

2623: Segmentation of User Feedback and Ratings

Lee Tian Xin 1211301744

Supervisor: Ts. Dr. Goh Hui Ngo, Moderator: Prof. Ts. Dr. Ting Choo Yee



A. Abstracts

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 study presents a comprehensive approach to analyzing user feedback through a combination of unsupervised and supervised learning techniques, focusing 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, identify areas for improvement, and understand user sentiment across different functionalities.

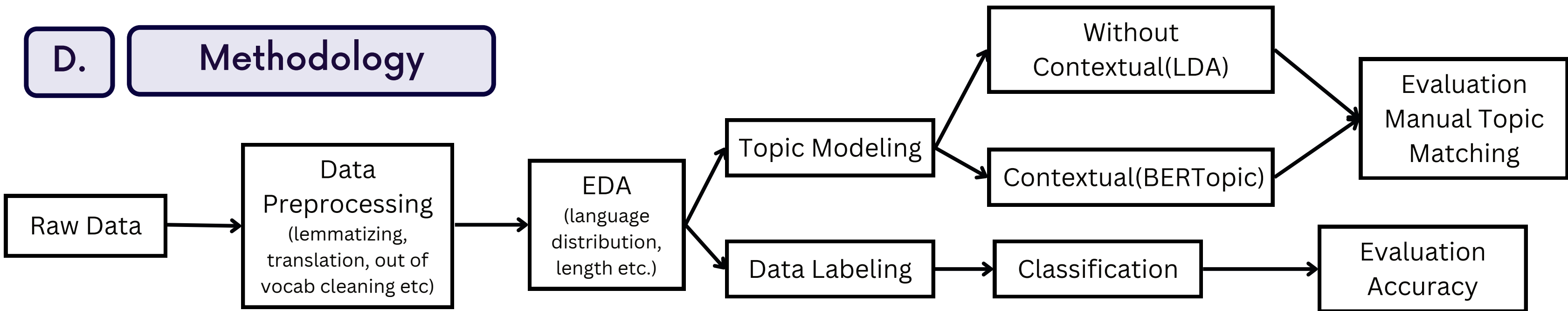
The results demonstrate the efficacy of this hybrid approach in extracting actionable insights from user reviews, thereby guiding targeted enhancements to the Selangkah app.

B. Problem Statement and Objectives

Problem Statement:
Manual sorting is time-consuming and susceptible to human error.

- Objectives:
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.

D. Methodology



Unsupervised Learning: Categorized user feedback into seven pre-defined topics (e.g., Sign-in, Health) to gain initial insights.
Supervised Learning: Defined category structure with five main categories (Account, Health, Finance, Rating, Others) and detailed subcategories for "Account" (Login, Sign up, Apk download, Performance, and Data Management).

C. Literature Review

- Topic Modeling:
- Traditional methods (LDA, DMM) struggle with short text sparsity.
 - LLMs (BERT, Llama2) capture contextual nuances and relationships.
- Classification:
- Pre-trained models (T5, Sentence T5, MiniLM, Sentence BERT) excel in tasks like matching app reviews; Sentence T5 performs best.
 - XLM-R is superior for multilingual Twitter data classification.
 - Combining Word2Vec and FastText with Bi-LSTM enhances sentiment analysis in film reviews.
 - SVM, Naive Bayes, and Decision Tree are effective for e-commerce product reviews.

E. Implementation and Evaluation

- Topic Modeling:
- LDA: Traditional method , requires preprocessing and optimal number of topics.
 - BERTopic: BERT-based, less preprocessing, auto-generates topics. Out of the 19 topics generated by BERTopic, 7 topics were found to have a match with the pre-defined topics.
- Classification:
- Embeddings: Both BERT and Multilingual Embeddings used to capture semantic information.
 - Classification Models: Logistic Regression, Random Forest, XGBoost, GRU, BiLSTM, standard BERT Transformer, and Multilingual Transformer.
 - Best performance: Multilingual Embedding + BiLSTM and GRU (up to 93.10% accuracy on balanced data).
- Handling Class Imbalance:
- SMOTE (Synthetic Minority Over-sampling Technique): Used to balance the dataset, significantly improving the performance of classification models.

F. Conclusion

- Hybrid approach combining unsupervised and supervised learning effectively analyzed Selangkah app user feedback.
- Key Learnings:
- BERTopic: Superior for topic modeling.
 - Multilingual Embeddings: Effective for handling diverse languages.
- Future Work:
- Explore advanced AI techniques and balanced data strategies.
 - Refine classification models and integrate semi-supervised learning.