactions on Computer Science and Engineering*, (ICICEE 2017), 396–402.

Thangaraj, M., & Sivakami, M. (2018). Text classification techniques: A literature review. *Interdisciplinary Journal of Information, Knowledge, and Management*, *13*, 117–135.

Vidhya, A. (2016). Stacked bar chart. Retrieved May 25, 2020, from https://www.analyticsvidhya.com/blog/2016/03/questions-ggplot2-packager/rplot78-2/

Wahome, R. (2020). This is how Twitter sees the world : Sentiment analysis part one. Retrieved Jan 25, 2021, from https://towardsdatascience.com/the-real-world-as-seen-on-twitter-sentiment-analysis-part-one-5ac2d06b63fb

Wang, Y., Chi, R., & Yang, Y. (2004). An integrated framework for customer value and customer-relationship-management performance: A customer-based perspective from China. *Managing Service Quality: An International Journal*, 169–182.

Wang, Y. (2021). System usability scale : A quick and efficient user study methodology. Retrieved Jan 25, 2021, from http://ixd.prattsi.org/2018/04/system-usability-scale-a-quick-and-efficient-user-study-methodology/

Womack, R. (2015). Data visualization and information literacy. *International Association for Social Science Information Service and Technology Quarterly*, *38*(1), 12.

Yaakub, M. R., Latiffi, M. I. A., & Zaabar, L. S. (2019). A review on sentiment analysis techniques and applications. *IOP Conference Series: Materials Science and Engineering*, *551*(1).

YouGov-BrandIndex. (2019). Malaysia top rankings index rankings by country.

Retrieved May 25, 2020, from https://www.brandindex.com/ranking/malaysia/2019-index/category/insurance

Zerium, A. Z. (2019). Artificial neural network explained. Retrieved May 25, 2020, from https://blog.goodaudience.com/artificial-neural-networks-explained436fcf36e75

Zulkifly, N.A.Q.Z., Kasim, Z. K., & Bidin, J. B. (2019). Selection of personal medical and health insurance company by using fuzzy topsis. *Jurnal Intelek, 14*(1).

APPENDICES

APPENDIX A

SURVEY

Sentiment Analysis of Malaysian Insurance Companies (SAMIC): A Visualization and Prediction using Support Vector Machine Algorithm

Assalamualaikum and Salam Sejahtera. I am Nur Farhana Binti Ahmad from Faculty of Computer and Mathematical Sciences at Universiti Teknologi Mara (UiTM) Cawangan Melaka, Kampus Jasin. I am currently pursuing a degree program in Bachelor of Computer Science (HONS.) and are required to conduct a study for my Final Year Project (FYP) on the above-mentioned title.

This survey is an initiative form to gather information for subject Project Formulation (CSP600) and is aimed specifically for people who are interested in buying insurance or already had one. I would really appreciate it if you can spend a few minutes answering this survey.

May the kindness you spread keep returning to you. Thank you !

* Required

Q1) Which of the following is your status? *

- ◯ A student
- ◯ An employee

Q2) Do you have/bought any personal insurance policy? *

- ◯ Yes
- ◯ No

Q3) Do you face any problem or difficulties in choosing the insurance companies in Malaysia to purchase an insurance policy? *

- ○ Yes
- ○ No

Q4) Which of the following is/are the problem/difficulties that you faced? *

- ☐ Needing a trusted advisor
- ☐ Finding the proper insurance policy
- ☐ Choosing an insurance company that has a financial solidity
- ☐ Other: _____

Q5) Which of the following method that you most likely to refer before purchasing an insurance policy?

- ☐ Online reviews of an insurance company
- ☐ Brochure distributed by the insurance companies
- ☐ The insurance agents
- ☐ Other: _____

Q6) During the surveying process to purchase the best insurance, do you agree that it was time-consuming to review the company's reputation and history?

- ○ Yes
- ○ No

Q7) Based on your experience, what do you expect the most from insurance companies?

- [ ] Good customer service
- [ ] Ability to promptly handling claims
- [ ] Reliable insurance policy quotes
- [ ] Recommendation of sufficient insurance protection
- [ ] Other: _____

Q8) In your opinion, which of the following is the best insurance company in Malaysia?

- ○ AIA
- ○ Prudential
- ○ Great Estern

Q9) If there is a platform that helps you with the decision making of purchasing insurance based on review, do you think it will be useful to overcome the current practise? *

- ○ Yes
- ○ No

Q10) Why do you think so?

Your answer

System Usability Testing for SAMIC Application

This survey is an initiative form to gather information about the system usability scale and is aimed specifically for people who have been selected to participate in the usability testing phases.

*All the responses and data gathered in this survey is for research purposes only and will be kept confidential and remain private. Your cooperation is highly appreciated.
Thank You !

* Required

Tester Name *

Your answer

Status *

- ◯ Potential Insurance Buyer
- ◯ Policyholder

1) I think that I would like to use this feature frequently *

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

2) I found the feature unnecessarily complex *

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

3) I found this system was meaningful *

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

4) I think that I would need the support of a technical person to be able to use the application's feature *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

5) The interface of this system is pleasant *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

6) I am not able to navigate to other pages easily *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

7) The content on the system was relevant *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

8) The content presented was not in the right format *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

9) The system has all the expected functions and capabilities *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

10) Overall, I am not satisfied with this system *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

APPENDIX B

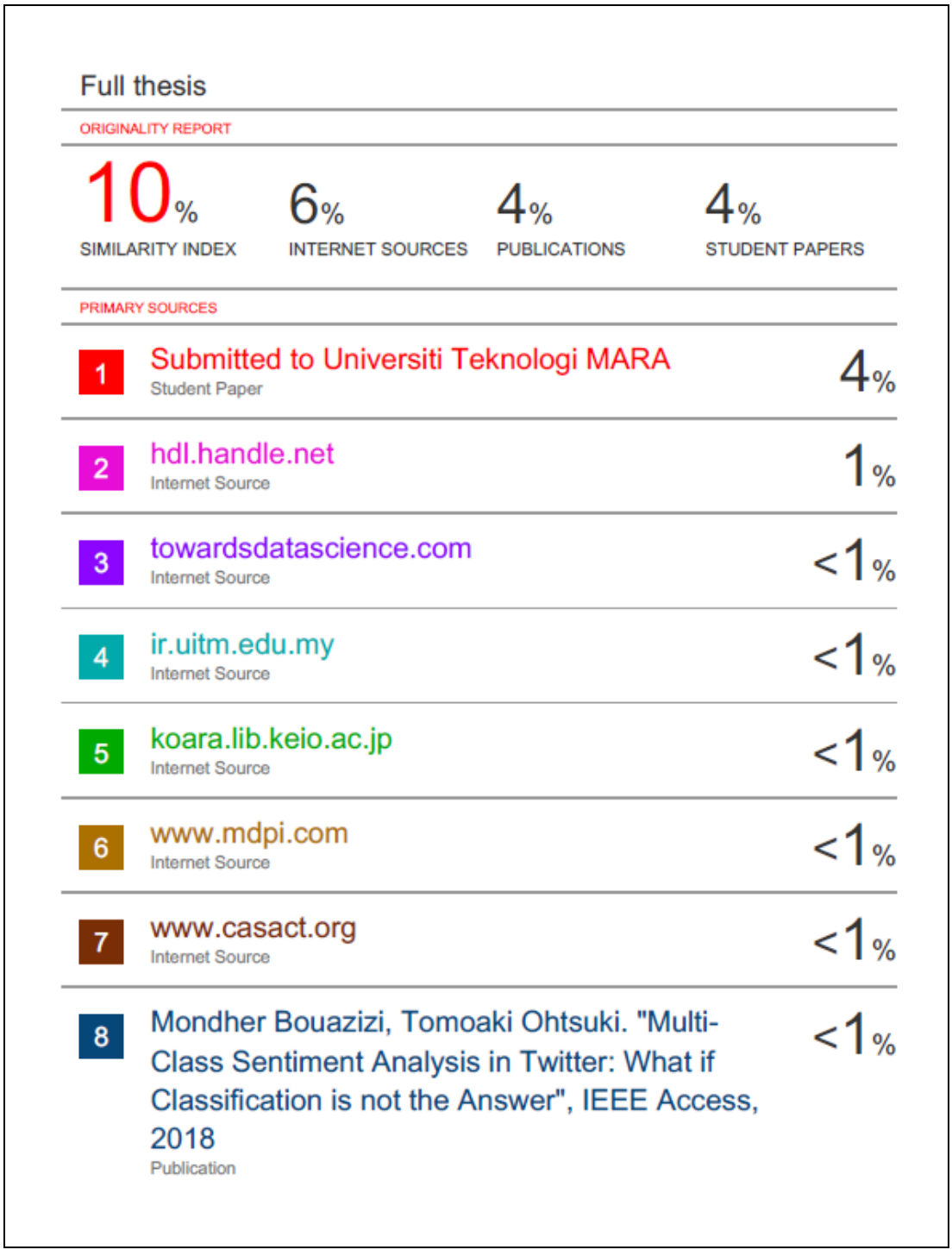
GANTT CHART

CSP600 Gantt Chart														
Description	Week													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Proposal Outline Form(F2)														
Mutual Acceptance Form(F1)														
Writing Chapter 1														
Chapter 1 Submission														
Assignment 1														
Writing Chapter 2														
Chapter 2 Submission														
Assignment 2														
Writing Chapter 3														
Chapter 3 Submission														
Plagiarism Checking														
Proposal Presentation and Submission														
Project Proposal Project														

CSP650 Gantt Chart														
Description	Week													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Develop Prediction Model														
Data Collection														
Data Transformation														
Accuracy Testing of Prediction Model Data														
Develop Interfaces														
Develop Decision Support System														
System Testing														
Writing Chapter 4														
Writing Chapter 5														
Writing Chapter 6														
Full Report														
Plagiarism Checking														

APPENDIX C

TURN-IT-IN RESULT



Full thesis

ORIGINALITY REPORT

10%	6%	4%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universiti Teknologi MARA Student Paper	4%
2	hdl.handle.net Internet Source	1%
3	towardsdatascience.com Internet Source	<1%
4	ir.uitm.edu.my Internet Source	<1%
5	koara.lib.keio.ac.jp Internet Source	<1%
6	www.mdpi.com Internet Source	<1%
7	www.casact.org Internet Source	<1%
8	Mondher Bouazizi, Tomoaki Ohtsuki. "Multi-Class Sentiment Analysis in Twitter: What if Classification is not the Answer", IEEE Access, 2018 Publication	<1%

9	Shahid Shayaa, Noor Ismawati Jaafar, Shamshul Bahri, Ainin Sulaiman et al. "Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges", IEEE Access, 2018 Publication	<1%
10	www.irjet.net Internet Source	<1%
11	Eline M. van den Broek-Altenburg, Adam J. Atherly. "Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season", Applied Sciences, 2019 Publication	<1%
12	arrow.tudublin.ie Internet Source	<1%
13	www.frontiersin.org Internet Source	<1%
14	"Emerging Trends in Computing and Expert Technology", Springer Science and Business Media LLC, 2020 Publication	<1%
15	dspace.alquds.edu Internet Source	<1%
16	"Proceedings of International Conference on	<1%

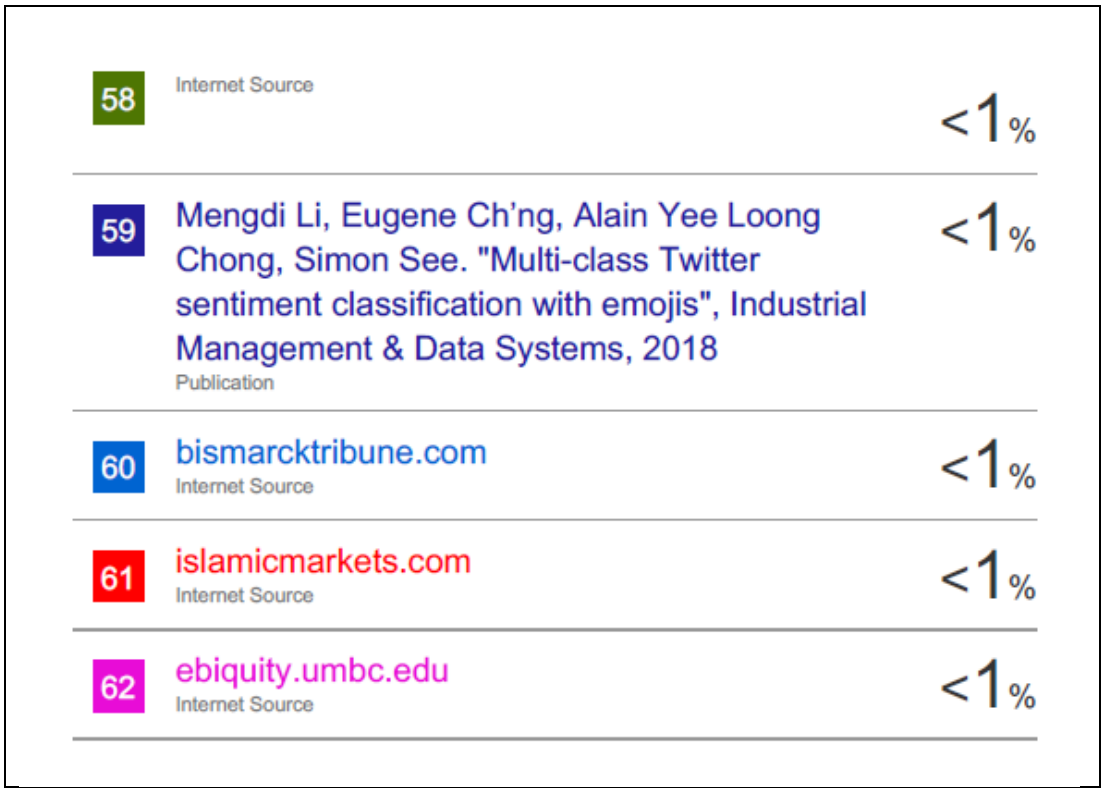
	Intelligent Computing, Information and Control Systems", Springer Science and Business Media LLC, 2021 Publication	
17	portal.fke.utm.my Internet Source	<1%
18	www.legalsectoralliance.co.uk Internet Source	<1%
19	Kuo-Cheng Kuo, Qian Long Kweh, Irene Wei Kiong Ting, Noor Azlinna Azizan. "Dynamic network performance evaluation of general insurance companies: an insight into risk management committee structure", Total Quality Management & Business Excellence, 2015 Publication	<1%
20	"Advances in Computing and Data Sciences", Springer Science and Business Media LLC, 2018 Publication	<1%
21	S de Yong, Y Kusumarini, P E D Tedjokoesoemo. "Interior design students' perception for AutoCAD, SketchUp and Rhinoceros software usability", IOP Conference Series: Earth and Environmental Science, 2020 Publication	<1%
22	link.springer.com Internet Source	<1%

23	www.researchgate.net Internet Source	<1%
24	O. F. W. Onifade, M. A. Malik. "SASM: A tool for sentiment analysis on Twitter", 2015 2nd World Symposium on Web Applications and Networking (WSWAN), 2015 Publication	<1%
25	www.coursehero.com Internet Source	<1%
26	cs229.stanford.edu Internet Source	<1%
27	www.ijitee.org Internet Source	<1%
28	doit.maryland.gov Internet Source	<1%
29	Alotaibi , Ashwaq Awad. "Security Issues in NoSQL Databases : Access Control (Authorization) for NoSQL Databases = المشاكل الأمنية في قواعد البيانات (NoSQL) : التحكم في الوصول لقواعد بيانات (صلاحية) (NoSQL)", King Abdulaziz University : Scientific Publishing Centre, 2020 Publication	<1%
30	Bayo Lawal. "Applied Statistical Methods in Agriculture, Health and Life Sciences", Springer Science and Business Media LLC, 2014 Publication	<1%

31	en.wikipedia.org Internet Source	<1%
32	medium.com Internet Source	<1%
33	repository.asu.edu Internet Source	<1%
34	ijcsi.org Internet Source	<1%
35	Hui Shan Lee, Fan Fah Cheng, Wai Mun Har, Annuar Md Nassir, Nazrul Hisyam Ab Razak. "Efficiency, firm-specific and corporate governance factors of the Takaful insurance", International Journal of Islamic and Middle Eastern Finance and Management, 2019 Publication	<1%
36	scikit-learn.org Internet Source	<1%
37	Aleksa Vukotic, James Goodwill. "Apache Tomcat 7", Springer Science and Business Media LLC, 2011 Publication	<1%
38	www.northdevon.gov.uk Internet Source	<1%
39	utpedia.utp.edu.my Internet Source	<1%

40	ivythesis.typepad.com Internet Source	<1%
41	"Computational Intelligence in Data Mining", Springer Science and Business Media LLC, 2019 Publication	<1%
42	ixd.prattsi.org Internet Source	<1%
43	scitepress.org Internet Source	<1%
44	wiredspace.wits.ac.za Internet Source	<1%
45	Muhammad Ali Hassan, Nhamo Mtetwa. "Feature Extraction and Classification of Spam Emails", 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI), 2018 Publication	<1%
46	Pratijnya Ajawan, Pooja Desai, Veena Desai. "Smart Sampark-An approach towards building a responsive system for Kisan Call Center", 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC), 2020 Publication	<1%
47	api.research-repository.uwa.edu.au Internet Source	<1%

48	nebula.wsimg.com Internet Source	<1%
49	spark.apache.org Internet Source	<1%
50	universaldebatingproject.blogspot.com Internet Source	<1%
51	www.freepatentsonline.com Internet Source	<1%
52	www.cabafx.com Internet Source	<1%
53	Steffi Ratanadewi, Marsellinus Bachtiar. "Inventory and Order System Development at PT.X", IOP Conference Series: Materials Science and Engineering, 2020 Publication	<1%
54	thesesups.ups-tlse.fr Internet Source	<1%
55	edocs.fu-berlin.de Internet Source	<1%
56	e-space.mmu.ac.uk Internet Source	<1%
57	www.ijmlc.org Internet Source	<1%
	www.fens.org	


58
Internet Source
<1%
59
Mengdi Li, Eugene Ch'ng, Alain Yee Loong Chong, Simon See. "Multi-class Twitter sentiment classification with emojis", Industrial Management & Data Systems, 2018
Publication
<1%
60
bismarcktribune.com
Internet Source
<1%
61
islamicmarkets.com
Internet Source
<1%
62
ebiquity.umbc.edu
Internet Source
<1%