# Malaysian fibre internet service provider: a naïve Bayes classification Twitter sentiment analysis

**Khyrina Airin Fariza Abu Samah[1], Muhamad Nabil Fahruddin[1], Raseeda Hamzah[1], Lala Septem Riza[2], Khairul Nurmazianna Ismail[1], Rosniza Roslan[1], Raihah Aminuddin[1], Nor Intan Shafini Nasaruddin[1]**
[1]College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA Melaka Branch, Jasin Campus, Melaka, Malaysia
[2]Department of Computer Science Education, Universitas Pendidikan Indonesia, Bandung, Indonesia

**Article Info**

***Article history:***

Received Jan 18, 2024
Revised Aug 9, 2024
Accepted Aug 30, 2024

***Keywords:***

Classification
English and Malay
Internet service provider
Naïve Bayes
Sentiment analysis
Twitter

## ABSTRACT

In the highly competitive landscape of Malaysian internet service providers (ISPs), users seek efficient ways to assess service quality. While various websites allow visual comparisons of fiber ISPs, a direct side-by-side evaluation remains elusive. A survey of 101 respondents revealed that 92.1% found researching a company's reputation time-consuming. Additionally, relying on English-centric online ratings may lead to skewed outcomes, disregarding reviews in diverse languages. In response, we developed a web-based dashboard utilizing Twitter sentiment analysis (SA) and the naïve Bayes (NB) algorithm to classify Malaysia's best fiber ISPs. The SA focused on four key factors: package price, internet speed, coverage area, and customer service, simplifying the comparison process. The system's usability and functionality tests showed that both the English and Malay models could classify scraped Twitter data with an accuracy of 80%. The system's remarkable usability score of 94.58% on the system usability scale (SUS) confirms its acceptability and excellent performance in achieving research goals.

***Corresponding Author:***

Khyrina Airin Fariza Abu Samah
College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA Melaka Branch
Jasin Campus, Merlimau, 77300, Melaka, Malaysia
Email: khyrina783@uitm.edu.my

## 1. INTRODUCTION

An internet service provider (ISP) is a company that offers basic internet connectivity to both homes and businesses [1]. They use various methods, such as fiber optics, copper cables, and satellites, to provide internet access to their customers [2]. In Malaysia, ISPs need a license from the Malaysian Communications and Multimedia Commission (MCMC) to operate. Home fibre, for example, uses fiber optic cables and networking devices to transfer data at much higher speeds than traditional copper wires [3], with speeds reaching up to 1 Gigabit per second, thanks to its advanced technology.

In Malaysia, when discussing fixed-fibre broadband infrastructure providers, names like Unifi (TM), TIME, and Allo City Broadband (TNB) are commonly mentioned. TIME's fibre network is tailored for high-density and high-rise buildings. TM, on the other hand, rents out its fibre-optic capacity through wholesale agreements with telecom firms such as Maxis Bhd, Digi.com Bhd, and Celcom Axiata Bhd at regulated rates [4]. Allo is extending its broadband Internet services to other regions of the country after completing the national fiberization and connectivity plan (NFCP) pilot in Jasin, Melaka, Malaysia. The government's *jalinan digital negara* (JENDELA) plan encourages broadband infrastructure providers with an open-access concept, allowing other companies to become retail service providers. Companies like Astro, Digi, and Maxis have announced plans to expand their fibre networks accordingly.

To address the nation's goal of improving broadband quality, expanding coverage, and lowering costs, the government introduced the NFCP [5]. This initiative extends to the fiber internet industry, where users face challenges in selecting the right ISP due to competition. One significant challenge for customers is the time-consuming process of manually comparing service information when deciding on a subscription. In a survey, 92.1% of 101 respondents currently subscribed to a Malaysian fiber ISP agreed that this comparison process is tedious. Several websites, like *imoney.my* and *ringgitplus.com*, allow users to compare fiber ISPs. However, these sites lack a direct comparison feature due to scattered data across multiple sources. The quality of information is crucial for customer satisfaction [6], [7]. In addition, much of the data on these sites is outdated and unreliable. This tedious task is potentially impacting their interest in subscribing to any service. User satisfaction with internet service influences their likelihood of continued use and willingness to repurchase it [8].

Many online ratings stem from an English-language platform, potentially yielding inaccurate results as reviews in other languages are not always considered [9]. Konno *et al.* [10] highlighted that excluding non-English studies from systematic reviews can introduce bias, as articles in languages other than English are often overlooked. Therefore, to ensure accuracy, evaluations of online ratings should encompass other languages, especially Malay.

The internet has transformed how people share their thoughts, using platforms like blogs, forums, and product reviews [11]. Microblogging sites, in particular, have become rich sources of information [12], [13]. Rathore and Ilavarasan [14] highlight that Twitter, a major social media platform, is commonly used to express feelings and opinions about brands, products, or services. With over 190 million tweets daily, users discuss various topics openly [15]. Due to social media's open nature, Twitter is also utilized by many establishments and organizations for information [16], [17]. Hence, if products or services fall short of expectations, customers are likely to express dissatisfaction on social networks, providing genuine feedback to the public [18].

This project involves creating a web-based dashboard that analyzes Twitter sentiment to classify and visualize the performance of the leading fiber ISPs in Malaysia. Web-based data visualization offers numerous benefits, simplifying complex data for better understanding, and quicker decision-making. Visualizations help identify trends and anomalies, providing valuable insights for planning and forecasting. Clear communication of data makes it effective for reaching diverse audiences and supporting collaborative decision-making. Interactive features enable dynamic exploration, while visual dashboards allow real-time monitoring of key indicators [19], [20]. The model focuses on popular opinions about Malaysian fiber ISPs regarding package price, internet speed, coverage, and customer service, using English and Malay datasets for sentiment analysis (SA).

We use naïve Bayes (NB) classifier that is favoured in supervised machine learning due to its simplicity and efficiency, making it suitable for use with limited computing resources. It performs well with high-dimensional data, such as text classification tasks, and offers fast training speeds [21]. The algorithm's assumption of feature independence contributes to its effectiveness and provides probabilistic outputs, aiding in understanding prediction confidence [22]. The paper is outlined as follows: section 1 introduces the topic, while section 2 details the research procedures. Section 3 examines the findings, focusing on precision, practicality, and applicability. Finally, section 4 briefly discusses potential future improvements.

## 2. RESEARCH METHOD

We started the research method with the web-based dashboard design, creating the diagram for use cases, and designing the system interface. These steps are crucial to achieving the study objectives and ensuring they are attainable. The back-end development, depicted in Figure 1, includes data collection, data pre-processing, the NB classifier, evaluation performance metrics, and model deployment for the web-based dashboard. The classification process is further divided into three sub-tasks: factor analysis, visualization, and reporting.

### 2.1. Data collection

Machine learning techniques are applied to develop text classification models for both Malay and English languages. The English model utilizes a dataset consisting of 800,000 positive and 800,000 negative entries obtained from [23]. Conversely, the Malay model uses a dataset sourced from the GitHub website, containing 344,733 negative entries and 312,985 positive entries. Both datasets, as depicted in Table 1, consist solely of positive and negative binary classification data. An extra dataset of 13,900 records was acquired from the Kaggle website to provide unbiased training and testing data for both models. The impartial data collected for the English model is translated into Malay using Google Translate to give unbiased data for the Malay model. Consequently, there are 1,613,900 training and testing data for the English model and 732,764 for the Malay model.

The majority of Malaysian Twitter users express their opinions in Malay. Real-world data from Twitter is used to research three fiber ISPs: Unifi, TIME, and Allo. Tweets containing the hashtags "#Unifi," "#TIMEInternet," and "#AlloWifi" are collected between January 1, 2022, and December 31, 2022. These hashtags represent the respective fiber ISPs. Using Twint in Jupyter Notebook, tweets are collected without case sensitivity. The collected tweets are stored as comma-separated values (CSV) files. There were 1,354 records for Unifi, 455 for Allo, and 325 for TIME.

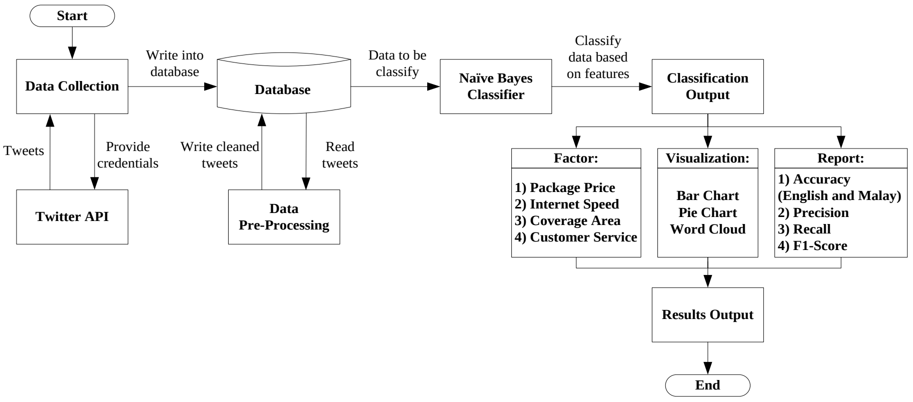


Figure 1. Research design for back-end development

Table 1. Summarization of the English and Malay model datasets

| Model | Data description | | |
|---|---|---|---|
| English | Source | https://www.kaggle.com/kazanova/sentiment140 [23] | |
| | Number of data | Positive | 800,000 |
| | | Negative | 800,000 |
| | | Neutral | 13,900 |
| | | Total | 1,613,900 |
| Malay | Source | https://github.com/mesolitica/malaysian-dataset [24] | |
| | Number of data | Positive | 312,985 |
| | | Negative | 344,733 |
| | | Neutral | - |
| | | Total | 657,718 |

### 2.2. Data pre-processing

Text pre-processing is a crucial step in refining and organizing data by removing irrelevant elements that do not contribute to its meaning, thereby enhancing the quality of the derived model. This process is facilitated using Python libraries natural language toolkit (NLTK) and RE. The final dataset is streamlined to include only four columns (date, username, tweet, and language) and discarded columns with redundant data. All characters are changed to lowercase to prevent problems associated with case sensitivity. Undesirable elements such as emojis, punctuation, and excess whitespace are then eliminated. We exclude the terms like hashtags, links, and mentions. Furthermore, null values and duplicate tweets are removed from the dataset to simplify the data further.

This dataset, however, remains high-dimensional, and in order to reduce the dimension, the stop words that do not contribute anything of value are eliminated. Stop words such as "the," "and," "of," and "on" are examples of stop words that are used in the English language. The NLTK library provides access to English stop words using a pre-built function. The Malay model excludes stop words in Malay by importing them from [24].

### 2.2.1. Naïve Bayes classification

The NB classifier model is employed in this study. It begins with data collection to gain information about fibre internet services. Twint, a Twitter scraping programme for Python, is used to extract tweets in both Malay and English relating to the three fibre ISPs, Unifi, TIME, and Allo (TNB). The data will then be pre-processed and classified into positive, neutral, and negative categories. Its goal is to eliminate any extraneous data that might impact the result. The model employs the NB classifier to assess the dataset, with the algorithm categorizing the dataset. The model classifies the dataset by applying the knowledge acquired from the labeled data in the training set. The NB theorem functions by determining the likelihood of an event through the probabilistic joint distribution of preceding events [25]. We utilize the pre-labeled training dataset to train the model to differentiate between positive, neutral, and negative utterances in this study.

The structured presentation of phrases and words, known as text representation, actively counts the occurrences of the bag of words (BOW) phrase. It transforms the retrieved word tokens into vectors, enabling the ML model to acquire this feature. Utilizing the three steps of BOW, the term frequency (TF), inverse document frequency (IDF), and unit length of the vectors are determined. The first two steps, collectively referred to as term frequency-inverse document frequency (TF-IDF), serve as a statistical measure to assess the significance of a word in a document. TF-IDF weight is commonly employed in information retrieval and text mining as a weighting factor. TF gauges the frequency of a term within a document, while IDF assesses a term’s relevance. In (1) and (2) encompass the formulas for computing TF and IDF, respectively.

$$TF(t) = \frac{(Number\ of\ times\ term\ appears\ in\ document)}{(Total\ number\ of\ terms\ in\ the\ document)} \tag{1}$$

$$IDF(t) = log_e \frac{(Total\ number\ of\ documents)}{(Number\ of\ documents\ with\ term\ t\ in\ it)} \tag{2}$$

### 2.2.2. Performance metrics evaluation

The subsequent phase involves deploying the model in a real-world scenario to evaluate its performance. This assessment utilizes the test holdout dataset, generating performance metrics such as a classification report and a confusion matrix [15], [25]. It provides results that encompass observed metrics, including accuracy, the confusion matrix, and the classification report. Ultimately, after the model processes the acquired data through the Jupyter Notebook for sentiment predictions, its performance undergoes evaluation before proceeding to the data visualization phase.

### 2.2.3. Web-based dashboard model deployment

Model deployment involves rendering an ML model that is accessible for use. During this stage, the model generates classified tweets with sentiment labels of ‘0’, ‘2’, and ‘4’, denoting negative, neutral, and positive feelings, respectively. Following sentiment predictions on the acquired data using the model classifier and performance evaluation, the data is visualized through Plotly, an open-source interactive graphics toolkit for Python. The process is initiated by importing the acquired data into Pandas data frames in Python. The outcome includes generating charts with details presented through coding. The analysis results are then leveraged to construct an interactive visualization tool illustrating real-world data analysis outcomes. The text data for the proposed fibre ISP system is depicted through a bar chart, pie chart, and word cloud. The words are displayed in varied colors, with word size indicating the frequency of appearance in the text data. This cloudy presentation aims to facilitate the easy identification of terms associated with fibre ISP. Following is a discussion of the system’s front-end development.

### 2.3. Development of front-end

Translating data into a graphical interface for user interaction is achieved through front-end web development or client-side development. The essential building blocks for creating websites, namely HTML, CSS, and JavaScript, facilitate this process. Within the Python web application environment, sentiment data charts and graphs can be effortlessly visualized. The development comprises ten modules: the overview page, Unifi page, TIME page, Allo page, package price page, internet speed page, coverage area page, customer service page, real-time tweet page, and competitive analysis page.

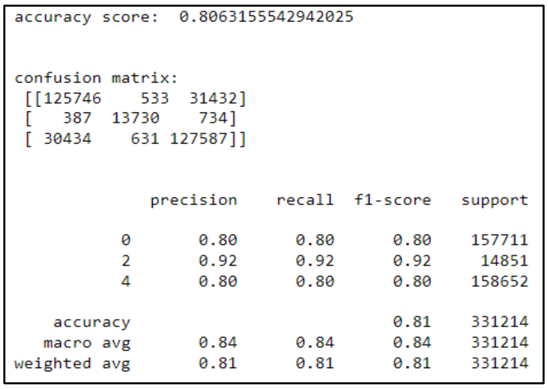
## 3. RESULTS AND DISCUSSION

The main issue in this study is that due to competition among ISPs in Malaysia, users need to get reviews and feedback on the quality of their services. However, customers faced challenges and time-consuming in selecting and evaluating each fibre ISP when making a subscription decision. Also, most online ratings are derived from an English-language online platform. Thus, this project entails creating a web-based dashboard that classifies and visualizes the top fibre ISPs in Malaysia performance of by analyzing sentiment from Twitter. As a result, this section is divided into four sections: accuracy testing, dashboard visualization, functionality testing, and usability testing.

### 3.1. Accuracy testing

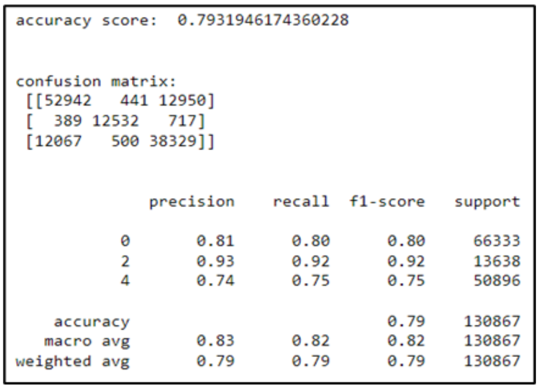
The classification was done by modelling the NB classifier model using a simple Python program. In order to ensure the model avoids overfitting, the cross-validation process utilizes the training data [26]. Various hyperparameter configurations are evaluated to partition the model randomly. Eight parameter configurations are tested with 10 K fold validations, resulting in the model being trained and assessed a total of eighty times. The dataset is bifurcated into training and testing sets, comprising 1,340,036 tweets and 335,009 tweets, respectively.

Figure 2 displays the testing accuracy for the English model, yielding a score of 80%. This accuracy score translates to 80%, indicating that the sentiment results are 80% accurate. In practical terms, this means that out of 10 trials, the model correctly identified eight answers as either “positive,” “neutral,” or “negative”. The numerical representations for these classes are 4 for “positive,” 2 for “neutral,” and 0 for “negative”. Figure 3 illustrates the accuracy score for the Malay model. Similar to the English model, the confusion matrix displays an 80% accuracy score, signifying that the sentiment results are 80% accurate. The result implies that the model accurately classified eight out of ten results as either “positive,” “neutral,” or “negative,” using the numerical representations 4, 2, and 0, respectively.



```
accuracy score:  0.8063155542942025


confusion matrix:
 [[125746    533  31432]
 [   387  13730    734]
 [ 30434    631 127587]]


              precision    recall  f1-score   support

           0       0.80      0.80      0.80    157711
           2       0.92      0.92      0.92     14851
           4       0.80      0.80      0.80    158652

    accuracy                           0.81    331214
   macro avg       0.84      0.84      0.84    331214
weighted avg       0.81      0.81      0.81    331214
```

Figure 2. English model accuracy test result



```
accuracy score:  0.7931946174360228


confusion matrix:
 [[52942   441 12950]
 [  389 12532   717]
 [12067   500 38329]]


              precision    recall  f1-score   support

           0       0.81      0.80      0.80     66333
           2       0.93      0.92      0.92     13638
           4       0.74      0.75      0.75     50896

    accuracy                           0.79    130867
   macro avg       0.83      0.82      0.82    130867
weighted avg       0.79      0.79      0.79    130867
```

Figure 3. Malay model accuracy test result

### 3.2. Dashboard visualization

To overcome the difficulties of making the comparison, displaying and illustrating real-world data analysis occurs on the system’s dashboard. Diverse visuals were used, including bar charts, pie charts, gauge charts, and word clouds. Based on data collected between January 1, 2022, and December 31, 2022, it has been determined that the market share of the three fibre ISPs is comparable since there is no significant difference between their market share scores. However, it is noticeable that TIME has the lowest market share compared to Unifi and Allo, with only 30% of their market share, as in Table 2.

Table 2. Comparison results for Unifi, Allo, and TIME in 2022

| Fibre ISP | Total mentions | Market share (%) |
|---|---|---|
| Unifi | 584 | 35 |
| Allo | 118 | 35 |
| TIME | 19 | 30 |

#### 3.2.1. Overall sentiment analysis

Figure 4 on the top dashboard actively presents the market situation for each fiber ISP in 2022 based on Twitter mentions. Unifi and TIME receive more negative mentions than neutral or positive. In contrast, Allo garners more positive mentions than neutral and negative, as illustrated in Figure 5 a bar chart portraying the overall sentiment of fiber ISPs in 2022, measured through Twitter mentions. Each dashboard is dedicated to fiber ISPs and related to the total number of sentiment breakdowns.

#### 3.2.2. Sentiment analysis visualization

Figure 6 shows the comprehensive word cloud for SA of the fibre ISP. Figure 6(a) visually displays positive sentiment text data using a green-colored word cloud. Figure 6(b) represents negative thoughts using a red color. Figure 6(c) depicts neutral sentiments with a grey color. Word size corresponds to its frequency in the dataset, with greater sizes representing more occurrences. ‘laju’, ‘dekat’, and ‘betul’ depicted in the world cloud are the linked positive terms. Red symbolizes words linked to a sour emotion. The word cloud contains the terms ‘trouble’, ‘slow’, and ‘line’. The grey word cloud representing neutral sentiment shows the terms linked to neutral references.

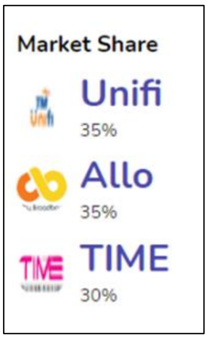


Figure 4. Market share for the fibre ISP

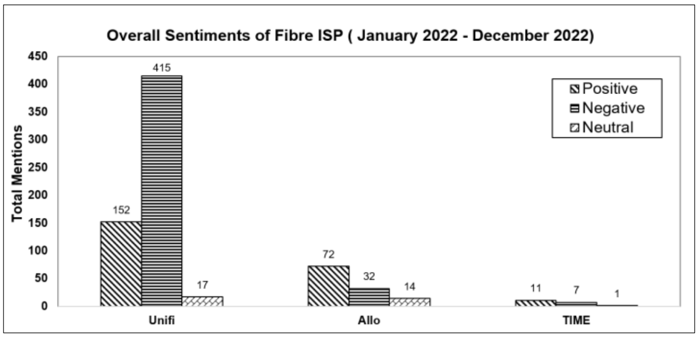
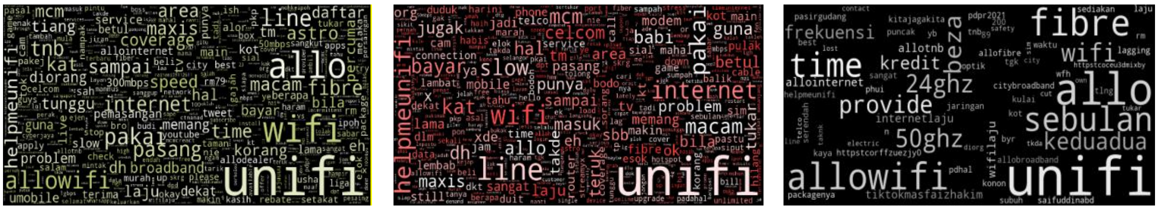


Figure 5. Bar graph: overall sentiments for the fibre ISP



(a)
(b)
(c)

Figure 6. Classification of overall sentiments for the fibre ISP for; (a) positive tweets, (b) negative tweets, and (c) neutral tweets

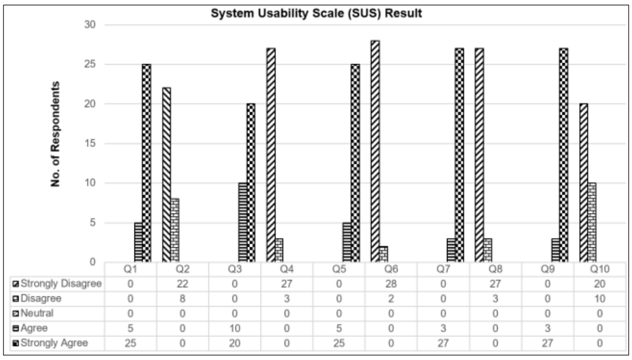
### 3.3. Functionality testing

Functionality testing is a type of software testing that evaluates the system based on its functional requirements. Users rigorously verify the system’s capabilities by executing test cases, inputting data, and examining the results. This procedure evaluates the effectiveness of the functionalities to ensure that the system’s functionality meets the expected standards. As a result, the functionality testing demonstrates that the system is running correctly and successfully.

### 3.4. Usability testing

A bar chart, illustrated in Figure 7, displays the scores for the ten system usability scale (SUS) statements, reflecting the scale of user rankings. The graph reveals that for odd-numbered items, which constitute positive comments, most users select scale 5, representing “strongly agree.” Conversely, even-numbered questions predominantly receive a score of 1 out of 5, indicating users “strongly disagree” with the negative claims. This suggests that users are well-acquainted with the system and find it unnecessary to seek technical support for its functions. Overall, users express satisfaction with the system. Figure 8 depicts the SUS scores using a histogram, where the frequency of SUS response represented by y-axis, and the x-axis signifies the percentage of the SUS score range.

As observed from the histogram, the data is distributed within the range of 90% to 97.5%. The graph exhibits a normal distribution, spanning from 90% to 97.5%, with a 1.9% interval. The peak of the histogram lies between 93.8% and 95.6%, comprising a total of 14 responders. Among these responders, nine fall below the central value, while seven fall above it. The 30 respondents who completed the SUS questionnaire yielded an overall SUS score of 94.58%. The baseline for the SUS average score is set at 70%, classifying scores below this threshold as below average, necessitating attention to design faults and additional studies. Scores exceeding 70% are considered above average and acceptable, while those surpassing 80% are deemed excellent [27]. Achieving a score of 94.58% indicates that the research is both acceptable and excellent.



| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 |
|---|---|---|---|---|---|---|---|---|---|---|
| ▨ Strongly Disagree | 0 | 22 | 0 | 27 | 0 | 28 | 0 | 27 | 0 | 20 |
| ▨ Disagree | 0 | 8 | 0 | 3 | 0 | 2 | 0 | 3 | 0 | 10 |
| ▨ Neutral | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ▤ Agree | 5 | 0 | 10 | 0 | 5 | 0 | 3 | 0 | 3 | 0 |
| ▩ Strongly Agree | 25 | 0 | 20 | 0 | 25 | 0 | 27 | 0 | 27 | 0 |

Figure 7. A bar chart for SUS result

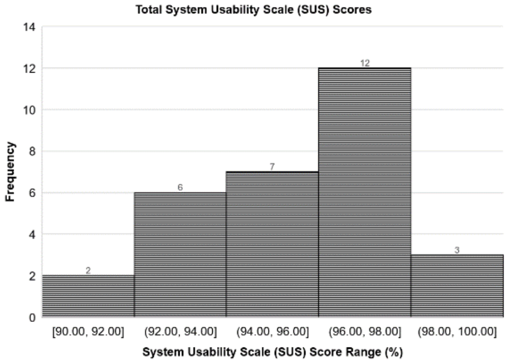


Figure 8. A histogram for SUS result

## 4. CONCLUSION

Creating a web-based dashboard was the initiative to present the SA portraying Twitter users' perceptions of Unifi, Allo, and TIME, spanning from January 1, 2022, to December 31, 2022. The NB classifier model, an integral part of the system application, empowers users to apply SA to any textual data. Three classifications—positive, neutral, and negative—facilitate insights into fiber ISPs' performance, aiding users in decision-making. The system's diverse representations simplify obtaining a profound understanding of each fiber ISP. During functionality testing, the system's functionality is rigorously examined and validated against the system requirements, achieving an accuracy of 80% for both the English and Malay models. Simultaneously, usability testing assesses the system workflow. The application undergoes comprehensive evaluation, performs as anticipated, and attains an acceptability score of 94.5% according to the SUS. For future research endeavors, it is recommended to define slang, abbreviations, and sarcastic phrases in the dictionary to transform them into relevant values, enhancing the determination of tone.

## ACKNOWLEDGEMENTS

This research was funded by a grant from Universiti Teknologi MARA Cawangan Melaka (TEJA Grant 2024 GDT 2024/1-15).

## REFERENCES

[1] T. Myllykangas, "Analysis of the residential smart building market from an internet service provider's perspective: A district heating business case," Aalto University, Master, 2021.

[2] M. Graydon and L. Parks, "Connecting the unconnected': a critical assessment of US satellite Internet services," *Media, Cult. Soc.*, vol. 42, no. 2, pp. 260–276, 2020, doi: 10.1177/016344371986183.

[3] H. Kaur, S. Singh, R. Kaur, and R. Singh, "Advances in fronthauling of communication technologies: A review," *J. Netw. Comput. Appl.*, vol. 223, pp. 1–40, 2024, doi: 10.1016/j.jnca.2023.103806.

[4] G. Kana, "TFP solutions' fibre broadband play," 2021. [Online]. Available: https://www.thestar.com.my/business/business-news/2021/09/04/tfp-solutions-fibre-broadband-play (accessed Nov. 28, 2022).

[5] R. Gong, "Digital inclusion: Assessing connectivity in Malaysia," in *Navigating Challenges, Realising Opportunities of Digital Transformation*, Malaysia, 2021, pp. 5–38.

[6] G. Kankam, E. Kyeremeh, G. N. K. Som, and I. T. Charnor, "Information quality and supply chain performance: The mediating role of information sharing," *Supply Chain Anal.*, vol. 2, pp. 1–8, 2023, doi: 10.1016/j.sca.2023.100005.

[7] B. Kim, M. Yoo, and W. Yang, "Online engagement among restaurant customers: The importance of enhancing flow for social media users," *J. Hosp. Tour. Res.*, vol. 44, no. 2, pp. 252–277, 2020, doi: 10.1177/1096348019887202.

[8] S.-M. Tseng, "Understanding the impact of the relationship quality on customer loyalty: The moderating effect of online service recovery," *Int. J. Qual. Serv. Sci.*, vol. 13, no. 2, pp. 300–320, 2021, doi: 10.1108/IJQSS-07-2020-0115.

[9] K. A. F. A. Samah, N. F. A. Misdan, M. N. H. H. Jono, and L. S. Riza, "The best Malaysian airline companies visualization through bilingual twitter sentiment analysis: a machine learning classification," *JOIV Int. J. Informatics Vis.*, vol. 6, no. 1, pp. 130–137, 2022, doi: 10.30630/joiv.6.1.879.

[10] K. Konno *et al.*, "Ignoring non-English-language studies may bias ecological meta-analyses," *Ecol. Evol.*, vol. 10, no. 13, pp. 6373–6384, 2020, doi: 10.1002/ece3.6368.

[11] M. S. Hossain, M. F. Rahman, and X. Zhou, "Impact of customers' interpersonal interactions in social commerce on customer relationship management performance," *J. Contemp. Mark. Sci.*, vol. 4, no. 1, pp. 161–181, 2021, doi: 10.1108/jcmars-12-2020-0050.

[12] S. Saranya and G. Usha, "A machine learning-based technique with intelligent wordnet lemmatize for twitter sentiment analysis," *Intell. Autom. Soft Comput.*, vol. 36, no. 1, pp. 339–352, 2023, doi: 10.32604/iasc.2023.031987.

[13] M. S. Hossain and M. F. Rahman, "Customer sentiment analysis and prediction of insurance products' reviews using machine learning approaches," *FIIB Bus. Rev.*, vol. 12, no. 4, pp. 386–402, 2023, doi: 10.1177/23197145221115793.

[14] A. K. Rathore and P. V. Ilavarasan, "Pre- and post-launch emotions in new product development: Insights from twitter analytics

of three products,” *Int. J. Inf. Manage.*, vol. 50, pp. 111–127, 2020, doi: 10.1016/j.ijinfomgt.2019.05.015.

[15] K. A. F. A. Samah, N. M. N. Azharludin, L. S. Rizan, M. N. H. H. Jono, and N. A. Moketar, “Classification and visualization: Twitter sentiment analysis of Malaysia’s private hospitals,” *Int. J. Artif. Intell.*, vol. 12, no. 4, pp. 1793–1802, 2023, doi: 10.11591/ijai.v12.i4.pp1793-1802.

[16] P. Gómez-Carraso, E. Guillamón-Saorí, and B. G. Osma, “Stakeholders versus firm communication in social media: The case of Twitter and corporate social responsibility information,” *Eur. Account. Rev.*, vol. 30, no. 1, pp. 31–62, 2021, doi: 10.1080/09638180.2019.1708428.

[17] E. Chandrasekaran, R., Mehta, V., Valkunde, T., and Moustakas, “Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study,” *J. Med. Internet Res.*, vol. 22, no. 10, pp. 1–37, 2020, doi: 10.2196/22624.

[18] S. McCarthy, W. Rowan, C. Mahony, and A. Vergne, “The dark side of digitalization and social media platform governance: a citizen engagement study,” *Internet Res.*, vol. 33, no. 6, pp. 2172–2204, 2023, doi: 10.1108/INTR-03-2022-0142.

[19] X. Qin, Y. Luo, N. Tang, and G. Li, “Making data visualization more efficient and effective: a survey,” *VLDB J.*, vol. 29, pp. 93–117, 2020, doi: 10.1007/s00778-019-00588-3.

[20] A. Thakkar and K. Chaudhari, “Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks,” *Appl. Soft Comput. J.*, vol. 96, pp. 1–13, 2020.

[21] U. Rahman and M. U. Mahbub, “Application of classification models on maintenance records through text mining approach in industrial environment,” *J. Qual. Maint. Eng.*, vol. 29, no. 1, pp. 203–219, 2023, doi: 10.1108/JQME-08-2021-0064.

[22] P. C. Sen, M. Hajra, and M. Ghosh, “Supervised classification algorithms in machine learning: A survey and review.,” in *Emerging Technology in Modelling and Graphics (IEM Graph)*, 2020, pp. 99–111, doi: 10.1007/978-981-13-7403-6_11.

[23] Kaggle, “Sentiment140 dataset with 1.6 million tweets,” 2022. [Online]. Available: https://www.kaggle.com/kazanova/sentiment140 (accessed Jan. 05, 2023).

[24] Github, “Malaysian-dataset,” 2023. [Online]. Available: https://github.com/mesolitica/malaysian-dataset (accessed Jan. 15, 2023).

[25] A. Agarwal, P. Sharma, M. Alshehri, A. A. Mohamed, and O. Alfarraj, “Classification model for accuracy and intrusion detection using machine learning approach,” *PeerJ Comput. Sci.*, vol. 7, 2021, doi: 10.7717/peerj-cs.437.

[26] A. M. López O., A. M. López, and J. Crossa, “Overfitting, model tuning, and evaluation of prediction performance,” in *Multi. stat. mach. learn. metho. for genomic pred.*, Cham: Springer International Publishing, 2022, pp. 109–139.

[27] A. A. Salih and A. M. Abdulazeez, “Evaluation of classification algorithms for intrusion detection system: A review,” *J. Soft Comput. Data Min.*, vol. 02, no. 01, pp. 31–40, 2021.

## BIOGRAPHIES OF AUTHORS

**Khyrina Airin Fariza Abu Samah** is a senior lecturer at the College of Computing, Informatics and Mathematics in Universiti Teknologi MARA (UiTM) Melaka Branch, Jasin Campus. Before joining UiTM, she had 13 years of working experience in the semiconductor industry. She has a Diploma, Bachelor’s degree and Master’s degree in Computer Science and Ph.D. in Information Technology. Her research interests are in artificial intelligence, algorithm analysis, machine learning optimization, and evacuation algorithms. She can be contacted at email: khyrina783@uitm.edu.my.

**Muhamad Nabil Fahruddin** is a recent graduate with a Bachelor’s Degree in Computer Science from Universiti Teknologi MARA (UiTM). His research interests include machine learning algorithms, data analysis, and visualization. Previously, he became a research assistant at the College of Computing, Informatics and Mathematics in Universiti Teknologi MARA (UiTM) Melaka Branch, Jasin Campus. Currently, he is working as an RPA Operation Specialist at RHB Banking Group. He can be contacted at email: nabilfahruddin98@gmail.com.

**Raseeda Hamzah** is currently working as a senior lecturer at the College of Computing, Informatics and Mathematics in Universiti Teknologi MARA (UiTM) Melaka Branch, Jasin Campus. She received her Bachelor and Master degrees from Universiti Teknikal Malaysia Melaka and the University of Malaya. Her Ph.D. in (Computational Linguistics) from UiTM. She is actively researching signal processing, pattern recognition and feature analysis for various areas and different types of data. She can be contacted at email: raseeda@uitm.edu.my.

**Lala Septem Riza** received Ph.D. in Computer Science from Universidad de Granada, Spain, in 2015. He works in the Department of Computer Science Education, Universitas Pendidikan Indonesia, Indonesia. He teaches machine learning, big data platforms, and statistical data science. His research interests are in machine learning, data science, and education. He can be contacted at email: lala.s.riza@upi.edu.

**Khairul Nurmazianna Ismail** was born in 1984 and holds a Master of Science (M.Sc) degree in Information Technology and Quantitative Science from Universiti Teknologi MARA (UiTM), Malaysia. She also holds a Bachelor of Science (B.Sc.) with Honours in Computer Science and a Diploma in Computer Science, both from UiTM. With 11 years of working experience, including time spent in thte industry, she currently serves as a lecturer in the College of Computing, Informatics and Mathematics, UiTM Melaka Branch, Jasin Campus. Her primary research interests lie in the field of machine learning. She can be contacted at email: khairul_nur@uitm.edu.my.

**Rosniza Roslan** is a senior lecturer at the Computing Sciences Studies, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM) Melaka Branch, Malaysia. She received her Bachelor’s Degree of Science in Computational Mathematics and Master’s Degree in Information Technology and Quantitative Sciences from Universiti Teknologi MARA (UiTM), Malaysia. Her research interests include image processing, medical imaging, pattern recognition, artificial intelligence, machine learning, and the internet of things. She can be contacted at email: rosniza@tmsk.uitm.edu.my.

**Raihah Aminuddin** is a senior lecturer at the College of Computing, Informatics and Mathematics, MARA University of Technology, Malacca Branch (Jasin Campus), Malaysia. She received her Diploma, Bachelor’s Degree, and Master’s in Computer Science from Universiti Teknologi MARA, Malaysia. She earned her Ph.D. in Computer Science from the University of Sheffield, United Kingdom, in 2019. She is actively researching image processing with deep learning and big data. She can be contacted at email: raihah1@uitm.edu.my.

**Nor Intan Shafini Nasaruddin** is a senior lecturer at the College of Computing, Informatics and Mathematics in Universiti Teknologi MARA (UiTM) Cawangan Melaka Branch, Jasin Campus. She completed her M.Sc. in Computer Science and B.Sc. in Information Technology. Her research interests are educational technology, the internet of things, and machine learning. She can be contacted at email: norintan4463@uitm.edu.my.