SEGMENTATION OF USER FEEDBACK AND RATINGS

1211301744 LEE TIAN XIN

BACHELOR OF COMPUTER SCIENCE (DATA SCIENCE)

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY

JUNE 2024

SEGMENTATION OF USER FEEDBACK AND RATINGS

BY

1211301744 LEE TIAN XIN

PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENT FOR THE DEGREE OF
COMPUTER SCIENCE (DATA SCIENCE)
in the
Faculty of Computing and Informatics

MULTIMEDIA UNIVERSITY
MALAYSIA

JUNE 2024

Declaration

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

Lee Tian Xin

Name of candidate: Lee Tian Xin

Faculty of Computing & Informatics

Multimedia University

Date: 27: 06: 2024

Abstract

The ubiquity of mobile applications in the digital age has underscored the critical role of user feedback in guiding app development and enhancing user satisfaction. This paper presents a comprehensive approach to analyzing user feedback through a combination of unsupervised and supervised learning techniques, with a focus on the Selangkah app—a multifunctional platform offering diverse services including health, welfare, and user management. Initially, unsupervised learning techniques, such as BERTopic, were employed for topic modeling to categorize user feedback. However, due to limitations in providing clear labels or insights into specific topics, the methodology transitioned to a supervised learning approach. By manually labeling a subset of the data and training classifiers on this labeled dataset, the study aims to efficiently categorize user feedback to identify areas for improvement and understand user sentiment across different functionalities. The results demonstrate the efficacy of this hybrid approach in extracting actionable insights from user reviews, thereby guiding targeted enhancements to the Selangkah app.

Table of Contents

List of Tables

List of Figures

List of Abbreviations/Symbols

LLMs	Large Language Models
LDA	Latent Dirichlet Allocation
BERT	Bidirectional Encoder Representations from Transformers
DMM	Dirichlet Multinomial Mixture
Llama2	Large Language Model Meta AI 2
NLP	Natural Language Processing
GRU	Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
SVM	Support Vector Machines
TFIDF	Term Frequency-Inverse Document Frequency
OCTIS	Optimizing and Comparing Topic Models is Simple
OOV	Out-of-Vocabulary
EDA	Exploratory Data Analysis

Chapter 1: Introduction

In the rapidly evolving landscape of digital technology, mobile applications have become integral to user engagement and satisfaction. The importance of timely and effective analysis of user feedback is highlighted by instances where apps faced significant backlash due to unaddressed issues or undesirable updates, as seen in cases like Pokémon GO and Gray Raven Punishing. (Gao et al., 2019; Wang et al., 2022). Such situations can be alleviated with timely problem-solving. Therefore, this project aims to rapidly understand user feedback to enhance the overall user experience.

1.1 Problem Statement

Despite the wealth of data available through user reviews and ratings, extracting meaningful insights remains a challenge due to the volume, diversity, and brevity of the feedback. Short text reviews often contain noise, sparsity, and lack context, making it difficult to parse and interpret manually. Traditional manual analysis methods are not only labor-intensive but also prone to inaccuracies. This study seeks to address these challenges by applying a combination of unsupervised and supervised learning techniques to automate the analysis of user feedback on the Selangkah app, focusing on its distinct functionalities: Account, Health, Finance, Sign up, and Sign in.

1.2 Research Question

1. What themes emerge from the preliminary analysis of user feedback on the Selangkah app?
2. How can unsupervised and supervised learning techniques be effectively utilized to extract key phrases and words related to the app's functionalities?

1.3 Objectives

Two objectives are to be achieved in this research project which are following:

1. To perform exploratory data analysis to uncover themes within the Selangkah user feedback dataset.
2. To identify key phrases and words associated with the app's specific functionalities using a combination of unsupervised and supervised learning techniques, enhancing the understanding of user sentiment and feedback.

1.4 Scope

The scope of this project is discovering themes in the 'Selangkah' dataset, which only focuses on non-null data, comprising around 2000 rows. It focuses on Malay, Chinese, and English user reviews.

1.5 Project Timeline

The project timeline is outlined in Figure 1.5-1, illustrating the tasks and their respective durations over two trimesters for the Final Year Project (FYP). The timeline is divided into two main sections corresponding to Trimester 1 and Trimester 2.

		Trimester 1 23/24 FYP 1														
	Week / Task	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No.																
1	Initial Research and Literature Review															
2	Exploratory Data Analysis (EDA)															
3	Data Preprocessing															
4	Development of Proposed Framework															
5	Experimentation with LDA															
6	Experimentation with BERTopic															
7	Write Report															
8	Preparation for Presentation															
9	Presentation															
		Trimester 2 23/24 FYP 2														
1	Parameter Tuning (BERTopic)															
2	Labeling															
3	Data Preparation for Classification															
4	Model Training (Traditional)															
5	Model Training (GRU/ BiLSTM)															
6	Streamlit (Interface)															
7	Write Report															
8	Preparation for Presentation															
9	Presentation															

Figure 1.5-1 Project Timeline

1.6 Chapter Organization

- Chapter 1: Introduction - Introduces the project, outlining the problem statement, research questions, objectives, scope, and project timeline. It also includes the expected findings and an overview of the chapter organization.
- Chapter 2: Literature Review - Reviews relevant literature on text classification, topic modeling, and the use of these techniques in analyzing app reviews. Establishes the theoretical foundation for the proposed framework.
- Chapter 3: Proposed Framework - Presents the framework used in the project, detailing the dataset, preprocessing methods, topic modeling techniques, and classification approaches.
- Chapter 4: Result and Discussion - Discusses the results of the analysis, including EDA results, preprocessing outcomes, topic modeling results, and classification performance. Compares different models and discusses their implications.
- Chapter 5: Conclusion - Summarizes the findings of the project, discusses the limitations, and provides suggestions for future work.

Chapter 2: Literature Review

2.1 Introduction

In the digital age, the explosion of user-generated data, particularly in the form of online reviews, has created a wealth of information for businesses and consumers alike. These reviews influence customer decisions, guide product development efforts, and provide valuable market research insights. However, analyzing this vast and often unstructured data presents a significant challenge.

This literature review explores the application of text analysis techniques for uncovering themes and gaining deeper understanding from user reviews. It examines both unsupervised and supervised learning approaches. The focus will be on topic modeling, a powerful unsupervised technique for identifying latent topics within a collection of user reviews. Additionally, the review will examine classification, a supervised learning approach for categorizing reviews or other relevant aspects.

By exploring these techniques and their applications in user review analysis, this review aims to provide a comprehensive understanding of how to extract valuable insights from this rich source of data. The review will discuss the strengths and weaknesses of each approach, along with their suitability for different research objectives. Finally, it will highlight potential future directions and advancements in user review analysis techniques.

2.2 Text Analysis Techniques

Text analysis, also known as Natural Language Processing (NLP), plays a crucial role in extracting and analyzing meaningful information from textual data. This information is valuable across various domains, including: Customer Relationship Management, Social Media Monitoring, Scientific Research.

Text analysis leverages various techniques from different fields, such as linguistics, statistics, and machine learning. This section delves into two prominent learning approaches used for text analysis: unsupervised and supervised learning.

Unsupervised learning algorithms work with data that lacks predefined categories or labels. The primary goal is to uncover hidden patterns or relationships within the data itself. Supervised learning algorithms rely on labeled data where each data point has a predefined category or label. The algorithm learns from this labeled data and can then be used to predict the category or label for new, unseen data.

2.3 Topic Modeling

Topic modelling is a statistical model for discovering abstract topics that occur in a collection of documents. It is a popular tool in text mining and NLP, used for organizing, understanding, and summarizing large datasets of textual information. There are two main topic modelling methods such as LDA and DMM.

Recent advancements in topic modelling have been influenced significantly by the advent of LLMs. These models, such as BERT (Devlin et al., 2018a) and Llama2 (Large Language Model Meta AI 2) (Touvron et al., n.d.) have set new benchmarks for understanding contextual nuances in text.

2.3.1 Latent Dirichlet Allocation (LDA)

LDA (Blei et al., 2003) is one of the most popular topic modelling techniques that view each document as a mixture of various topics and each topic as a collection of keywords. By this method, LDA can infer both the topics present in a corpus and the degree to which each document exhibits these topics.

In the realm of user reviews, (Zhou et al., 2022) employed online Latent Dirichlet Allocation (OLDA) to capture topic evolution and identify emerging topics, while

(Vallurupalli & Bose, n.d.) utilized a similar approach to explore the thematic composition of online reviews.

Beyond user reviews, similar methodologies have been applied in different contexts, such as analyzing COVID-19 trends. For instance, (Amara et al., 2021) utilized LDA to assess multilingual social media content on Facebook, demonstrating its capacity to discern evolving public discourse themes across various languages. This study underscores LDA's versatility in handling complex, real-world datasets. Another example is provided by (Mutiah et al., 2022).who also analyzed COVID-19 data using similar methodologies.

Additionally, LDA's utility extends to Twitter data analysis, as explored in studies by (Garcia & Berton, 2021; Guerzo et al., 2021)These works underline LDA's effectiveness in topic detection and sentiment analysis in different linguistic contexts, offering insights into public reactions during the pandemic.

Furthermore, (Guerzo et al., 2021) extend LDA's application to a bilingual context, combining it with clustering algorithms to examine disaster-related tweets in both English and Filipino. This approach demonstrates LDA's flexibility in handling multilingual and multifaceted datasets.

A. Word Representation

For LDA, it typically requires word representations such as Term Frequency-Inverse Document Frequency (TFIDF) (Ramos, n.d.), doc2bow, and count vectorizers before fitting into the model. Additionally, other techniques like Word2Vec and FastText have been used in combination with LDA in previous studies.

2.3.2 Dirichlet Multinomial Mixture (DMM)

Besides LDA, DMM is also a popular model, which is used in data extraction from extremely short texts (Li et al., 2021), it mentioned that LDA, loses effectiveness on

the sparseness of short texts, and DMM is sensitive to noisy words therefore it learns inaccurate topic representations. Besides, (N. Agarwal et al., 2020) this paper also used DMM to overcome short text problems.

2.3.3 Large Language Models

A. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a groundbreaking model in the field of NLP, developed by Google. Unlike traditional language models that process text in a linear fashion (either left-to-right or right-to-left), BERT is designed to understand the context of a word in a sentence by looking at the words that come before and after it. This bidirectional approach is a key differentiator and a significant advancement over previous models.

Bert can overcome the limitations of traditional methods such as LDA, that fail to capture the semantic relationship and are not specific to any domain(Udupa et al., 2022; Vasudeva Raju et al., 2022). Besides that, LDA ignores topic correlation since it believes all topics are independent, and with large corpora, and it has issues with sparsity.(Vasudeva Raju et al., 2022) Refer to (Rashid et al., 2019), Conventional approaches also falter in capturing word co-occurrence patterns in short texts found on the web, social media platforms like Twitter, and news headlines.

In contrast, BERTopic distinguishes itself by its ability to infer topics while preserving the original text structure with remarkable efficacy (Egger & Yu, 2022). When compared to Gibbs Sampling Dirichlet Multinomial Mixture Model (GSDMM) in terms of short text analysis, (Udupa et al., 2022) found that BERTopic yielded better results.

Specifically for topic modelling, BERTopic is often employed, as it is a subset of BERT. BERTopic has been utilized in various contexts, demonstrating its versatility and effectiveness across different domains. (Hendry et al., 2021) utilized BERTopic in customer service chats, highlighting its utility in analyzing conversational data.

Similarly, (Tounsi et al., 2023)applied BERTopic to analyze users of ChatGPT, demonstrating its suitability for understanding user interactions in online platforms. Additionally, BERTopic has been utilized in topic modeling tasks aimed at recommending candidate technologies for digital therapeutics (DTx), as demonstrated by (Jeon et al., 2023). In this study, patent documents were leveraged as a source of text data, showcasing BERTopic's adaptability to diverse text sources and its potential for generating actionable insights across different domains.

The incorporation of BERT-based methods marks a significant advancement in topic modelling. (George & Sumathy, 2023) illustrate the synergy between BERT and traditional methods like LDA, noting enhanced accuracy and topic coherence. This integration showcases the evolution of topic modelling techniques, blending the strengths of neural network-based models with established statistical methods.

In terms of Application in Domain-Specific Literature, BERTopic's effectiveness in domain-specific literature is highlighted by (Parlina & Maryati, 2023). Their study exemplifies how BERTopic excels in extracting specialized topics, demonstrating its adaptability to various fields of research. This underscores the model's capability to handle nuanced and specific thematic content.

In Financial Text Analysis, the work by(Vasudeva Raju et al., 2022) underlines the effectiveness of BERTopic approaches in dealing with intricate datasets, an essential feature for financial text analysis. Additionally,(Hristova & and Netov, 2022)employ BERTopic to analyze the impact of COVID-19 on education, focusing on media coverage and public perception of distance learning.

B. LLAMA2 (Large Language Model Meta AI 2)

LLAMA2, another prominent large language model, released by Meta, represents an updated iteration of LLAMA. Trained on 40% more data than its predecessor and boasting twice the text context length. LLAMA2 has been utilized in chat applications, as reported by (Touvron et al., 2023).

2.4 Evaluation

To assess LDA, commonly employed metrics include Topic Coherence and Perplexity. In the case of BERTopic, as recommended by the author (Grootendorst, 2022), the suggested evaluation method involves utilizing OCTIS (Optimizing and Comparing Topic Models is Simple).

2.4.1 Topic Coherence

Topic Coherence measures how similar these words are to each other. The coherence based on pointwise mutual information (PMI) gave the largest correlations with human ratings (Röder et al., 2015). Widely used coherence metrics include UMass and UCI.

A. UMass

The formula shown in Figure 2.4-1 is adapted from the paper by del (del Gobbo et al., 2021) and represents the UMASS metric for topic coherence. The UMASS metric measures the coherence of topics in a topic model by evaluating the co-occurrence of words within the same topic.

The formula is given by:

$$C(k, W^{(k)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D\left(w_m^{(k)}\right), D\left(w_l^{(k)}\right) + 1}{D\left(w_l^{(k)}\right)}$$

Figure 2.4-1 Formula for UMASS

B. UCI

The formula shown in Figure 2.4-2 is adapted from the paper by (Röder et al., 2015) represents the UCI metric for topic coherence. The UCI metric evaluates the coherence of topics by analyzing word pairwise co-occurrence statistics across documents.

$$C_{\text{UCI}} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{PMI}(w_i, w_j)$$

$$\text{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}$$

Figure 2.4-2 Formula for UCI

2.4.2 Perplexity and log-likelihood

Perplexity serves as an indicator of how effectively a probability distribution or model predicts a given sample, with a higher perplexity value indicating better performance. Conversely, log likelihood is a measure used to evaluate the goodness of fit of a statistical model, where a lower value is indicative of a better fit.

2.4.3 OCTIS

OCTIS a Python library, standardizes and automates topic modeling experiments, optimizing hyperparameters with Bayesian algorithms. Integrated with Weights & Biases, it allows for flexible metric evaluation and seamless result tracking and export. Ideal for efficient and comprehensive topic modelling analysis (Terragni et al., n.d.).

2.5 Classification

This section delves into text classification, a fundamental technique in text analysis. It explores various applications, embedding techniques, classification algorithms, and comparative studies to understand their effectiveness.

2.5.1 Text Classification Techniques

Text classification encompasses various techniques for assigning predefined categories to text data. Here, we explore traditional methods and highlight the growing importance of pre-trained models for tasks like app review categorization.

Traditional Techniques:

- Logistic Regression: A foundational model known for its simplicity and interpretability. It serves as a common baseline for evaluating the performance of more complex models. While interpretable, logistic regression might struggle with complex non-linear relationships present in text data.
- Random Forest: This ensemble learning method combines multiple decision trees, offering robustness and the ability to handle non-linear text data. Random forests are often praised for their accuracy and ability to handle high-dimensional data. However, interpreting their results can be challenging.
- XGBoost: An optimized gradient boosting algorithm known for its high performance and ability to handle complex classification tasks. XGBoost excels at tasks involving large datasets and provides some interpretability features compared to other boosting algorithms.
- GRU (Gated Recurrent Unit): An RNN architecture that captures temporal dependencies in text data. It's simpler than LSTMs but can be effective for tasks where long-term dependencies aren't crucial.
- BiLSTM (Bidirectional Long Short-Term Memory): A variant of LSTM that processes text in both directions (forward and backward), allowing it to capture context more effectively than traditional LSTMs.

Pre-trained Models: A Shift in Focus

While traditional techniques have served NLP well, recent advancements in pre-trained models have revolutionized text classification. These models are trained on massive amounts of text data, allowing them to capture complex linguistic patterns and contextual information. This is particularly valuable for tasks like app review categorization, where understanding user intent and sentiment is crucial.

One prominent example is BERT (Bidirectional Encoder Representations from Transformers)(Devlin et al., 2018b). BERT's ability to process text bidirectionally,

considering both preceding and following words, allows it to grasp the nuances of language. This is crucial for app reviews where a single word's meaning can depend on the surrounding context. Additionally, BERT can be fine-tuned for specific tasks like app review categorization with relatively small amounts of labeled data, making it highly adaptable for real-world applications.

2.5.2 Evaluation

Several evaluation metrics, such as Accuracy, Precision, Recall, and F1-score, are commonly used to assess the performance of text classification models. The reviewed studies demonstrate the effectiveness of pre-trained models and various machine learning techniques in achieving high performance on text classification tasks across diverse domains.

A. Accuracy

The proportion of correctly classified data points (True Positives + True Negatives) divided by the total number of data points.
Accuracy = (TP + TN) / (TP + TN + FP + FN)

B. Precision

The proportion of correctly predicted positive cases (True Positives) divided by the total number of predicted positive cases (True Positives + False Positives).
Precision = TP / (TP + FP)

C. Recall

The proportion of correctly identified positive cases (True Positives) divided by the total number of actual positive cases (True Positives + False Negatives).
Recall = TP / (TP + FN)

D. F1-score

A harmonic mean between Precision and Recall, providing a balanced view of both metrics.

F1-score = 2 * (Precision * Recall) / (Precision + Recall)

2.5.3 Recent advancements in text classification

Recent advancements in NLP have leveraged pre-trained models to enhance text classification tasks, including app review categorization. For instance, Wang et al. (Wang et al., 2022) used models like Sentence T5 to match app reviews with bug reports, demonstrating the potential of these techniques for categorizing review content. Similarly, XLM-R achieved promising results in classifying multilingual Twitter data (Manias et al., 2023)

Beyond app review matching, sentiment analysis is another crucial aspect. (Mouthami et al., 2023) explored sentiment analysis of film reviews using various word embeddings and models like Bi-LSTM. This highlights the importance of considering user sentiment when working with app reviews. Techniques like Bi-LSTM can be valuable for categorizing reviews based on positive/negative sentiment alongside identifying the review's subject (feature request/bug report).

Furthermore, (Ng & Carley, 2021) utilized BERT embeddings for classifying fact-checked coronavirus stories, showcasing the versatility of pre-trained models in different domains. Text classification research also explores various algorithms. While studies like (J. Agarwal et al., 2023) compare Naive Bayes variants for news sentiment analysis, (Dharrao et al., 2023) demonstrate the effectiveness of Support Vector Machines (SVM) and other classifiers in handling product review data. These studies collectively underscore the potential of pre-trained models, sentiment analysis techniques, and various classification algorithms for app review categorization tasks.

Future directions in user review analysis techniques may involve integrating more advanced models and approaches to further enhance the accuracy and reliability of classifications. Potential advancements could include the development of more sophisticated multilingual models, improved semi-supervised learning methods, and the application of deep learning architectures tailored for specific review contexts. Exploring these avenues could lead to more nuanced understanding and better handling of user reviews, ultimately contributing to improved customer insights and enhanced user experiences.

Chapter 3: Proposed Framework

3.1 Introduction

This chapter outlines the proposed framework for sentiment analysis of user reviews for the Selangkah app. The framework is divided into several sections, each addressing a specific component of the process.

3.2 Proposed Framework Overview

The proposed framework for our research integrates two main methodologies: Topic Modeling and Classification. Each methodology is systematically structured to handle the nuances of short text reviews, ensuring efficient preprocessing, analysis, and evaluation. The framework leverages a blend of traditional and advanced techniques, tailored specifically for multilingual datasets comprising Malay, English, and Chinese texts. The workflow is divided into several key stages, as illustrated in the accompanying Proposed Framework (Topic Modeling) (Figure 3.2-1) and Proposed Framework (Classification) (Figure 3.2-2).

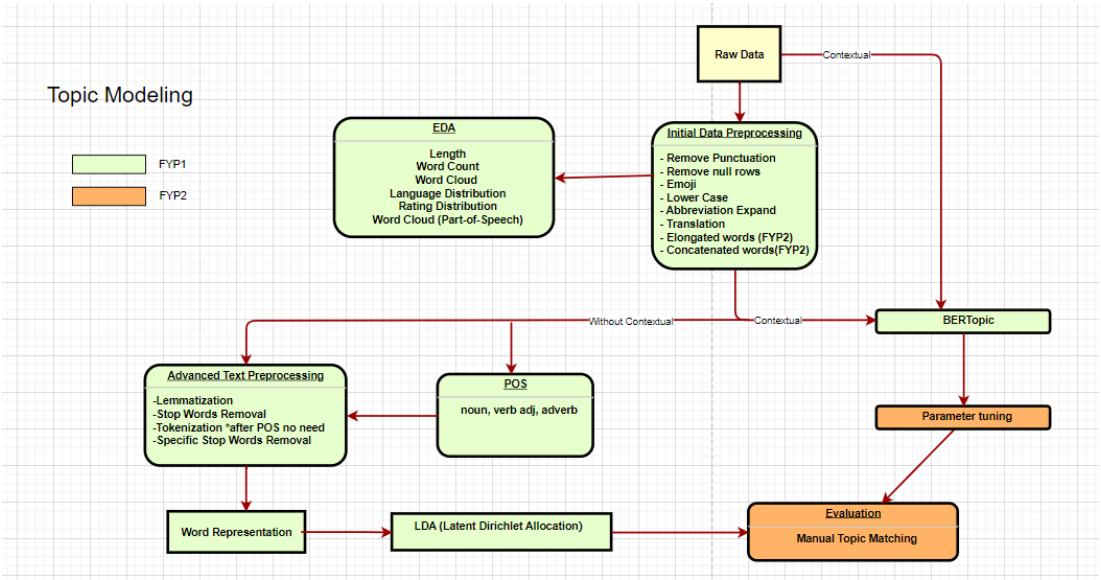


Figure 3.2-1 Proposed Framework (Topic Modeling)

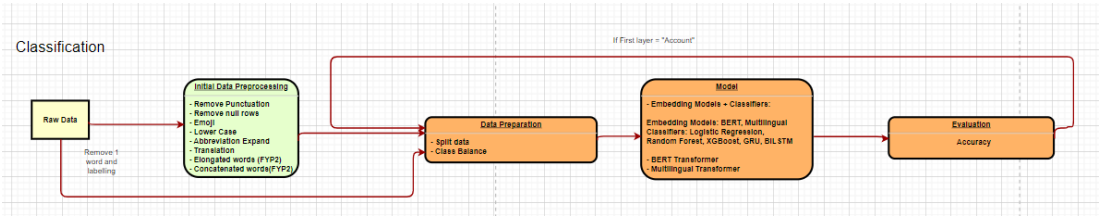


Figure 3.2-2 Proposed Framework (Classification)

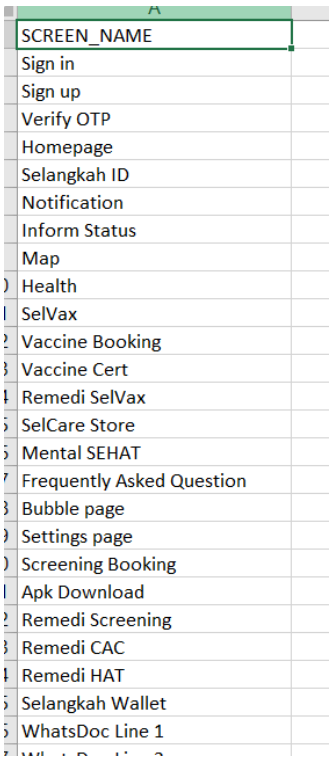
3.3 Data Set

Two datasets are utilized for this project:

- *Review* Dataset: Sourced from the Selangkah application on the Google Play Store, consisting of 6449 rows (as shown in Figure 3.3-1). This dataset includes user reviews in Malay, Chinese, and English.
- *Function Names* Dataset: Consists of keywords representing various functions provided by Selangkah, originally containing 50 functions (as shown in Figure 3.3-2) but refined to a core set of 5 main categories and their subcategories.

A	B	C	D	E
Review Text	Star Rating	Review Submit Date and Time		
Nice.. great apps!	5	2021-01-17T05:22:10Z		
	5	2021-01-17T06:08:23Z		
EVERYONE SHOULD HAVE THIS ALL NEW & UPDATED SELANGKAH APP! WHY? I'LL TELL YOU WHY! ----- I find it aw	5	2021-01-20T06:03:04Z		
Can't even sign up as new user. Blank screen when click on sign-up. And can't retrieve password also. Same blank s	1	2021-01-30T17:13:52Z		
Very fast repond for support, people centric and easy to use. Kudos to the development team and those who came	5	2021-02-09T05:08:58Z		
All in one goodness, great pandemic tool	5	2021-02-02T11:47:21Z		
If everyone do download the app, we find more interesting features on it .	5	2021-02-09T05:52:24Z		
All-in-one, so handy! My favourite feature is of course the super-cool scanner that can read both Selangkah and M	5	2021-02-09T05:53:53Z		
Nice app with a lot of useful features	5	2021-02-09T05:59:37Z		
This approach of putting all pandemic related services in one place is most welcomed. So far I have no issue registe	5	2021-02-09T05:59:55Z		
I applaud the concept and the sustainability of the app beyond the pandemic. But I can't register for Selangor ID.	3	2021-02-09T07:16:49Z		
	5	2021-02-09T07:27:08Z		
	5	2021-02-09T08:37:43Z		
Seems like there's bug in the app. After registering and came out from the registration page, the account doesn't se	3	2021-02-09T09:49:04Z		
ID sign up not working!	2	2021-02-09T09:58:14Z		
Wow ðŸ˜³ðŸ˜³	5	2021-02-09T07:06:39Z		
Yang bagus app ni,bile masuk, terus keluar scan. Tak payah nak tekan lagi check in. Setakat ni load pon cepat.	5	2021-02-10T00:16:35Z		
Please tambah boleh add partner, mysejahtera ada. Benda ni sangat convenient	4	2021-01-16T02:48:29Z		

Figure 3.3-1 Snipped of the *Review* Dataset



A
SCREEN_NAME
Sign in
Sign up
Verify OTP
Homepage
Selangkah ID
Notification
Inform Status
Map
Health
SelVax
Vaccine Booking
Vaccine Cert
Remedi SelVax
SelCare Store
Mental SEHAT
Frequently Asked Question
Bubble page
Settings page
Screening Booking
Apk Download
Remedi Screening
Remedi CAC
Remedi HAT
Selangkah Wallet
WhatsDoc Line 1

Figure 3.3-2 Snipped of the *Function Names* Dataset

3.4 Function Defined

This section outlines the core functionalities of the Selangkah app and provides a detailed breakdown of the subcategories for the 'Account' functionality. Understanding these categories is crucial for accurate topic modeling and classification of user reviews.

3.4.1 Selangkah App Functionalities

The functionalities of the Selangkah app have been redefined into five main categories, with detailed subcategories for 'Account'.

Table 3.4-1 summarizes the main functionalities.

Table 3.4-1 Function Category Description

Category	Description
Account	Issues related to sign in
Finance	Welfare-related issues such as Bingkas (support for low-income families) and ASAS (support for children under six years old)/ E-Wallet
Health	Issues related to Covid-19 or other health functionalities in Selangkah such as Selcare
Others	Suggestions
Rating	General ratings without specific details

Table 3.4-2 details the subcategories within 'Account'.

Table 3.4-2 Account Function Subcategory Descriptions

Subcategory	Description
Sign In	Issues related to signing into the app
Sign Up	Problems with the registration and password processes
APK Download	Difficulties in downloading the app from Google Play Store
Data Management	Complaints about data not being updated or incorrect information
Loading and Performance	Feedback on app loading and performance

Justification for No Subclasses in Other Categories: While "Account" functionalities benefit from further subclassification due to their detailed nature, other categories are not subclassified for the reasons explained in Table 3.4-3.

Table 3.4-3 Reasons for Not Having Subclasses

Category	Reason for No Subclasses
Health	The data is skewed towards Covid-19 (70%), which is no longer a predominant concern, and general health (30%), making further subclassification less insightful.
Finance	Potential subcategories (welfare, e-wallet) exist, but dataset size is too small for meaningful analysis.
Rating	General nature of the category makes subclassification unnecessary.
Others	Varied and unspecific feedback within the category makes subclassification unproductive.

3.4.2 Mapping Reviews to Functionalities

Reviews are categorized based on their relevance to the specific functionalities of the app. This involves:

- Manual Labeling: A subset of reviews is manually labeled to create a training dataset.
- Keyword Identification: Keywords and phrases indicative of each functionality are identified to facilitate automated categorization. For instance, words like "login," "register," and "password" might be associated with the 'Account' functionality, while "appointment," "Covid-19," and "health record" might relate to the 'Health' functionality.

3.5 Initial Data Preprocessing

This section details the data preprocessing steps employed to prepare the text data for analysis. Preprocessing is crucial for ensuring the data is clean and suitable for machine learning or other analytical techniques.

3.5.1 Text Cleaning

The initial stage of preprocessing involves text cleaning, which addresses various inconsistencies and noise within the data. The following cleaning steps are applied:

- Punctuation Removal: Punctuation marks such as commas, periods, and exclamation points are removed from the text. This helps standardize the text and focus on the core meaning of the words.
- Null Row Removal: Empty rows or rows with missing values (null) are identified and removed from the dataset. These rows do not contribute to the analysis and can potentially skew the results.
- Emoji Removal: Emojis are removed from the text as their interpretation can be subjective and may not be directly relevant to the sentiment analysis.

- Lowercasing: All text is converted to lowercase. This ensures consistency in the data and avoids potential bias due to capitalization differences.

3.5.2 Abbreviation Handling

The dataset contains various abbreviations specific to both Malay and English languages. These abbreviations can hinder the analysis process if left unexpanded. To address this, a predefined list of abbreviations is utilized. This list, detailed in Figures 3.5-1 (Malay) and 3.5-2 (English), provides the full forms for each abbreviation encountered in the data. By replacing abbreviations with their corresponding full forms, the meaning of the text becomes clearer and more consistent.

Table 3.5-1 List of Malay Abbreviations

"Abbr": Abbreviation

No	Abbr	Description	No	Abbr	Description	No	Abbr	Description
1	ady	already	34	kenapa	kenapa	67	pn	puan
2	ape	apa	35	klr	keluar	68	ptg	petang
3	bbrp	beberapa	36	knapa	kenapa	69	sbb	sebab
4	bg	bagi	37	knp	kenapa	70	sdh	sudah
5	bgus	bagus	38	korang	kamu orang	71	sgt	sangat
6	bkn	bukan	39	kpd	kepada	72	skrg	sekarang
7	blh	boleh	40	krg	kurang	73	slh	salah
8	blm	belum	41	krj	kerja	74	sm	sama
9	bnyk	banyak	42	krn	kerana	75	smlm	semalam
10	brg	barang	43	kt	dekat	76	spj	sampai
11	brp	berapa	44	kwn	kawan	77	ssh	susah
12	byk	banyak	45	lbh	lebih	78	sy	saya
13	cik	encik	46	lg	lagi	79	tak	tidak
14	cpt	cepat	47	lm	lama	80	tdr	tidur

15	d	already	48	lps	lepas	81	tgh	tengah
16	dgn	dengan	49	mcm	macam	82	tlg	tolong
17	dh	dah	50	mdh	mudah	83	tmpt	tempat
18	dkt	dekat	51	mkn	makan	84	tp	tapi
19	dlm	dalam	52	mknn	makanan	85	tq	thank
20	dr	dari	53	mlm	malam	86	ttg	tentang
21	fhm	faham	54	mn	mana	87	utk	untuk
22	gak	juga	55	msh	masih	88	x	tidak
23	hdp	hidup	56	msk	masuk	89	xbole	tidak boleh
24	hr	hari	57	nak	mahu	90	xboleh	tidak boleh
25	hrp	harap	58	ni	ini	91	xdapt	tidak dapat
26	htr	hantar	59	nk	nak	92	xde	tidak kerja
27	je	sahaja	60	org	orang	93	xdpt	tidak dapat
28	jgn	jangan	61	pd	pada	94	xjadi	tak jadi
29	jln	jalan	62	pdhl	padahal	95	xleh	tidak
30	jom	jom	63	pgi	pagi	96	xlepas	tidak lepas
31	kali	kali	64	pgr	pengarah	97	xupdate	tidak update
32	kat	dekat	65	pk	pukul			
33	keje	kerja	66	pls	please			

Table 3.5-2 List of English Abbreviations

No	Abbr	Description	No	Abbr	Description	No	Abbr	Description
1	afaik	as far as I know	20	idk	I don't know	39	tba	to be announced
2	atm	at the moment	21	idc	I don't care	40	tbh	to be honest
3	brb	be right back	22	imo	in my opinion	41	tl;dr	too long; didn't read
4	btw	by the way	23	irl	in real life	42	tq	thank you

5	cov	covid -19	24	iso	isolation	43	ty	thank you
6	c-19	covid-19	25	lmk	let me know	44	u	you
7	demo	demonstration	26	l8r	later	45	ui	user interface
8	dev	developer	27	obv	obviously	46	ux	user experience
9	dl	download	28	omg	oh my god	47	vax	vaccine
10	dwl	download	29	pls	please	48	vent	ventilator
11	e2e	end-to-end	30	pwd	password	49	ver.	version
12	easy	easy	31	quar.	quarantine	50	yr	your
13	eta	estimated time of arrival	32	rip	rest in peace	51	ur	your
14	fav	favorite	33	afaik	as far as I know	52	5*	5 stars
15	ft	feature	34	sry	sorry	53	a+	excellent
16	fyi	for your information	35	sd	social distancing	54	f-	fail
17	gr8	great	36	smh	shaking my head	55	n	and
18	grat	great	37	thx	thanks			
19	hrs	hours	38	tbd	to be determined			

3.5.3 Out-of-Vocabulary (OOV) Cleaning

To further improve the quality of the data and translation accuracy, Out-of-Vocabulary (OOV) cleaning is performed. This step focuses on handling words that might not be present in the standard vocabulary used for analysis. Two specific types of OOV words are addressed:

- Elongated Words (English & Malay): Words where characters are excessively repeated (e.g., looooong in English, baaaaikkkk in Malay) can cause issues during

processing. These words are normalized by replacing repetitions with a single character (looooong becomes long, baaaaikkkk becomes baik).

- Concatenated Words (English): Words formed by joining multiple words without spaces (e.g., goodmorning) can be misinterpreted by translation models. These words are segmented based on context or predefined rules (goodmorning becomes good morning).

By addressing OOV words, the data becomes more consistent and suitable for further analysis, particularly for machine learning tasks like sentiment analysis.

3.5.4 Language Detection and Translation

The dataset includes reviews written in languages other than English. To ensure all reviews contribute to the analysis, a two-step approach is implemented:

- Language Detection: The Lingua library, a comprehensive language detection tool, is used to identify reviews written in Malay and Chinese. This allows for focusing on these specific languages besides English.
- Translation: Reviews identified as non-English undergo translation into English using the Google Translate API. This API offers wide accessibility and the capability to handle the nuances of translation. The goal is to retain the original sentiment and intent of the feedback during the translation process.

By implementing these data preprocessing steps, the text data is cleaned, standardized, and prepared for further analysis. This ensures the analysis captures the true meaning and sentiment expressed within the reviews.

3.6 Exploratory Data Analysis (EDA)

EDA focuses on exploring the overall properties of the text data. The following analyses are performed:

- Length: The distribution of review lengths (number of characters or words) is analyzed to understand the typical length of reviews and identify any outliers. This can provide insights into user behavior and writing styles.
- Word Count: The total number of words in the entire dataset is calculated. This provides a basic measure of the text volume available for analysis.
- Word Cloud: A word cloud is generated to visualize the most frequent words used in the reviews. This helps identify prominent topics and keywords that might be relevant for further analysis. To gain deeper insights into the content, the word cloud can be further enriched by incorporating information from the POS tagging process. This allows us to visualize high-frequency nouns, verbs, adjectives, and adverbs, providing a more nuanced understanding of the vocabulary used in the reviews.
- Language Distribution: The proportion of reviews written in different languages (e.g., English, Malay, Chinese) is identified. This information is crucial for understanding the multilingual nature of the data and determining if translation steps were necessary during preprocessing.
- Rating Distribution: The distribution of user ratings is analyzed. This helps visualize user sentiment and identify potential biases towards positive or negative feedback.

3.7 Topic Modeling

This section explores the topic modeling techniques employed within the proposed framework for sentiment analysis. The framework utilizes two main approaches: Latent Dirichlet Allocation (LDA) and BERTopic. These techniques cluster reviews based on subtopics within the Account, Health, and Finance categories, a total of seven subtopics, which will be further illustrated in the evaluation section. The topic

modeling focuses on identifying themes within the user feedback and categorizing them accordingly, aiming to verify alignment with predefined subtopics.

3.7.1 LDA

Text classification often involves uncovering the underlying themes or topics within a collection of documents. LDA is a powerful statistical technique for topic modeling. It analyzes a corpus of text data and identifies latent topics, which are hidden thematic structures that explain the generation of documents.

LDA assumes that each document is a mixture of these latent topics, and each word within a document is generated based on the topic distribution of that document. By analyzing word co-occurrence patterns and document structure, LDA helps us discover these hidden topics and understand the thematic landscape of a text collection.

However, before applying LDA, effective text preprocessing and feature representation are crucial steps. Here, the focus will be on these essential processes:

- Text Preprocessing: This stage prepares the raw text data for LDA by removing noise and inconsistencies. Two approaches to preprocessing will be explored: direct preprocessing and POS filtering with preprocessing.
- Feature Representation: This involves transforming the preprocessed text data into a numerical representation suitable for LDA analysis. Two commonly used feature representation techniques will be discussed.

A. Text Preprocessing

Data Preprocessing for LDA offers two pathways:

1. Direct Preprocessing: This path bypasses POS tagging and directly applies essential steps like lemmatization (converting words to base form), stop word removal (general and specific stop words - details in Figure 3.7-1), and tokenization.

```
['ok','okay','oke', 'okk', 'okayyy','kk','k','okei', 'good',
'can', 'please', 'even', 'excellent', 'thank', 'want', 'well',
'i', 'I','good', 'la'
,'also', 'ye', 'still','app','mysejahtera','not','ler']
```

Figure 3.7-1 List of Specific Stop Words

2. POS Filtering with Preprocessing: This path incorporates Part-of-Speech (POS) tagging to identify the grammatical function of each word (e.g., noun, verb, adjective). Words with specific POS tags (e.g., pronouns, conjunctions) might be excluded before applying lemmatization, stop word removal, and tokenization.

POS tagging enhances the model's ability to understand the relationships between words. By focusing on relevant parts of speech, the framework aims to improve the quality of the data for topic modeling. The overall goal of text Preprocessing is to prepare the text data for subsequent tasks (word embedding and topic modeling) by removing unnecessary elements and enhancing its suitability for analysis.

B. Feature Representation for LDA

The choice of feature representation technique can significantly impact the effectiveness of LDA for sentiment analysis. This section discusses two commonly used techniques:

- TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF assigns weights to words based on their frequency within a document and their rarity across the entire corpus. This helps to emphasize terms that are distinctive to specific sentiments.
- Count Vectorizer: This technique simply counts the occurrences of each word within a document. While straightforward, it might not capture the relative importance of words for sentiment analysis.

3.7.2 BERTopic

The proposed framework utilizes BERTopic (Grootendorst, 2022), a powerful tool that leverages BERT-derived embeddings for topic discovery in text data. BERTopic offers a flexible approach to uncovering and characterizing topics, making it particularly well-suited for sentiment analysis tasks. The framework provides two paths: one with preprocessing and another without preprocessing.

A. Core Functionalities:

BERTopic employs several key functionalities within the framework:

- Document Embeddings: Semantic representations of documents are created using Sentence-BERT (SBERT), capturing the meaning of each document.
- Topic Grouping: HDBSCAN, a density-based clustering algorithm, groups documents into distinct topics without requiring prior knowledge of the number of topics.
- Topic Characterization: Significant words for each topic are identified using class-based Term Frequency-Inverse Document Frequency (c-TF-IDF), with BERTopic providing descriptive labels to aid interpretation.
- Fine-tuning (Optional): Topic representations can be fine-tuned using other large language models (LLMs) like GPT or LLAMA 2 for specialized topics or more detailed analysis.

B. Experimentation and Tuning:

The proposed framework leverages BERTopic's capabilities for topic modeling, which involves parameter tuning. While BERTopic offers default settings, this section explores the effects of tuning parameters like the number of topics and the minimum topic size to potentially optimize model performance.

Parameters Tuned:

- Number of Topics: This parameter adjusts the number of topics identified beyond the default "auto" setting. Tuning this parameter allows for exploring different granularities of topic representation.
- Minimum Topic Size: This parameter sets the smallest size for a topic to be considered valid. Tuning this parameter can influence the granularity of topics and potentially reduce noise.

Evaluation Metrics:

To evaluate the results of parameter tuning, the framework utilizes the code from BERTopic_evaluation for OCTIS (Optimizing and Comparing Topic Models is Simple). The evaluation involves the following metrics:

- Normalized Pointwise Mutual Information (NPMI): This metric measures the coherence of topics by evaluating the co-occurrence of words within a topic. Higher NPMI indicates better topic coherence, meaning words within a topic are more semantically related.
- Diversity Score: This metric assesses the uniqueness of the topics by measuring the difference in terms between topics. A higher diversity score indicates topics are more distinct and cover a wider range of concepts.
- Computation Time: This metric evaluates the efficiency of the model in terms of processing time. Tuning parameters can sometimes impact processing speed, so this metric is considered.

3.7.3 Evaluation

The discovered topics are manually compared to seven predefined subcategories based on the functionalities of the Selangkah app (described in section 3.4). These subcategories include: Sign In, Register, APK Download, General Health, COVID-19 E-wallet, Welfare. This manual evaluation ensures the topics discovered by the

clustering models (e.g., LDA, BERTopic) capture the specific concerns users have within each subcategory.

3.8 Transition to Classification

Clustering techniques provided valuable insights into user review themes, but limitations arose in precisely categorizing feedback for specific Selangkah app functionalities. To address this, the project shifted to a supervised learning approach, focusing on classification techniques. Supervised learning leverages labeled data to train models that assign new, unlabeled data points to predefined categories. This allows categorization of user reviews based on discussed functionalities, leading to a more granular understanding of user concerns and preferences.

3.9 Classification

This section explores the classification approach employed within the framework for sentiment analysis. The objective is to categorize user reviews into various categories to facilitate a more granular understanding of user feedback. The first layer will classify each review into one of the five main categories (account, finance, health, others, rating) as predefined in Section 3.4. The second layer will further categorize reviews into subcategories for the account category. Both layers utilize the same procedure for classification.

3.9.1 Self-Labeling

Creating a reliable dataset with labeled reviews was necessary to train classification models. Manually labeling the entire dataset would be time-consuming, so a self-labeling approach was employed. Single-word reviews were removed, resulting in a 1433-review dataset. These reviews were then manually labeled according to their discussed functionalities. This labeled subset serves as the training data for the classification models.

3.9.2 Model Training

The classification models were categorized based on the text embedding techniques they utilize. Text embeddings represent textual data numerically, enabling machine learning models to process and analyze reviews. Three main groups of embeddings were explored: BERT embeddings, Multilingual embeddings, and Transformers.

A. BERT Embeddings:

For reviews processed using BERT embeddings, various classification models were trained, including Logistic Regression, Random Forest, XGBoost, GRU, and BiLSTM. These models offer varying strengths in classification tasks, allowing for a comprehensive exploration of the most effective approach for Selangkah user review classification.

B. Multilingual Embeddings:

Similar to the BERT embedding approach, models employing Multilingual embeddings were trained, including Logistic Regression, Random Forest, XGBoost, GRU, and BiLSTM. Comparing the performance of models trained with different embedding techniques will reveal the most suitable representation for capturing the nuances of user reviews within the Selangkah app.

C. Transformers:

Transformers, a powerful class of deep learning models, were also employed for user review classification. This included training a BERT Transformer and a Multilingual Transformer on the labeled data. Evaluating the performance of these models will reveal their effectiveness in classifying user reviews compared to the simpler models trained with individual embedding techniques.

3.9.3 Evaluation

Standard performance metrics like accuracy, precision, recall, and F1-score will be used to assess the effectiveness of the trained classification models. Accuracy measures the overall proportion of correctly classified reviews. Precision reflects the ratio of correctly classified reviews within the positive predictions made by the model. Recall indicates the proportion of relevant reviews identified by the model. Finally, F1-score provides a harmonic mean of precision and recall, offering a balanced view of the model's performance. Analyzing these metrics will determine the most successful classification model for categorizing Selangkah user reviews based on their discussed functionalities.

Chapter 4: Result and Discussion

4.1 Introduction

This chapter presents the results of the sentiment analysis and discusses the findings, highlighting the effectiveness of the proposed framework in analyzing user reviews of the Selangkah app.

4.2 EDA Results

This section presents an exploration of the Selangkah user review dataset to understand its characteristics and gain insights into user sentiment.

A. Number of rows and columns

The initial dataset contains 6448 rows (reviews). However, 4439 rows have missing values in the "review text" column. To ensure data quality, these rows were removed, leaving a final dataset of 2009 reviews for further analysis. Figure 4.2-1 shows the overall columns present in the raw data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6449 entries, 0 to 6448
Data columns (total 3 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Review Text                 2009 non-null   object
 1   Star Rating                 6449 non-null   int64
 2   Review Submit Date and Time  6449 non-null   object
dtypes: int64(1), object(2)
memory usage: 151.3+ KB
```

Figure 4.2-1 Overall Columns for Raw Data

B. Number of Rating Each Year

The dataset includes reviews spanning from the app's launch in January 2021 to September 2023. A decline in the number of reviews over time is observed, with approximately 1700 in 2021, 200 in 2022, and 100 in 2023 (Figure 4.2-2).

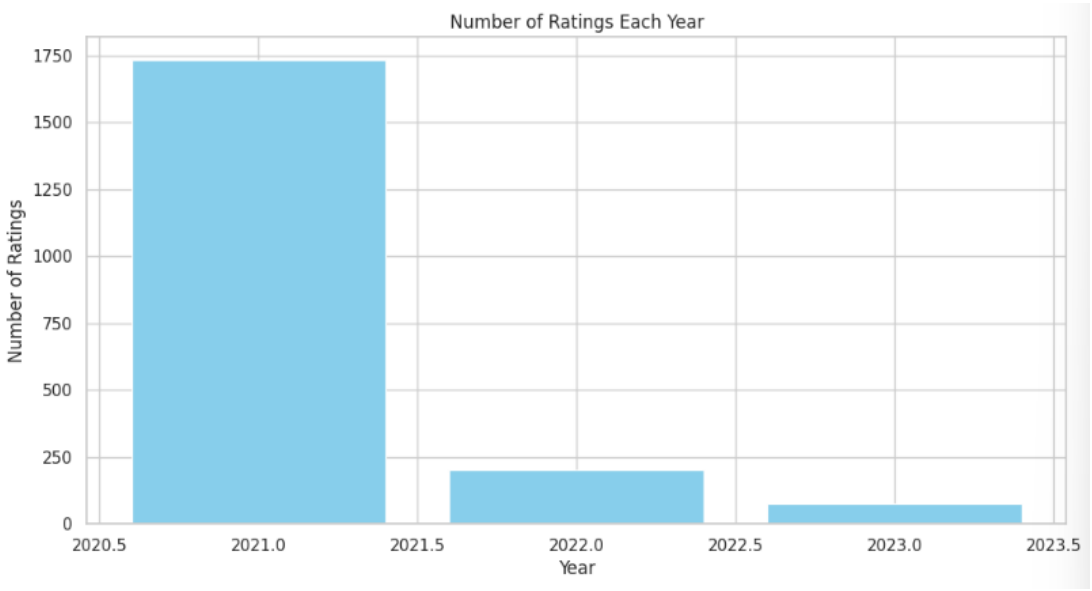
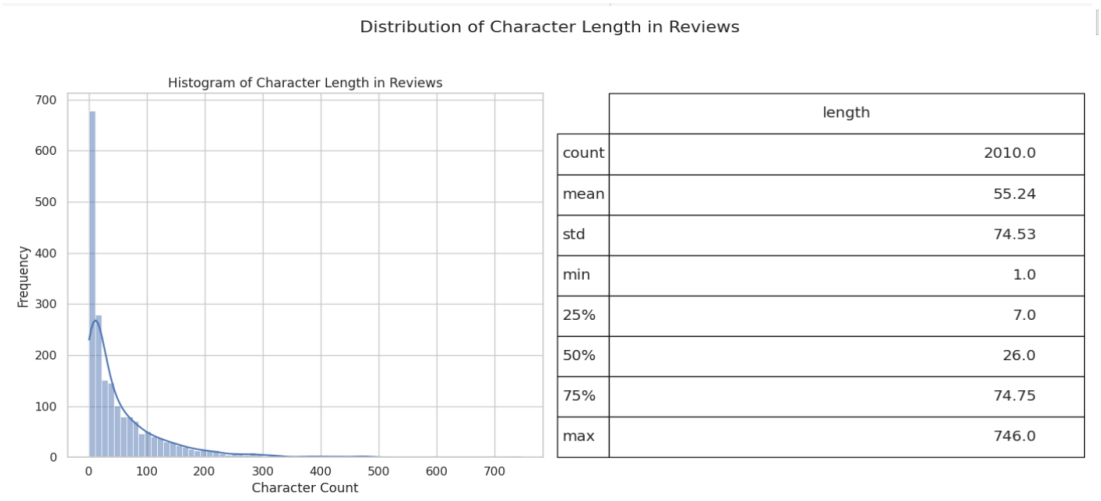


Figure 4.2-2 Number of Ratings Each Year After Removing Null Rows

C. Length (Number character of words) & Word Count

The analysis of word count and character count distribution revealed that most reviews are short. A significant portion (around 575) consists of only one word (e.g., "good"). The average review length is around 10 words, with a median of 4 words. Interestingly, a trend emerges when we consider word count in relation to rating. Reviews with lower ratings (1-star and 2-star) tend to be longer compared to those with higher ratings (4-star and 5-star). These findings are further supported by the distribution of character length and word count visualized in Figures 4.2-3 and 4.2-6.

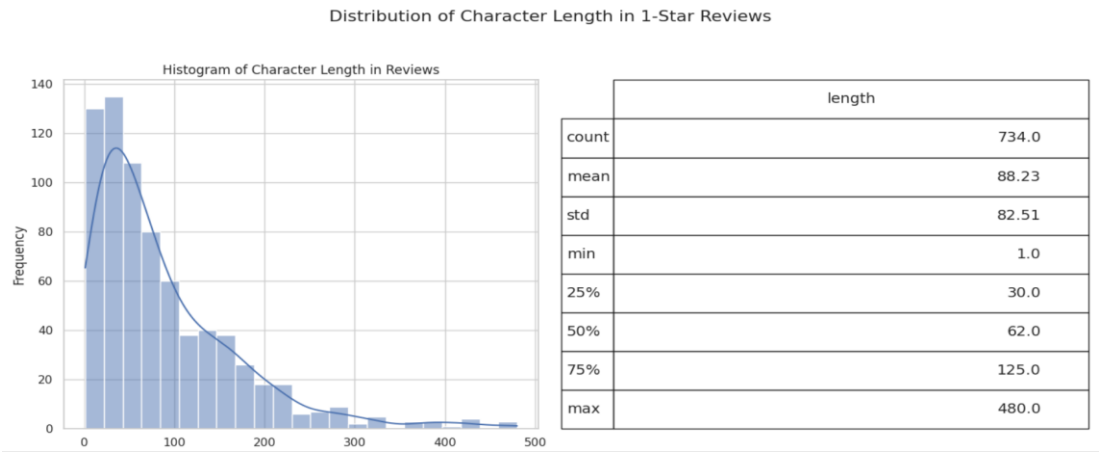
Figure 4.2-3 illustrates the histogram of character lengths in all reviews, showing the frequency of reviews with varying character counts.



	length
count	2010.0
mean	55.24
std	74.53
min	1.0
25%	7.0
50%	26.0
75%	74.75
max	746.0

Figure 4.2-3 Distribution of Character Length in All Reviews

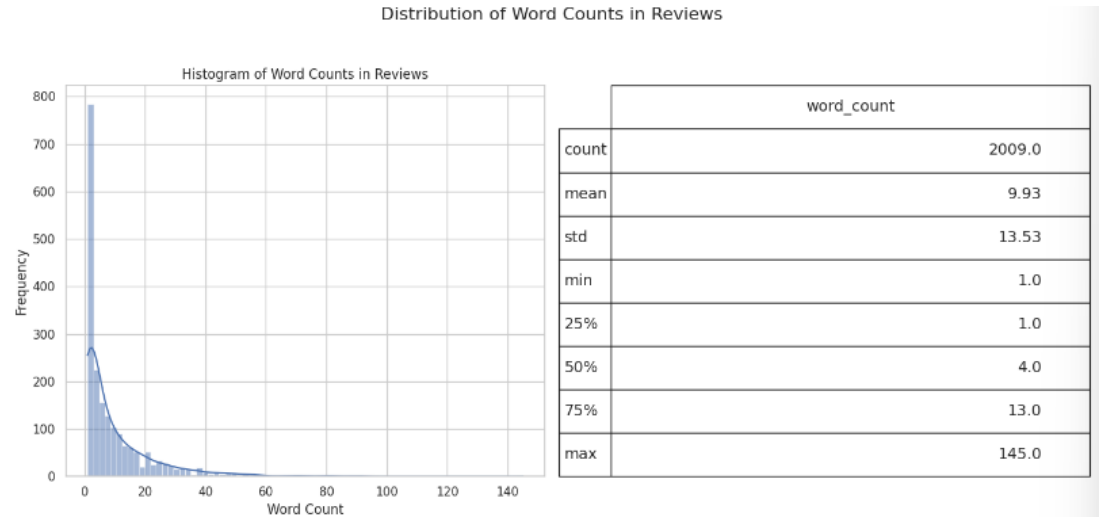
Figure 4.2-4 illustrates the histogram of character lengths in 1- Star reviews, showing the frequency of reviews with varying character counts.



	length
count	734.0
mean	88.23
std	82.51
min	1.0
25%	30.0
50%	62.0
75%	125.0
max	480.0

Figure 4.2-4 Distribution of Character Length in 1-Star Reviews

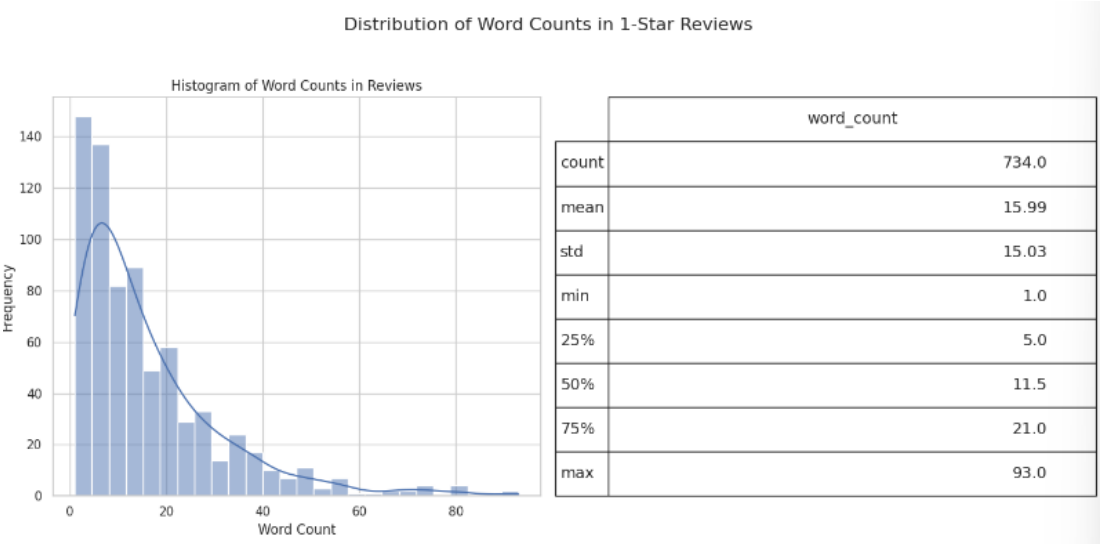
Figure 4.2-5 illustrates the histogram of word counts in all reviews, showing the frequency of reviews with varying word counts.



	word_count
count	2009.0
mean	9.93
std	13.53
min	1.0
25%	1.0
50%	4.0
75%	13.0
max	145.0

Figure 4.2-5 Distribution of Word Counts in All Reviews

Figure 4.2-6 illustrates the histogram of word counts in 1-Star reviews, showing the frequency of reviews with varying word counts.



	word_count
count	734.0
mean	15.99
std	15.03
min	1.0
25%	5.0
50%	11.5
75%	21.0
max	93.0

Figure 4.2-6 Distribution of Word Counts in 1-Star Reviews

D. Language Distribution

Analysis of the "review text" column revealed the presence of three languages: Malay, English, and Chinese. The specific distribution of each language can be found in Figure 4.2-7.

```
Number of Malay reviews: 855
Number of English reviews: 1098
Number of Chinese reviews: 25
Number of Tamil reviews: 0
```

Figure 4.2-7 Language Distribution

E. Word Cloud

This section summarizes the key findings from the word cloud analysis of user reviews after translation.

Overall Word Frequency:

The most frequent words across all reviews include "app," "update," "time," "login," "register," "unable," and "vaccine" (Figure 4.2-8) (Figure 4.2-9).

- Nouns: "update," "app," "vaccine," "time"
- Adjectives: "update," "app," "many," "bad"
- Verbs: "register," "open," "update," "use"
- Adverbs: "already," "really," "still," "always"

Word Frequency by Rating:

- Rating 1 (Figure 4.2-10): Words like "update," "register," "open," "vaccine" are prominent. Adverbs like "already," "still" suggest frustration.
- Rating 2 (Figure 4.2-11): Words like "update," "scan," "password," "slow" are common. Adverbs like "still," "always" suggest ongoing issues.
- Rating 3 (Figure 4.2-12): Words like "update," "open," "time," "phone," "vaccine" are frequent. Verbs like "open," "try" suggest attempts to use the app.
- Rating 4 (Figure 4.2-13): Words like "update," "open," "app," "easy," "use" are prominent. Adjectives like "easy," "friendly" indicate positive experiences.

- Rating 5 (Figure 4.2-14): Words like "app," "nice," "helpful" are dominant. Adverbs like "fast" suggest a positive impression.

These observations are further corroborated by the word cloud visualizations presented in Figures 4.2-8 to 4.2-14.

Figures 4.2-8 visualises the most common words found in all user reviews after initial preprocessing, highlighting the prominent terms used by users. Words are displayed in varying sizes, where larger words appear more frequently in the user reviews.

Figure 4.2-8 Word Cloud for All User Reviews after Initial Preprocessing (after Initial Preprocessing)

Figure 4.2-9 presents a word cloud that visualizes the most common words found in user reviews after the initial preprocessing steps, with a focus on part-of-speech (POS) tagging.

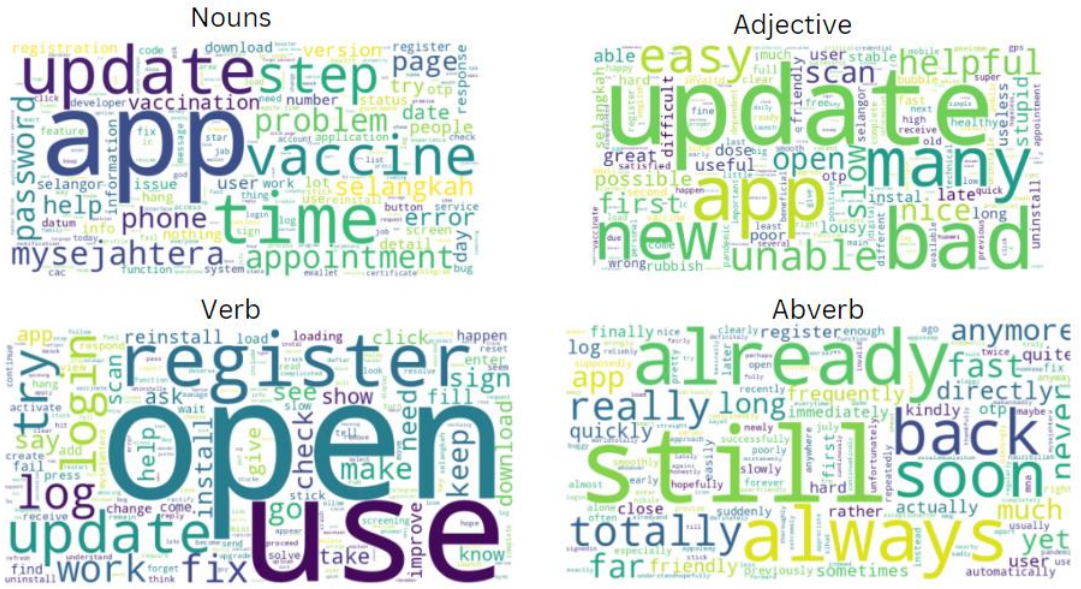


Figure 4.2-9 POS Tagging Word Cloud for All User Reviews (after Initial Preprocessing)

Figures 4.2-10 to 4.2-14 contain several word clouds that visualize the most common words found in user reviews after initial preprocessing, categorized by their part-of-speech (POS) tags. Each figure is divided into four sections: All Words, Adverbs, Adjectives (ADJ), Nouns, and Verbs.

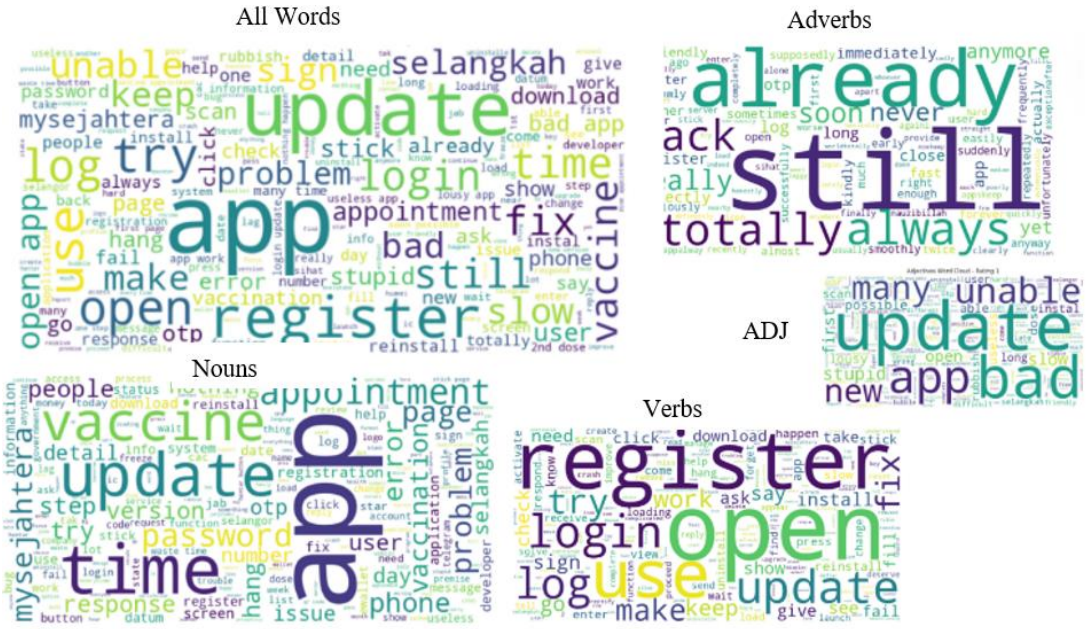


Figure 4.2-10 Word Clouds for Rating 1 Reviews (after Initial Preprocessing)

Figure 4.2-11 Word Clouds for Rating 2 Reviews (after Initial Preprocessing)

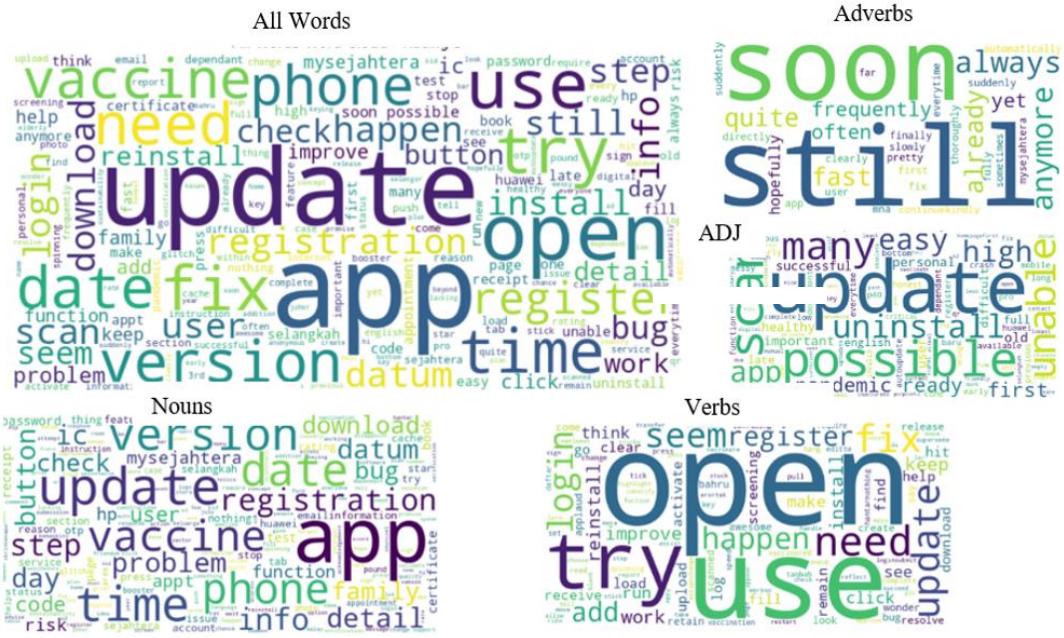


Figure 4.2-12 Word Clouds for Rating 3 Reviews (after Initial Preprocessing)

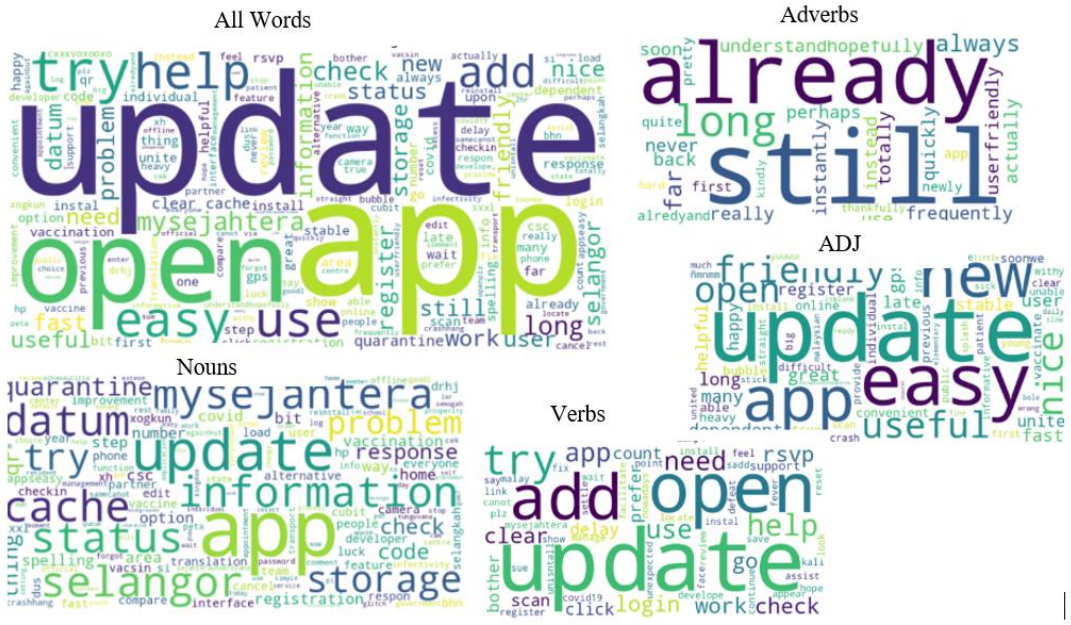


Figure 4.2-13 Word Clouds for Rating 4 Reviews (after Initial Preprocessing)

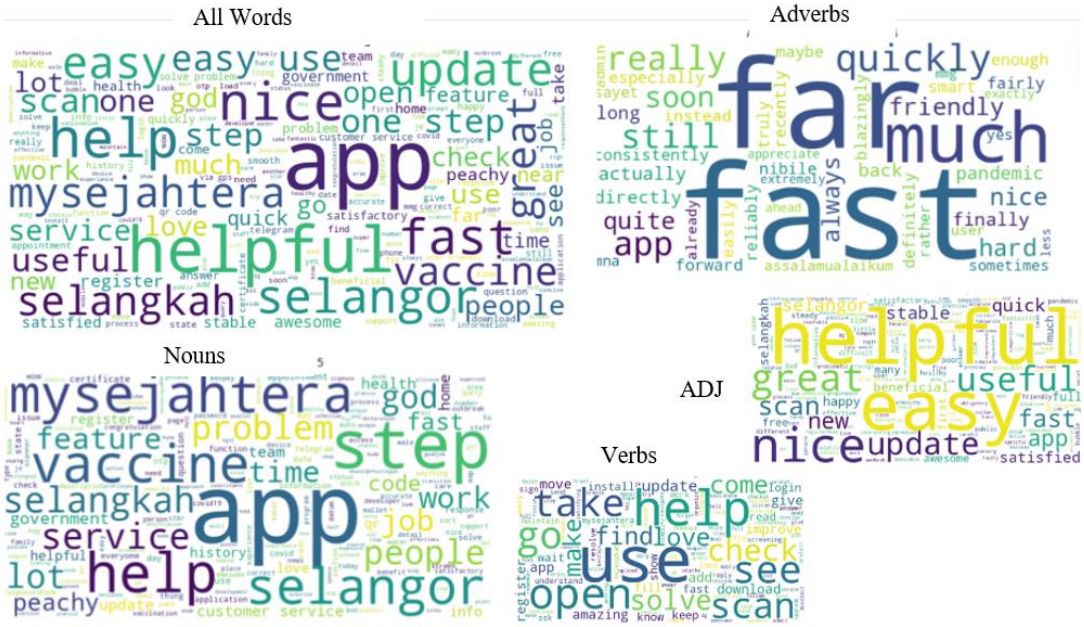


Figure 4.2-14 Word Clouds for Rating 5 Reviews (after Initial Preprocessing)

4.3 Initial Preprocessing Result

Following the completion of preprocessing, presented below is a sample of the obtained results. This section highlights the outcomes of various preprocessing steps,

including data cleaning, normalization, and transformation, which are essential for preparing the dataset for subsequent analysis.

Following the completion of Preprocessing 1, presented below is a sample of the obtained results (Table 4.3-1).

Table 4.3-1 Initial Data Preprocessing Result

Field	Review
Original Review	keputusan swab test di dewan mbsa seksyen 28 pada 27022021 tak keluar dalam apps terlepas flight ke kota kinabalu pagi ni 28022021 malang sungguh
After Preprocessing result	swab test results at the mbsa hall section 28 on 27022021 were not out in the apps to miss a flight to kota kinabalu this morning 28022021 unfortunately

4.4 Topic Modeling Results

This section delves into the results obtained by applying topic modeling techniques to cluster function names for LDA and BERTopic.

4.4.1 Analysis of Function Names in BERTopic Results

The key finding from (Table 4.4-1) is that preprocessing the data by removing stop words and potentially lemmatizing the function names significantly improves the alignment between the automatically generated topics and the actual function names. With preprocessing, a higher percentage of topics directly matched the function names (e.g., 63% vs. 57% for 7 topics). This suggests that preprocessing helps BERTopic focus on the core meaning of the function names and reduces the influence of irrelevant words.

Even without preprocessing, BERTopic was able to identify meaningful topics within the function names, with over 50% of the topics matching function names. This indicates that BERTopic can be effective in extracting semantic structure from

function names. However, preprocessing provides a significant boost in the accuracy and interpretability of the results.

Table 4.4-1BERTopic Results

Preprocessing	Total Rows	No. of Selected Topics	Total Topics Generated by BERTopic	No. of Topics Match with Function Names	Matched Percentage
No	2000	7	40 (auto)	5	63%
		7	20	4	57%
Yes	1433	7	33 (auto)	6	85%
		7	19	7	100%

Table 4.4-2 presents additional BERTopic results for comparison with LDA when the number of selected topics is set to 11. With preprocessing, 57% of the topics matched the function names, indicating a notable alignment even when the number of topics was adjusted. Further discussion will be provided in section 4.6.

Table 4.4-2Additional BERTopic Results

Preprocessing	Total Rows	No. of Selected Topics	Total Topics Generated by BERTopic	No. of Topics Match with Function Names	Matched Percentage
Yes	1433	7	11	4	57%

4.4.2 Analysis of Function Names in LDA Results

A. Advanced Preprocessing Results

Table 4.4-3 shows the completion of Advanced Text Preprocessing, presented below is a sample of the obtained results.

Table 4.4-3 Advanced Text Preprocessing Result

Field	Review
After-Translate Review	Swab test results at the MBSA Hall Section 28 on 27022021 did not go out in the apps to miss the flight to Kota Kinabalu this morning 28022021 unfortunately
After Direct Preprocessing 2	['swab', 'test', 'result', 'mbsa', 'hall', 'section', '28', '27022021', 'go', 'miss', 'flight', 'kota', 'kinabalu', 'morning', '28022021', 'unfortunately']
POS Filtering Path	[('swab', 'JJ'),('test', 'NN'), ('result', 'NN'), ('mbsa', 'NN'), ('hall', 'NN'), ('section', 'NN'), ('28', 'CD'), ('27022021', 'CD'), ('go', 'VB'), ('miss', 'VB'), ('flight', 'NN'), ('kota', 'VB'), ('kinabalu', 'VB'), ('morning', 'NN'), ('28022021', 'CD'), ('unfortunately', 'RB')]

B. LDA Results

Table 4.4-4 presents the results of applying LDA to user reviews, focusing on the matching percentage between automatically generated topics and function names. The table includes results with preprocessing applied to the text data.

Unlike BERTopic, LDA achieved a low matching percentage (only 28%) between the automatically generated topics and the function names, regardless of preprocessing. This suggests that LDA might not be well-suited for this specific task.

Table 4.4-4 LDA Results

Preprocessing	Total Rows	No. of Selected Topics	Total Topics Generated by LDA (manual)	No. of Topics Match with Function Names	Matched Percentage
		7	11	2	28%
Yes	1433	7	11	2	28%

4.5 Classification Results

This section evaluates the model performance on first layer categories and second layer categories.

4.5.1 Data Preparation

Both the first layer category dataset and the second layer category dataset exhibited significant class imbalance. The number of reviews in each category varied greatly:

- First Layer Categories (Figure 4.5-1): Account (474), Finance (30), Health (170), Others (89)
- Second Layer (Subcategories of Account Functions) (Figure 4.5-3): Apk download (35), Data Management (13), Loading and Performance (119), Sign in (197), Sign up (111)

To address this imbalance and ensure all categories contribute equally during model training, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data for both datasets. SMOTE generates synthetic examples for the minority classes, resulting in a balanced distribution where each class contains an equal number of samples (474 for first layer, 197 for second layer).

As shown below in Figure 4.5-1 to 4.5-4, the bar chart illustrates the imbalanced class distribution of app review categories before and after applying SMOTE.

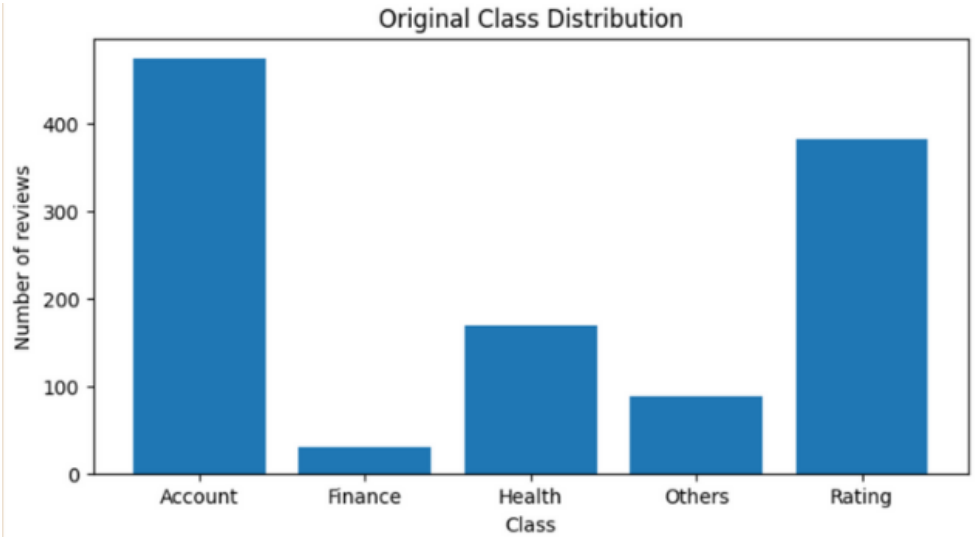


Figure 4.5-1 Bar Chart - Class Distribution of App Review Categories Before SMOTE (First Layer)

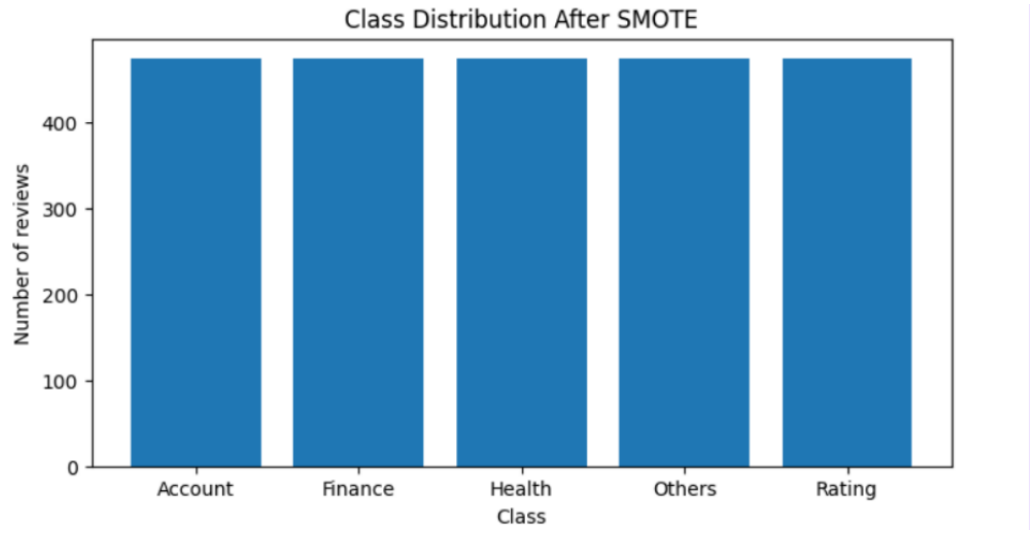


Figure 4.5-2 Bar Chart - Class Distribution of App Review Categories After SMOTE (First Layer)

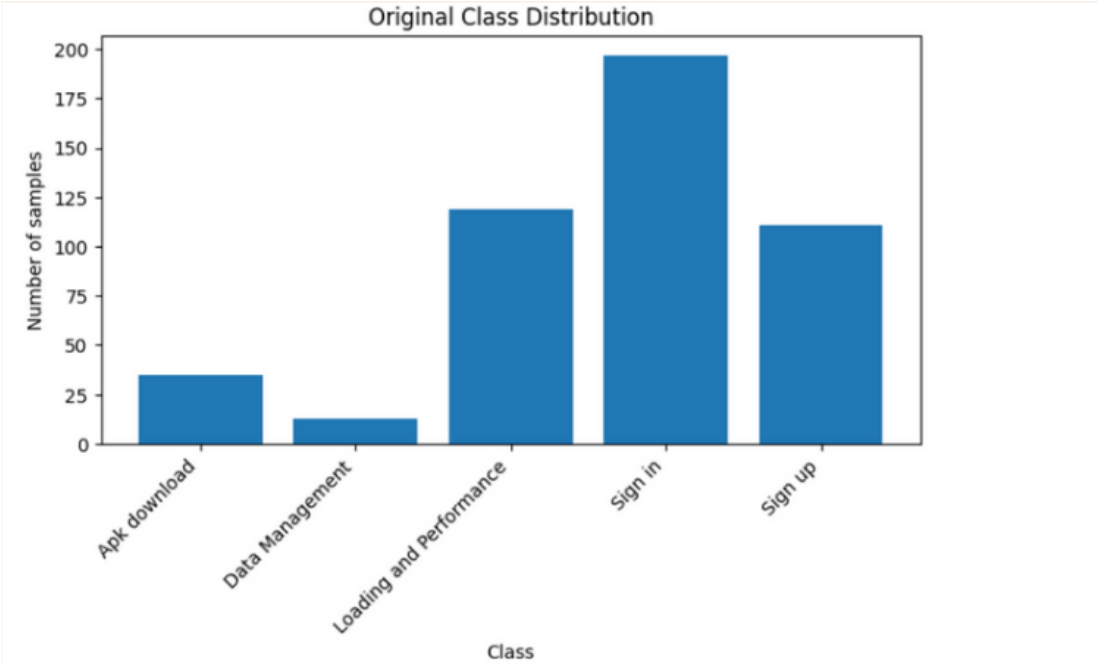


Figure 4.5-3 Bar Chart - Class Distribution of App Usage Categories Before SMOTE (Second Layer)

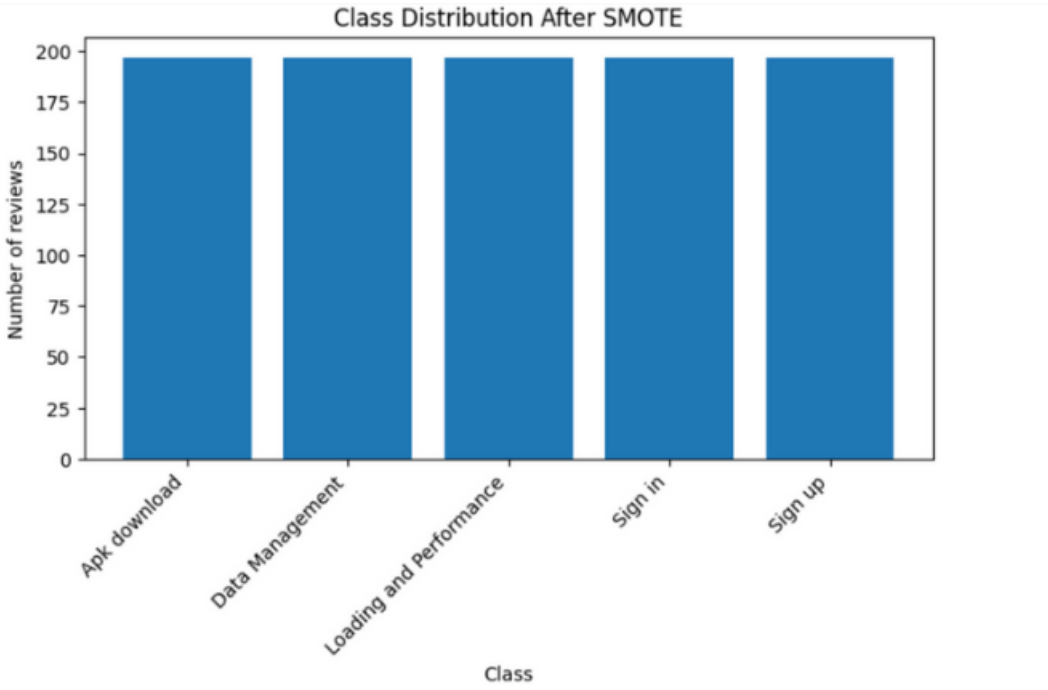


Figure 4.5-4 Bar Chart - Class Distribution of App Usage Categories After SMOTE (Second Layer)

4.5.2 Classification

Table 4.5-1 compares the performance of various classification models for a multi-class text classification task. When trained on balanced data, Multilingual Embedding + BiLSTM achieved a significant accuracy improvement of nearly 19% (from 76.66% to 93.10%) compared to imbalanced data. Other models like Multilingual Embedding + GRU (92.26%) also showed substantial improvement on balanced datasets.

Table 4.5-1 Classification Model Performance Comparison (First Layer)

Method	Data Type	Accuracy (Before Translation)	Accuracy (After Translation)
BERT Embedding + Logistic Regression	Imbalance	-	75%
	Balance	-	75%
BERT Embedding + Random Forest	Imbalance	-	70%
	Balance	-	69%
BERT Embedding + XGBOOST	Imbalance	-	75%
	Balance	-	71%
BERT Embedding + GRU	Imbalance	-	75.26%
	Balance	-	87.71%
BERT Embedding + BiLSTM	Imbalance	-	74.56%
	Balance	-	86.70%
Multilingual Embedding + Logistic Regression	Imbalance	78%	78%
	Balance	74%	77%
Multilingual Embedding + Random Forest	Imbalance	75%	79%
	Balance	76%	77%
Multilingual Embedding +XGBOOST	Imbalance	82%	83%
	Balance	79%	81%
Multilingual Embedding +GRU	Imbalance	78.40%	76.66%
	Balance	92.26%	92.09%
Multilingual Embedding + BiLSTM	Imbalance	76.31%	76.66%
	Balance	**93.10%**	**93.10%**
BERT Transformer	Imbalance	-	76%
	Balance	-	75%
Multilingual Transformer	Imbalance	79%	79%
	Balance	75%	78%

Table 4.5-2 compares the performance of various classification models for a text classification task on 'Account' data. Similar to the previous table, Balancing datasets is a powerful technique to improve the accuracy of GRU and LSTM models, especially for tasks with imbalanced data. However, for other algorithms like Random Forest, Logistic Regression, XGBoost, and Transformers, the impact is less pronounced. Multilingual Embedding + GRU achieved the highest accuracy (90.48%) on balanced data, showing a substantial improvement of over 21% compared to imbalanced data (68.91%). Other models like Multilingual Embedding + BiLSTM (90.12%) also performed well on balanced datasets.

Table 4.5-2 Classification Model Performance Comparison (Second Layer- Account)

Method	Data Type	Accuracy (Before Translation)	Accuracy (After Translation)
BERT Embedding + Logistic Regression	Imbalance	-	75%
	Balance	-	75%
BERT Embedding + Random Forest	Imbalance	-	70%
	Balance	-	69%
BERT Embedding + XGBOOST	Imbalance	-	75%
	Balance	-	71%
BERT Embedding + GRU	Imbalance	-	74.49%
	Balance	-	82.30%
BERT Embedding + BiLSTM	Imbalance	-	71.43%
	Balance	-	81.89%
Multilingual Embedding + Logistic Regression	Imbalance	74%	76%
	Balance	71%	74%
Multilingual Embedding + Random Forest	Imbalance	72%	74%
	Balance	74%	74%
Multilingual Embedding +XGBOOST	Imbalance	68%	69%
	Balance	62%	68%
Multilingual Embedding +GRU	Imbalance	68.91%	78.99%
	Balance	**90.48%**	90.12%
Multilingual Embedding + BiLSTM	Imbalance	75.63%	74.79%
	Balance	90.12%	89.71%
BERT Transformer	Imbalance	-	71%
	Balance	-	74%
Multilingual Transformer	Imbalance	77%	82%
	Balance	80%	81%

4.5.3 Screen Displays in Streamlit

Figure 4.5-5 shows the user interface of the Streamlit app designed for review categorization. The interface allows users to input a review text and classify it into predefined categories through a two-layer classification process.

In this example, the input text is "Can't open the apps." Upon clicking the "Classify" button, the app performs the following:

- First Layer Classification: The app classifies the review into one of the main categories. In this instance, the result is "Account".
- Second Layer Classification: Once the review is classified under the "Account" category, the app further classifies it into a more specific subcategory. Here, the result is "Sign in".

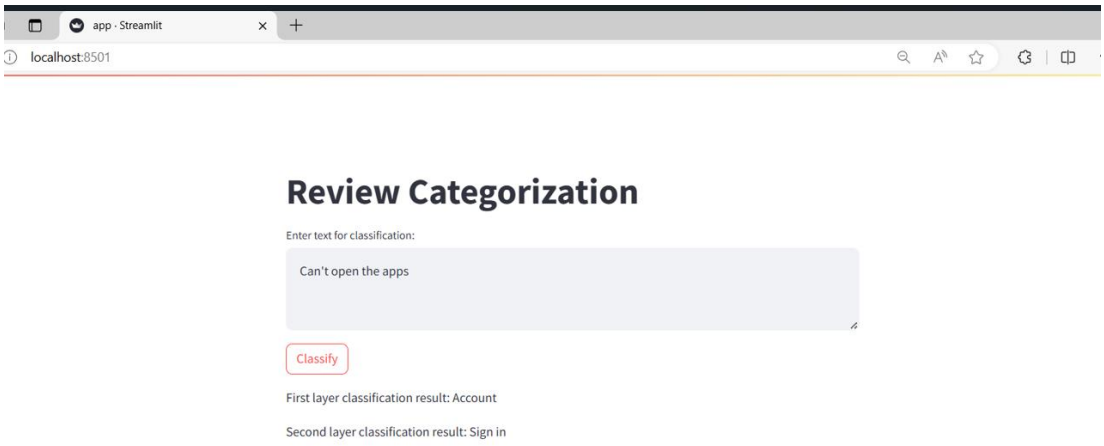


Figure 4.5-5 Streamlit App Interface for Review Categorization

4.6 Discussion

This section delves deeper into the findings from the previous sections, exploring the implications and potential explanations for the observed results.

4.6.1 Topic Modeling - LDA VS BERTopic

From the result, BERTopic outperformed LDA in clustering function names. As shown in the previous section (refer to Table 4.4-1 and Table 4.4-2), BERTopic achieved significantly higher matching percentages between automatically generated topics and actual function names compared to LDA (over 50% vs. 28%).

Word Representation Advantage:

A key reason for BERTopic's superior performance lies in its use of word embeddings. These capture semantic relationships between words, allowing BERTopic to group related function names more effectively. For instance, as shown in Table 4.6-1, in the category of "sign in," BERTopic identified relevant terms like "password," "reset," "forgot," and "login," whereas LDA included irrelevant words like "update," "date," and "work."

Table 4.6-1 Comparison of Word Representations for "Sign In" by BERTopic and LDA

Model	Word Representation
BERTopic	password, reset, forgot, forget, sign, character, say, unable, new, login
LDA	update, password, log, selangkah, date, new, go, work, change, reset

Possible Reasons for Lower LDA Performance:

- Limited Semantic Capture: LDA might struggle to capture the nuances of function names compared to BERTopic's word embedding-based approach.
- Short Text Challenges: Function names are often shorter and less detailed than typical documents LDA is designed for. This can make it difficult for LDA to identify distinct topics within function names.

4.6.2 Classification

This section analyzes the performance of various classification models for the text classification task. The key finding is that multilingual embedding models generally outperform BERT embedding models on balanced data. This trend is observed across different classification algorithms (Logistic Regression, Random Forest, XGBoost, GRU, BiLSTM).

Potential Reasons for Multilingual Embedding Advantage:

- Wider Language Exposure: Multilingual embedding models are trained on a broader range of languages compared to monolingual BERT models. This exposure to diverse linguistic structures might enhance their ability to generalize to unseen data, leading to better performance on balanced datasets.
- Focus on Semantics: Multilingual embeddings aim to capture semantic similarities across languages. This focus on meaning representation could be particularly beneficial for the specific classification task at hand, especially when dealing with balanced data that allows the model to learn the underlying relationships more effectively.

Limitations of Random Forest:

The results suggest that Random Forest performs similarly or even worse compared to other classifiers like XGBoost, GRU, and BiLSTM. This indicates that for this specific classification task of identifying functionalities from Selangkah user reviews, more complex models that can capture non-linear relationships between features (e.g., interactions between functionalities or user descriptions) and the target variable might be more suitable. While Logistic Regression offers interpretability, it might not be the best choice for capturing the intricacies of the text data in this classification problem.

4.6.3 Topic Modeling vs. Classification

A comparative analysis of topic modeling and classification methods reveals their respective strengths and weaknesses in the context of the Selangkah app reviews. Topic modeling methods like BERTopic are useful for identifying hidden patterns and groupings within the data without predefined labels. However, classification models provide more precise predictions when labels are available, making them more suitable for structured tasks like function name categorization.

4.6.4 Limitations and Challenges

A. Data-Related Challenges

- Sparsity and Malay Words: Both topic modeling and classification models struggled with the sparsity of the data, where many features have few occurrences. Additionally, the presence of Malay words, for which the models might not have adequate training data, affected performance. The lack of high-quality, annotated Malay datasets further exacerbated this issue.
- Imbalanced Data: Imbalanced data, where some categories have significantly more examples than others, posed a challenge for both topic modeling and classification. In topic modeling, dominant clusters overshadowed less frequent topics, reducing diversity. For classification, models struggled to learn from underrepresented categories due to limited training data.
- Noise and Irrelevant Information: Reviews often contained noise like typos, slang, and abbreviations. Preprocessing helped, but residual noise still hindered the models' ability to accurately identify and group similar topics for clustering or learn meaningful patterns for classification.

B. Model-Related Challenges

- Multilingual Data Handling: The multilingual nature of the data (Malay, English, Chinese) introduced complexity for both topic modeling and classification. Models might struggle to handle the variations in language structure and vocabulary, leading to issues with creating consistent embeddings and achieving optimal performance.
- Classification Model Complexity: Certain classification models, particularly Logistic Regression, were not well-suited for the complexities of the text data. They struggled to capture non-linear relationships and contextual nuances compared to more advanced models like GRU and BiLSTM, leading to lower accuracy.
- Limited Training Data for Some Categories: Similar to the data imbalance issue, categories with limited training data affected both topic modeling and

classification. With fewer examples, models couldn't learn effectively, resulting in poorer performance for those underrepresented categories.

Chapter 5: Conclusion

This project aimed to analyze and improve the understanding of user reviews for the Selangkah app through advanced data analysis and machine learning techniques. The project successfully achieved its research objectives, providing valuable insights and contributing to the field of text analysis and NLP.

Accomplishments of Research Objectives:

1. EDA:

The project conducted a thorough EDA, revealing key characteristics and trends in the user review data. By comparing the data before and after preprocessing, the analysis highlighted the impact of cleaning and preprocessing steps on data quality and insights.

2. Text Preprocessing and Handling Multilingual Data:

Comprehensive preprocessing steps were implemented to clean the text data, handle abbreviations, manage out-of-vocabulary words, and translate multilingual reviews. This ensured the data was prepared for subsequent analysis and modeling, effectively dealing with challenges posed by the presence of Malay, English, and Chinese text.

3. Topic Modeling:

The project utilized both LDA and BERTopic for topic modeling, with BERTopic outperforming LDA in clustering function names. This demonstrated the effectiveness of word embeddings in capturing semantic relationships and providing more accurate and meaningful topics.

4. Classification:

Various classification models were tested, revealing that multilingual embedding models generally outperformed BERT embedding models on balanced data. The study highlighted the advantages of multilingual embeddings in capturing semantic similarities across languages, which improved classification performance.

Major Learnings:

1. Importance of Preprocessing:

Effective text preprocessing is crucial for improving the quality of data and the performance of machine learning models. Handling multilingual data and domain-specific terms requires tailored preprocessing techniques.

2. Advantages of Advanced Embedding Techniques:

Word embeddings, especially those designed for multilingual contexts, significantly enhance the ability to capture semantic relationships and improve the performance of topic modeling and classification tasks.

3. Challenges with Imbalanced and Noisy Data:

Data imbalance and noise present significant challenges in topic modeling and classification. Addressing these issues requires advanced techniques and careful consideration of model selection and data handling strategies.

Future Work:

Future research can build on this project by exploring advanced AI models like ChatGPT for more sophisticated text analysis and interpretation. Improvements in data preprocessing techniques, particularly for handling low-resource languages like Malay, will be essential. Additionally, leveraging domain-specific language models and transfer learning can further enhance the accuracy and reliability of the analysis.

Addressing data imbalance through synthetic data generation, advanced sampling techniques, and exploring more sophisticated embedding methods will also be beneficial. Finally, refining classification algorithms and investigating new topic modeling approaches can provide deeper insights and improve the overall effectiveness of text analysis in multilingual and noisy data environments.

In conclusion, this project has successfully achieved its research objectives, providing a comprehensive analysis of user reviews for the Selangkah app. The findings highlight the importance of effective preprocessing, the advantages of advanced embedding techniques, and the challenges posed by multilingual and imbalanced data. The insights and methodologies developed in this project lay a solid foundation for future work, promising further advancements in the field of text analysis and natural language processing.

References

Agarwal, J., Christa, S., Pai, H. A., Kumar, M. A., & Prasad, M. S. G. (2023). Machine Learning Application for News Text Classification. *Proceedings of the 13th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2023*, 463–466. https://doi.org/10.1109/Confluence56041.2023.10048856

Agarwal, N., Sikka, G., & Awasthi, L. K. (2020). Evaluation of web service clustering using Dirichlet Multinomial Mixture model based approach for Dimensionality Reduction in service representation. *Information Processing and Management*, *57*(4). https://doi.org/10.1016/j.ipm.2020.102238

Amara, A., Hadj Taieb, M. A., & Ben Aouicha, M. (2021). Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. *Applied Intelligence*, *51*(5), 3052–3073. https://doi.org/10.1007/s10489-020-02033-3

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In *Journal of Machine Learning Research* (Vol. 3).

del Gobbo, E., Fontanella, S., Sarra, A., & Fontanella, L. (2021). Emerging Topics in Brexit Debate on Twitter Around the Deadlines: A Probabilistic Topic Modelling Approach. *Social Indicators Research*, *156*(2–3), 669–688. https://doi.org/10.1007/s11205-020-02442-4

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018a). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018b). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Dharrao, D., Deokate, S., Bongale, A. M., & Urolagin, S. (2023). E-Commerce Product Review Classification based on Supervised Machine Learning Techniques. *2023 9th International Conference on Advanced Computing and Communication Systems, ICACCS 2023*, 1934–1939. https://doi.org/10.1109/ICACCS57279.2023.10112717

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*. https://doi.org/10.3389/fsoc.2022.886498

Gao, C., Zheng, W., Deng, Y., Lo, D., Zeng, J., Lyu, M. R., & King, I. (2019). Emerging App Issue Identification from User Feedback: Experience on WeChat. *Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2019*, 279–288. https://doi.org/10.1109/ICSE-SEIP.2019.00040

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing*, *101*. https://doi.org/10.1016/j.asoc.2020.107057

George, L., & Sumathy, P. (2023). An integrated clustering and BERT framework for improved topic modeling. *International Journal of Information Technology (Singapore)*, *15*(4), 2187–2195. https://doi.org/10.1007/s41870-023-01268-w

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. http://arxiv.org/abs/2203.05794

Guerzo, L. A. B., Kilkenny, H. A. O., Osorio, R. N. D., Villegas, A. H. E., & Ponay, C. S. (2021). Topic Modelling and Clustering of Disaster-Related Tweets using Bilingual Latent Dirichlet Allocation and Incremental Clustering Algorithm with Support Vector Machines for Need Assessment. *Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021*, 189–193. https://doi.org/10.1109/ICSECS52883.2021.00041

Hendry, D., Darari, F., Nurfadillah, R., Khanna, G., Sun, M., Condylis, P. C., & Taufik, N. (2021). Topic Modeling for Customer Service Chats. *2021 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2021*. https://doi.org/10.1109/ICACSIS53237.2021.9631322

Hristova, G., & Netov, N. (2022). Media Coverage and Public Perception of Distance Learning During the COVID-19 Pandemic: A Topic Modeling Approach Based on BERTopic. *Proceedings - 2022 IEEE International*

Conference on Big Data, Big Data 2022, 2259–2264. https://doi.org/10.1109/BigData55660.2022.10020466

Jeon, E., Yoon, N., & Sohn, S. Y. (2023). Exploring new digital therapeutics technologies for psychiatric disorders using BERTopic and PatentSBERTa. *Technological Forecasting and Social Change*, *186*. https://doi.org/10.1016/j.techfore.2022.122130

Li, X., Wang, Y., Ouyang, J., & Wang, M. (2021). Topic extraction from extremely short texts with variational manifold regularization. *Machine Learning*, *110*(5), 1029–1066. https://doi.org/10.1007/s10994-021-05962-3

Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C., & Kyriazis, D. (2023). Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. *Neural Computing and Applications*, *35*(29), 21415–21431. https://doi.org/10.1007/s00521-023-08629-3

Mouthami, K., Yuvaraj, N., Thilaheswaran, K. K., & Lokeshvar, K. J. (2023). Text Sentiment Analysis of Film Reviews Using Bi-LSTM and GRU. *2023 4th International Conference on Electronics and Sustainable Communication Systems, ICESC 2023 - Proceedings*, 1379–1386. https://doi.org/10.1109/ICESC57686.2023.10193121

Mutiah, N., Prawira, D., & Rusi, I. (2022). Topic Modeling on Covid-19 Vaccination in Indonesia Using LDA Model. *2022 7th International Conference on Informatics and Computing, ICIC 2022*. https://doi.org/10.1109/ICIC56845.2022.10007005

Ng, L. H. X., & Carley, K. M. (2021). "The coronavirus is a bioweapon": classifying coronavirus stories on fact-checking sites. *Computational and Mathematical Organization Theory*, *27*(2), 179–194. https://doi.org/10.1007/s10588-021-09329-w

Parlina, A., & Maryati, I. (2023). Leveraging BERTopic for the Analysis of Scientific Papers on Seaweed. *Proceedings - 2023 10th International Conference on Computer, Control, Informatics and Its Applications: Exploring the Power of Data: Leveraging Information to Drive Digital Innovation, IC3INA 2023*, 279–283. https://doi.org/10.1109/IC3INA60834.2023.10285737

Ramos, J. (n.d.). *Using TF-IDF to Determine Word Relevance in Document Queries*.

Rashid, J., Shah, S. M. A., & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. *Information Processing and Management*, *56*(6). https://doi.org/10.1016/j.ipm.2019.102060

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Terragni, S., Fersini, E., Galuzzi, B., Tropeano, P., & Candelieri, A. (n.d.). *OCTIS: Comparing and Optimizing Topic Models is Simple!* http://people.csail.mit.edu/jrennie/

Tounsi, A., Elkefi, S., & Bhar, S. L. (2023). Exploring the Reactions of Early Users of ChatGPT to the Tool using Twitter Data: Sentiment and Topic Analyses. *Proceedings of the 2023 IEEE International Conference on Advanced Systems and Emergent Technologies, IC_ASET 2023*. https://doi.org/10.1109/IC_ASET58101.2023.10150870

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (n.d.). *Llama 2: Open Foundation and Fine-Tuned Chat Models*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. http://arxiv.org/abs/2307.09288

Udupa, A., Adarsh, K. N., Aravinda, A., Godihal, N. H., & Kayarvizhy, N. (2022). An Exploratory Analysis of GSDMM and BERTopic on Short Text Topic Modelling. *2022 4th International Conference on Cognitive Computing and Information Processing, CCIP 2022*. https://doi.org/10.1109/CCIP57447.2022.10058687

Vallurupalli, V., & Bose, I. (n.d.). *Exploring thematic composition of online reviews: A topic modeling approach*. https://doi.org/10.1007/s12525-020-00397-5/Published

Vasudeva Raju, S., Kumar Bolla, B., Nayak, D. K., & Jyothsna, K. H. (2022). Topic Modelling on Consumer Financial Protection Bureau Data: An Approach Using BERT Based Embeddings. *2022 IEEE 7th International Conference for Convergence in Technology, I2CT 2022*. https://doi.org/10.1109/I2CT54291.2022.9824873

Wang, X., Zhang, W., Lai, S., Ye, C., & Zhou, H. (2022). The Use of Pretrained Model for Matching App Reviews and Bug Reports. *IEEE International Conference on Software Quality, Reliability and Security, QRS, 2022-December*, 242–251. https://doi.org/10.1109/QRS57517.2022.00034

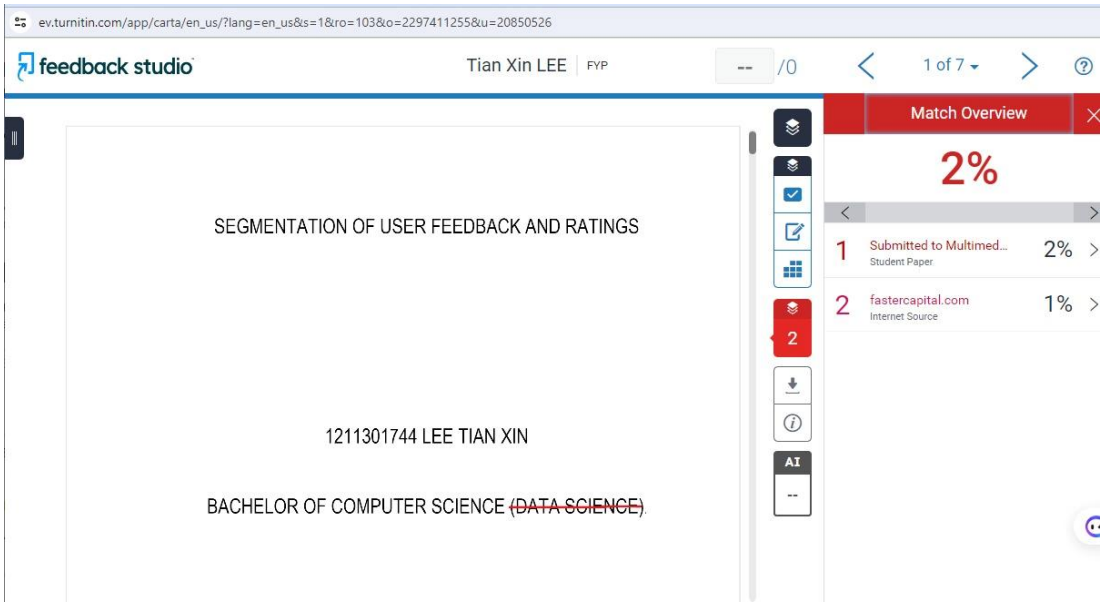
Zhou, W., Wang, Y., Gao, C., & Yang, F. (2022). Emerging topic identification from app reviews via adaptive online biterm topic modeling. *Frontiers of Information Technology and Electronic Engineering*, *23*(5), 678–691. https://doi.org/10.1631/FITEE.2100465

Appendices

Appendix A: Research Paper

The research paper will be attached at the end of this report.

Appendix B: Turnitin Similarity Index Page



Appendix C: Meeting Logs

1211301744_FYP2_Meeting_Log1.pdf

1211301744_FYP2_Meeting_Log2.pdf

1211301744_FYP2_Meeting_Log3.pdf

1211301744_FYP2_Meeting_Log4.pdf

1211301744_FYP2_Meeting_Log5.pdf

1211301744_FYP2_Meeting_Log6.pdf

Segmentation of User Feedback and Ratings

Lee Tian Xin
Faculty of Computing and Informatics
Multimedia University
Selangor, Malaysia
1211301744@student.mmu.edu.my

Goh Hui Ngo
Faculty of Computing and Informatics
Multimedia University
Selangor, Malaysia
hngoh@mmu.edu.my

***Abstract*— The ubiquity of mobile applications in the digital age has underscored the critical role of user feedback in guiding app development and enhancing user satisfaction. This paper presents a comprehensive approach to analyzing user feedback through a combination of unsupervised and supervised learning techniques, with a focus on the Selangkah app—a multifunctional platform offering diverse services including health, welfare, and user management. Initially, unsupervised learning techniques, such as BERTopic, were employed for topic modeling to categorize user feedback. However, due to limitations in providing clear labels or insights into specific topics, the methodology transitioned to a supervised learning approach. By manually labeling a subset of the data and training classifiers on this labeled dataset, the study aims to efficiently categorize user feedback to identify areas for improvement and understand user sentiment across different functionalities. The results demonstrate the efficacy of this hybrid approach in extracting actionable insights from user reviews, thereby guiding targeted enhancements to the Selangkah app.**

Keywords— natural language processing, topic modeling, text classification

I. INTRODUCTION

In the rapidly evolving landscape of digital technology, mobile applications have become integral to user engagement and satisfaction. The importance of timely and effective analysis of user feedback is highlighted by instances where apps faced significant backlash due to unaddressed issues or undesirable updates, as seen in cases like Pokémon GO and Gray Raven Punishing (Gao et al., 2019; Wang et al., 2022). Such situations can be alleviated with timely problem-solving. Therefore, this project aims to rapidly understand user feedback to enhance the overall user experience.

Despite the wealth of data available through user reviews and ratings, extracting meaningful insights remains a challenge due to the volume, diversity, and brevity of the feedback. Short text reviews often contain noise, sparsity, and lack context, making it difficult to parse and interpret manually. Traditional manual analysis methods are not only labor-intensive but also prone to inaccuracies. This study seeks to address these challenges by applying a combination of unsupervised and supervised learning techniques to automate the analysis of user feedback on the Selangkah app, focusing on its distinct functionalities: Account, Health, Finance, Sign up, and Sign in.

The research questions guiding this study are:

- What themes emerge from the preliminary analysis of user feedback on the Selangkah app?
- How can unsupervised and supervised learning techniques be effectively utilized to extract key phrases and words related to the app's functionalities?

The objectives of this research project are:

- To perform exploratory data analysis to uncover themes within the Selangkah user feedback dataset.
- To identify key phrases and words associated with the app's specific functionalities using a combination of unsupervised and supervised learning techniques, enhancing the understanding of user sentiment and feedback.

The scope of this project is discovering themes in the Selangkah dataset, which only focuses on non-null data, comprising around 2000 rows. It focuses on Malay, Chinese, and English user reviews.

II. RELATED WORK

The use of topic modeling and classification techniques has seen substantial advancements in recent years, significantly enhancing the analysis of textual data across various domains. Topic modeling, a statistical model for discovering abstract topics within a collection of documents, is widely employed in text mining and NLP to organize, understand, and summarize large datasets. Traditional methods such as Latent Dirichlet Allocation (LDA) and Dirichlet Multinomial Mixture (DMM) have been popular for their ability to uncover thematic structures within text. LDA, for example, views each document as a mixture of topics and has been applied to user reviews and COVID-19 data, demonstrating its versatility in capturing topic evolution and public discourse themes (Amara et al., 2021; Zhou et al., 2022). However, these methods face challenges with short texts' sparseness, where models like DMM have shown better sensitivity to noisy words (Agarwal et al., 2020).

The advent of Large Language Models (LLMs) such as BERT and Llama2 has marked a significant leap in topic modeling. These models excel in understanding contextual nuances, overcoming the limitations of traditional methods like LDA, which often fail to capture semantic relationships and topic (Udupa et al., 2022; Vasudeva Raju et al., 2022). BERT-based models, such as BERTopic, preserve the original text structure while effectively inferring topics, proving superior in various contexts, from customer service chats to analyzing user interactions on online platforms (Egger & Yu, 2022; Tounsi et al., 2023)

In the realm of text classification, pretrained models like T5, Sentence T5, Sentence MiniLM, and Sentence BERT have been instrumental in tasks such as matching app reviews and bug reports, with Sentence T5 outperforming other models(Wang et al., 2022). Comparative analyses of multilingual approaches, particularly using XLM-R, have

shown superior performance in classifying Twitter data compared to other BERT-based classifiers (Manias et al., 2023). Furthermore, sentiment analysis of film reviews using word embeddings like Word2Vec and FastText, combined with models like Bi-LSTM, has demonstrated significant improvements in accuracy (Mouthami et al., 2023).

The integration of clustering and classification techniques has further enriched the analysis of complex datasets. For example, the combination of K-Means clustering with topic modeling has been used to enhance the understanding of user reviews and social media content by grouping similar topics before applying classification techniques (Garcia & Berton, 2021; Guerzo et al., 2021).This approach helps in identifying patterns and improving the accuracy of classifiers by providing more structured input data. Future research directions may involve integrating more advanced models, improving semi-supervised learning methods, and tailoring deep learning architectures to further enhance text analysis techniques and provide deeper insights into user-generated content.

These advancements collectively illustrate the evolution of topic modeling and text classification methodologies, blending the strengths of traditional statistical methods with modern neural network-based models to achieve higher accuracy and reliability. Future research directions may involve integrating more advanced models, improving semi-supervised learning methods, and tailoring deep learning architectures to further enhance text analysis techniques and provide deeper insights into user-generated content.

III. METHODOLOGY

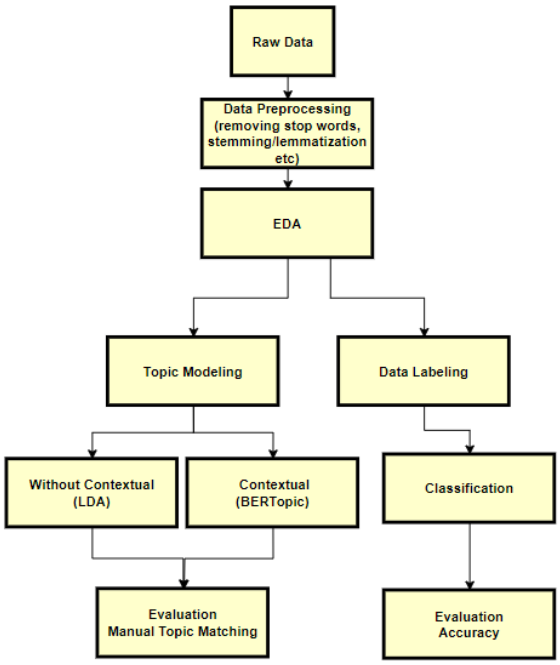


Fig. 1. Methodology

The methodology (Fig. 1.) of this research comprises several key steps, each designed to analyze user feedback from the Selangkah app efficiently. The overall process is divided into two main branches: Topic Modeling and Classification.

A. Data Preprocessing

Data preprocessing is a critical initial step in ensuring the quality and consistency of the dataset, especially given the multilingual nature of the user reviews which include Malay, Chinese, and English. This process involves several key stages:

- Remove null rows: The dataset used in this study comprises user reviews from the Selangkah app, available in Malay, Chinese, and English. Initially, the dataset contained approximately 6000 reviews. After removing null entries and irrelevant data, the dataset was reduced to around 2000 reviews.
- Text Cleaning: Removal of punctuation, numbers, and special characters.
- Abbreviation Handling: Expansion of common abbreviations.
- Out-of-Vocabulary Cleaning: Replacement of rare words with a placeholder.
- Language Detection and Translation: Identification of the language of each review and translation into a common language (English) for uniform processing.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is conducted to gain preliminary insights into the dataset and inform subsequent analysis steps. This involves:

- Distribution of Ratings: Examining the distribution of user ratings from 1 to 5 stars.
- Review Length Analysis: Analyzing the length of reviews in terms of the number of words and characters.
- Word Cloud Generation: Visualizing the most frequent words in the dataset.
- Language Distribution: Identifying the proportion of reviews in each language.

C. Topic Modeling

Unsupervised learning techniques were employed to discover underlying themes in user feedback:

- Latent Dirichlet Allocation (LDA): LDA is chosen for its established effectiveness in identifying sets of words that frequently co-occur. Parameter tuning is conducted to optimize the number of topics, alpha, and beta values, enhancing the model's performance.
- BERTopic(Grootendorst, 2022): BERTopic is used for its ability to leverage contextual embeddings, allowing for a more nuanced identification of topics based on semantic meanings. Similar to LDA, parameter tuning is performed to adjust the number of topics and embedding dimensions.

D. Classification

For categorizing user reviews, various classification methods are applied to a manually labeled subset of the dataset. This process begins with the manual labeling of a subset of reviews, focusing on predefined categories such as "Account," "Health," "Finance," "Sign up," and "Sign in."

The labeled data undergoes preprocessing to ensure consistency.

Several classification algorithms are trained on this preprocessed data. Logistic Regression serves as a simple yet effective baseline classifier. Random Forest, an ensemble method combining multiple decision trees, and XGBoost, a gradient boosting algorithm, are employed for their performance and speed. Additionally, recurrent neural network (RNN) variants like GRU and BiLSTM, which are suitable for handling sequential data, are also utilized. To address potential imbalances in the labeled data, SMOTE (Synthetic Minority Over-sampling Technique) have been employed to create synthetic samples for under-represented review categories, leading to a more balanced dataset for training the classification models.

This methodology combines unsupervised and supervised learning techniques to provide a thorough analysis of user feedback, extracting actionable insights to guide improvements for the Selangkah app. The structured approach ensures a detailed understanding of user sentiment and feedback across the app's various functionalities.

E. Evaluation

The performance of the topic modeling and classification models is evaluated using various metrics:

- Topic Modeling Evaluation: The topics identified by LDA and BERTopic are evaluated through manual topic matching. This involves assessing the coherence and relevance of the topics, ensuring they align with the expected themes and categories.
- Classification Evaluation: The classification models are evaluated based on their accuracy in predicting the correct category for each review.

IV. RESULTS AND DISCUSSION

A. EDA

A This initial exploration of the Selangkah app user review dataset revealed several interesting trends. The review volume shows a decline over time, with the highest number in 2021 and a significant drop in 2023 (Fig. 3.). This could suggest a decrease in user engagement or fewer app issues requiring feedback. Review length analysis indicates most reviews are concise, with a significant portion consisting of just one word. Interestingly, lower-rated reviews tend to be longer, suggesting users express frustrations in more detail.

The language distribution (Fig. 2.) reveals a mix of Malay, English, and Chinese reviews, with English being the most common. The word cloud analysis (Fig. 5.) provides valuable insights into user sentiment. Words like "app," "update," "time," and "vaccine" are frequent across all ratings. Analyzing word categories highlights the presence of nouns ("update," "app"), adjectives ("good," "bad"), verbs ("register," "open"), and adverbs ("already," "still") (Fig. 6.). Examining word frequency by rating paints a clearer picture. Lower ratings feature words like "update," "register," "open," and "vaccine" alongside adverbs like "already" and "still," suggesting user frustration. Conversely, higher ratings showcase words like "easy" and "helpful," indicating positive experiences(Fig. 7.) (Fig. 8.).

A bar chart analysis further categorizes the reviews into primary and secondary layers. The first layer consists of categories such as Account (474), Finance (30), Health (170), and Others (89) (Fig. 9). The second layer delves deeper into the subcategories of the Account function, including Apk download (35), Data Management (13), Loading and Performance (119), Sign in (197), and Sign up (111) (Fig. 10). This layered categorization helps in identifying specific areas where users encounter issues or express satisfaction.

Overall, the EDA findings offer a valuable starting point for understanding user sentiment and identifying areas for improvement in the Selangkah app.

```
Number of Malay reviews: 855
Number of English reviews: 1098
Number of Chinese reviews: 25
Number of Tamil reviews: 0
```

Fig. 2. Language Distribution

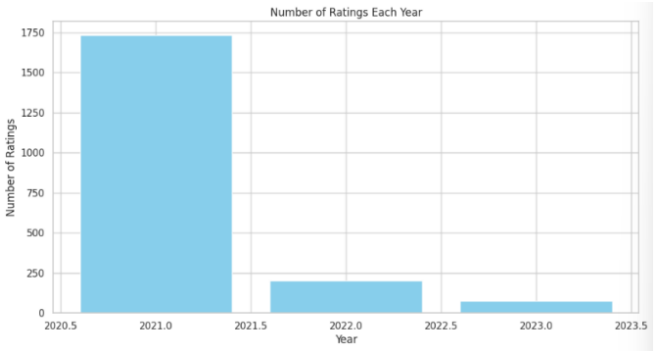


Fig. 3. Number of Ratings Each Year After Removing Null Rows

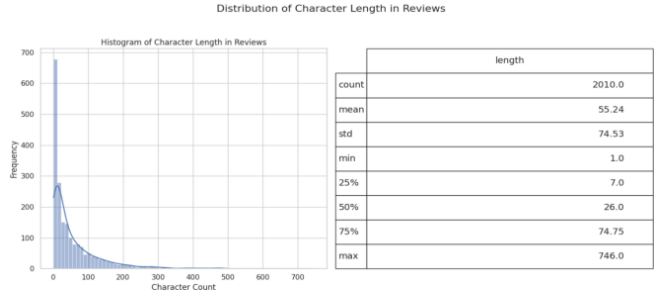


Fig. 4. Distribution of Character Length in All Reviews

Fig. 5. Word Cloud for All User Reviews after Initial Preprocessing (after Initial Preprocessing)

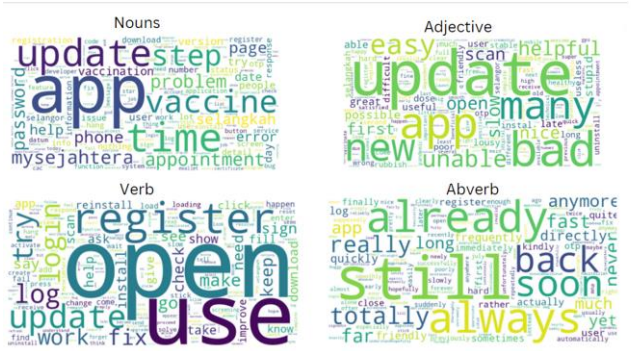


Fig. 6. POS Tagging Word Cloud for All User Reviews (after Initial Preprocessing)

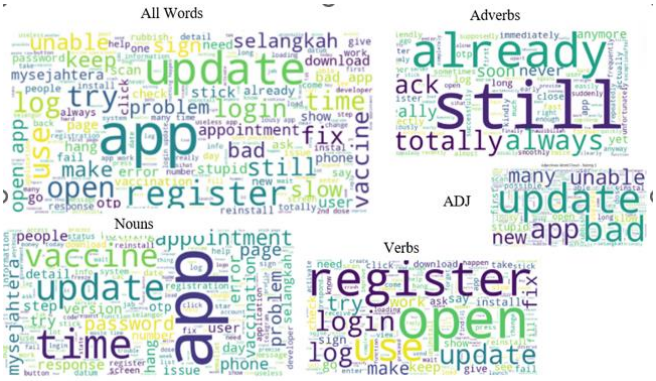


Fig. 7. Word Clouds for Rating 1 Reviews (after Initial Preprocessing)

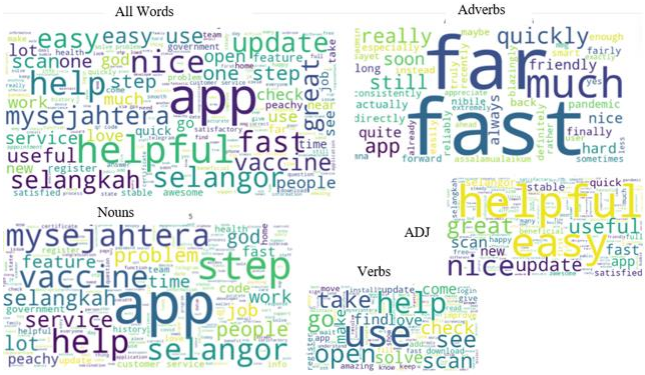


Fig. 8. Word Clouds for Rating 5 Reviews (after Initial Preprocessing)

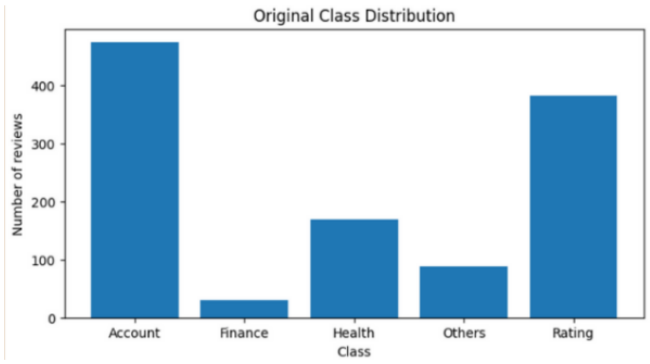


Fig. 9. Class Distribution of App Review Categories Before SMOTE (First Layer)

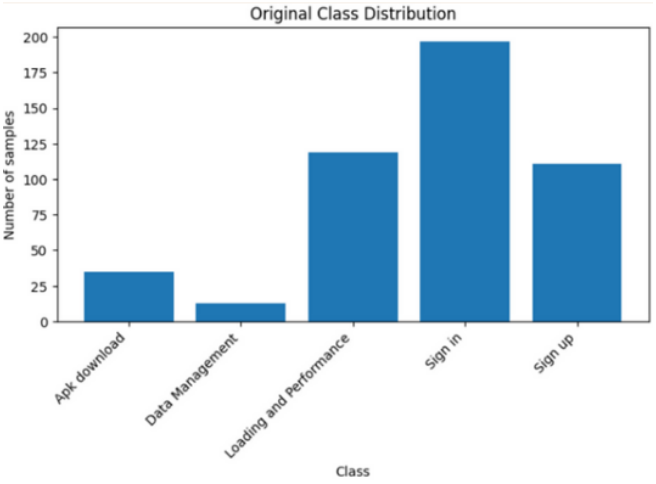


Fig. 10. Bar Chart - Class Distribution of App Usage Categories Before SMOTE (Second Layer)Topic Modeling Results

B. *Topic Modeling Results*

This analysis explored the effectiveness of BERTopic and LDA for topic modeling in Selangkah app user reviews. Hyperparameter tuning was applied to BERTopic to evaluate different model configurations and identify the optimal number of topics. Based on this tuning, BERTopic's performance was compared to LDA with a fixed number of 11 topics. The results revealed that BERTopic, even with hypertuning at various topic counts (40 and 33, 20 and 19), consistently identified more relevant topics, particularly those related to function names. In contrast, LDA with 11 topics struggled to capture the same level of detail and missed several function-related themes present in the reviews. This highlights the advantage of BERTopic's word embedding approach, which effectively captures semantic relationships and leads to a more accurate representation of user discussions, especially for specific app functionalities.

TABLE I. TOPIC MODELING RESULTS

Model	Preprocessing	***Total Rows***	***Total Topics generated by Model***	***No of Topics Match with Function Names***
BERTopic	No	2000	40 (auto)	5 (63%)
			20	4 (57%)
	Yes	1433	33 (auto)	6 (85%)
		1433	**19**	**7 (100%)**
		1433	11	4 (57%)
LDA	Yes	1433	11	2(28%)

C. *Classification Results*

This work investigates the effectiveness of various classification models for sentiment analysis of Selangkah app user reviews. A multi-class classification approach is employed to categorize reviews into classes. A crucial aspect was addressing data imbalance, where some sentiment categories might be overrepresented. To mitigate this, model performance was evaluated on both imbalanced and balanced datasets achieved through oversampling or undersampling techniques.

Tables II and III present the accuracy scores of different models under these data conditions. Balancing datasets is a powerful technique to improve the accuracy of GRU and LSTM models, especially for tasks with imbalanced data. However, for other algorithms like Random Forest, Logistic Regression, XGBoost, and Transformers, the impact is less pronounced.

In Table II, Multilingual Embedding + BiLSTM achieved the highest accuracy (93.10%) on the balanced dataset, a significant improvement of nearly 19% compared to its imbalanced data performance (76.66%). This suggests that combining multilingual embeddings with a BiLSTM architecture effectively addresses the challenges of sentiment classification in a multilingual context.

Similarly, in Table III (Account data), Multilingual Embedding + GRU achieved the highest accuracy (90.48%) on the balanced dataset, showcasing a substantial improvement of over 21% compared to imbalanced data (68.91%). Other models like Multilingual Embedding + BiLSTM also performed well (90.12%) on balanced datasets for Account data.

These findings highlight the effectiveness of multilingual embeddings, particularly when combined with architectures like BiLSTM and GRU, for sentiment classification in multilingual environments.

TABLE II. CLASSIFICATION MODEL PERFORMANCE COMPARISON (FIRST LAYER)

Method	*Data Type*	*Accuracy (Before Translation) (%)*	*Accuracy (After Translation) (%)*
BERT Embedding + Logistic	Imbalance	-	75
	Balance	-	75
BERT Embedding + Random Forest	Imbalance	-	70
	Balance	-	69
BERT Embedding + XGBOOST	Imbalance	-	75
	Balance	-	71
BERT Embedding + GRU	Imbalance	-	75.26
	Balance	-	87.71
BERT Embedding + BiLSTM	Imbalance	-	74.56
	Balance	-	86.70
Multilingual Embedding + Logistic	Imbalance	78	78
	Balance	74	77
Multilingual Embedding + Random	Imbalance	75	79
	Balance	76	77
Multilingual Embedding +XGBOOST	Imbalance	82	83
	Balance	79	81
Multilingual Embedding +GRU	Imbalance	78.4	76.66
	Balance	92.26	92.09
Multilingual Embedding + BiLSTM	Imbalance	76.31	76.66
	Balance	**93.10**	**93.10**
BERT Transformer	Imbalance	-	76
	Balance	-	75
Multilingual Transformer	Imbalance	79	79
	Balance	75	78

TABLE III. CLASSIFICATION MODEL PERFORMANCE COMPARISON (SECOND LAYER- ACCOUNT)

Method	*Data Type*	*Accuracy (Before Translation) (%)*	*Accuracy (After Translation) (%)*
BERT Embedding + Logistic	Imbalance	-	75
	Balance	-	75
BERT Embedding + Random Forest	Imbalance	-	70
	Balance	-	69
BERT Embedding + XGBOOST	Imbalance	-	75
	Balance	-	71
BERT Embedding + GRU	Imbalance	-	74.49
	Balance	-	82.30
BERT Embedding + BiLSTM	Imbalance	-	71.43
	Balance	-	81.89
Multilingual Embedding + Logistic	Imbalance	74	76
	Balance	71	74
Multilingual Embedding + Random	Imbalance	72	74
	Balance	74	74
Multilingual Embedding +XGBOOST	Imbalance	68	69
	Balance	62	68
Multilingual Embedding +GRU	Imbalance	68.91	78.99
	Balance	**90.48**	90.12
Multilingual Embedding + BiLSTM	Imbalance	75.63	74.79
	Balance	90.12	89.71
BERT Transformer	Imbalance	-	71
	Balance	-	74
Multilingual Transformer	Imbalance	77	82
	Balance	80	81

V. CONCLUSION

This project leveraged advanced data analysis and machine learning (ML) techniques to unlock user insights from Selangkah app reviews. Exploratory Data Analysis (EDA) revealed key trends, and comprehensive text preprocessing tackled multilingual data challenges. Both traditional (LDA) and advanced (BERTopic) topic modeling were utilized, with BERTopic outperforming LDA in clustering function names due to its effective use of word embeddings for capturing semantic relationships. Classification models were explored, showcasing the

advantages of multilingual embedding models for balanced datasets.

These findings highlight the crucial role of effective preprocessing in data quality and the superiority of word embeddings, especially multilingual ones, in capturing semantic relationships and leading to more accurate topic identification. The project also identified data imbalance and noise as key challenges requiring further investigation.

Future research can build upon this foundation by exploring advanced AI models for deeper text interpretation, improved preprocessing techniques for low-resource languages like Malay, and leveraging domain-specific language models and transfer learning for enhanced accuracy. Additionally, addressing data imbalance through synthetic data generation and advanced sampling techniques, alongside investigating new clustering approaches, holds promise for unlocking even deeper insights and improving the effectiveness of text analysis in multilingual and noisy environments. Overall, this project establishes a solid foundation for future advancements in user review analysis and contributes valuable insights to the fields of text analysis and natural language processing.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to the following individuals and institutions for their invaluable support throughout this project:

- My supervisor, Dr. Goh Hui Ngo, for her continuous guidance, insightful feedback, and unwavering encouragement. Their expertise and support were instrumental in shaping this research and ensuring its quality.
- My moderator, Prof Ting Choo Yee, for their helpful suggestions and constructive feedback during the development of this project. Their insights were crucial in refining my research approach and presentation.

I am also grateful to the Selangkah for providing the dataset used in this research. Their willingness to share this valuable data resource significantly contributed to the successful completion of this project.

REFERENCES

[1] Agarwal, N., Sikka, G., & Awasthi, L. K. (2020). Evaluation of web service clustering using Dirichlet Multinomial Mixture model based approach for Dimensionality Reduction in service representation. Information Processing and Management, 57(4). https://doi.org/10.1016/j.ipm.2020.102238

[2] Amara, A., Hadj Taieb, M. A., & Ben Aouicha, M. (2021). Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. Applied Intelligence, 51(5), 3052–3073. https://doi.org/10.1007/s10489-020-02033-3

[3] Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in Sociology, 7. https://doi.org/10.3389/fsoc.2022.886498

[4] Gao, C., Zheng, W., Deng, Y., Lo, D., Zeng, J., Lyu, M. R., & King, I. (2019). Emerging App Issue Identification from User Feedback: Experience on WeChat. Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2019, 279–288. https://doi.org/10.1109/ICSE-SEIP.2019.00040

[5] Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Applied Soft Computing, 101. https://doi.org/10.1016/j.asoc.2020.107057

[6] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. http://arxiv.org/abs/2203.05794

[7] Guerzo, L. A. B., Kilkenny, H. A. O., Osorio, R. N. D., Villegas, A. H. E., & Ponay, C. S. (2021). Topic Modelling and Clustering of Disaster-Related Tweets using Bilingual Latent Dirichlet Allocation and Incremental Clustering Algorithm with Support Vector Machines for Need Assessment. Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021, 189–193. https://doi.org/10.1109/ICSECS52883.2021.00041

[8] Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C., & Kyriazis, D. (2023). Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Computing and Applications, 35(29), 21415–21431. https://doi.org/10.1007/s00521-023-08629-3

[9] Mouthami, K., Yuvaraj, N., Thilaheswaran, K. K., & Lokeshvar, K. J. (2023). Text Sentiment Analysis of Film Reviews Using Bi-LSTM and GRU. 2023 4th International Conference on Electronics and Sustainable Communication Systems, ICESC 2023 - Proceedings, 1379–1386. https://doi.org/10.1109/ICESC57686.2023.10193121

[10] Tounsi, A., Elkefi, S., & Bhar, S. L. (2023). Exploring the Reactions of Early Users of ChatGPT to the Tool using Twitter Data: Sentiment and Topic Analyses. Proceedings of the 2023 IEEE International Conference on Advanced Systems and Emergent Technologies, IC_ASET 2023. https://doi.org/10.1109/IC_ASET58101.2023.10150870

[11] Udupa, A., Adarsh, K. N., Aravinda, A., Godihal, N. H., & Kayarvizhy, N. (2022). An Exploratory Analysis of GSDMM and BERTopic on Short Text Topic Modelling. 2022 4th International Conference on Cognitive Computing and Information Processing, CCIP 2022. https://doi.org/10.1109/CCIP57447.2022.10058687

[12] Vasudeva Raju, S., Kumar Bolla, B., Nayak, D. K., & Jyothsna, K. H. (2022). Topic Modelling on Consumer Financial Protection Bureau Data: An Approach Using BERT Based Embeddings. 2022 IEEE 7th International Conference for Convergence in Technology, I2CT 2022. https://doi.org/10.1109/I2CT54291.2022.9824873

[13] Wang, X., Zhang, W., Lai, S., Ye, C., & Zhou, H. (2022). The Use of Pretrained Model for Matching App Reviews and Bug Reports. IEEE International Conference on Software Quality, Reliability and Security, QRS, 2022-December, 242–251. https://doi.org/10.1109/QRS57517.2022.00034

[14] Zhou, W., Wang, Y., Gao, C., & Yang, F. (2022). Emerging topic identification from app reviews via adaptive online biterm topic modeling. Frontiers of Information Technology and Electronic Engineering, 23(5), 678–691. https://doi.org/10.1631/FITEE.2100465