SEGMENTATION OF USER FEEDBACK AND RATINGS

Presented By : Lee Tian Xin

2024

ABTRACTS

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 apps

PROBLEM STATEMENT

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

First Problem

Volume and Diversity of Data: Extracting insights from vast, varied, and brief user reviews is challenging.

Second Problem

Short Text Reviews Challenges: Noise, sparsity, and lack of context complicate manual interpretation.

OBJECTIVES

- **Objective 1**

  To perform exploratory data analysis to uncover themes within the Selangkah user feedback dataset

- **Objective 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.

LITERATURE REVIEW

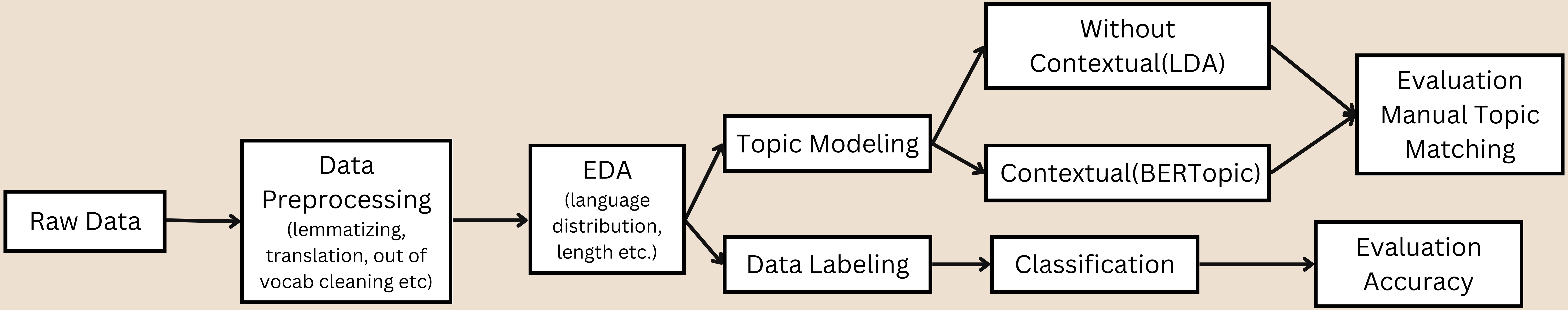
Topic Modeling

- **Traditional methods (LDA, DMM) struggle with short text sparsity.**
- **LLMs (BERT, Llama2) capture contextual nuances and relationships.**

Classification:

- **Pre-trained models like T5, Sentence T5, MiniLM, and Sentence BERT are used for matching app reviews;**
- **XLM-R for multilingual Twitter data classification;**
- **Word2Vec and FastText with Bi-LSTM for sentiment analysis in film reviews;**
- **SVM, Naive Bayes, and Decision Tree for e-commerce product review classification.**

METHODOLOGY


Raw Data
Data Preprocessing (lemmatizing, translation, out of vocab cleaning etc)
EDA (language distribution, length etc.)
Topic Modeling
Data Labeling
Without Contextual(LDA)
Contextual(BERTopic)
Evaluation Manual Topic Matching
Classification
Evaluation Accuracy

- **Unsupervised Learning: Categorized user feedback into seven pre-defined topics (Sign In, Register, APK Download, General Health, COVID-19, E-wallet, Welfare) 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).**

IMPLEMENTATION AND EVALUATION

Topic Modeling:

- **LDA: Traditional method , requires preprocessing and optimize number of topics.**
- **BERTopic: BERT-based, less preprocessing, auto-generates topics.**

IMPLEMENTATION AND EVALUATION

Topic Modeling:
BERTopic Performance:

- Out of the 19 topics generated by BERTopic, 7 topics were found to have a match with the pre-defined topics.
- Specifically, the "Rating" category, which was not a predefined functionality in BERTopic but used in classification, was allocated 7 out of the 19 topics.
- The remaining topics were either duplicates or noise.

IMPLEMENTATION AND EVALUATION

Topic Modeling:
LDA Performance:

- **Conversely, LDA generated 11 topics, out of which only 2 topics matched the pre-defined topics.**

IMPLEMENTATION AND EVALUATION

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**

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.

THANK YOU

Presented By : Lee Tian Xin