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

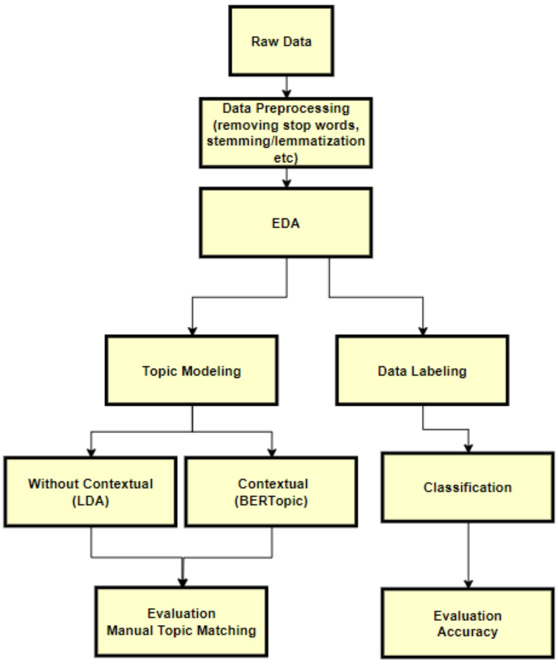


Fig. 1. Methodology

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

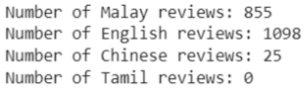
A. *EDA*

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

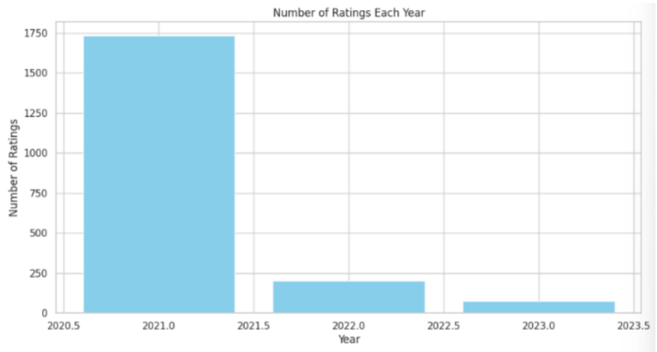
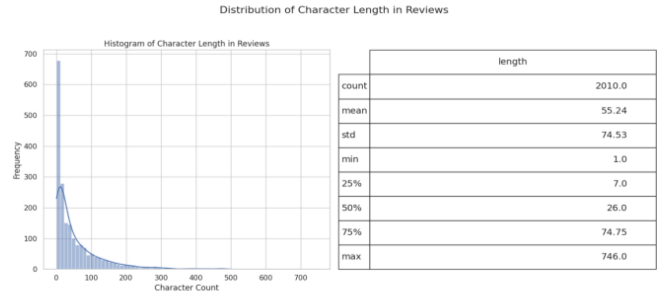


Fig. 3. Number of Ratings Each Year After Removing Null Rows



	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

Fig. 4. Distribution of Character Length in All Reviews

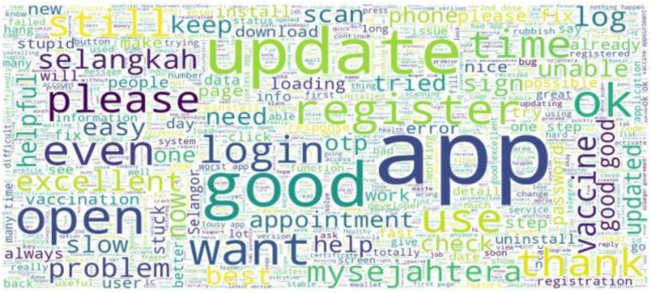


Fig. 5. Word Cloud for All User Reviews after Initial Preprocessing (after Initial Preprocessing)

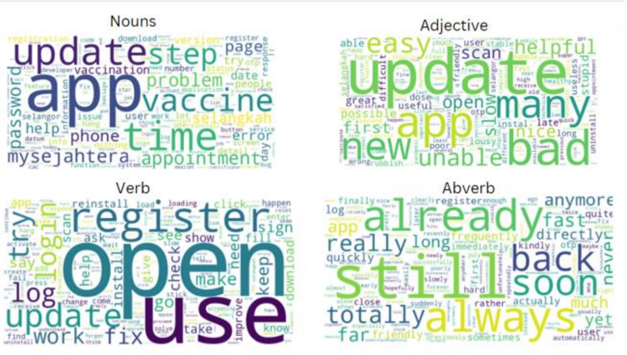


Fig. 6. POS Tagging Word Cloud for All User Reviews (after Initial Preprocessing)

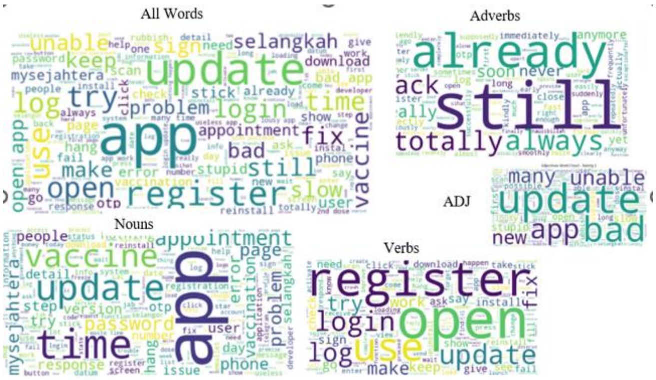


Fig. 7. Word Clouds for Rating 1 Reviews (after Initial Preprocessing)

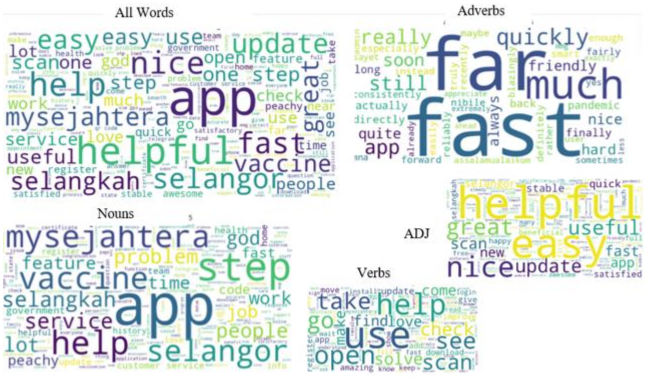


Fig. 8. Word Clouds for Rating 5 Reviews (after Initial Preprocessing)

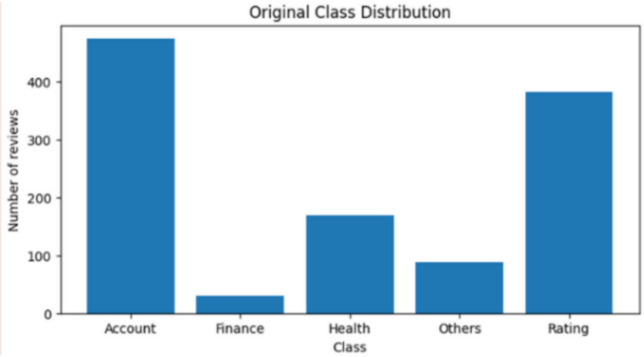


Fig. 9. Class Distribution of App Review Categories Before SMOTE (First Layer)

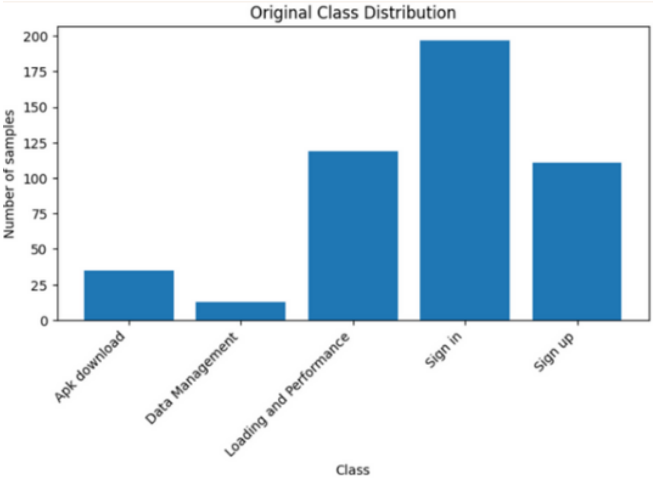


Fig. 10. Bar Chart - Class Distribution of App Usage Categories Before SMOTE (Second Layer)Topic Modeling Results

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