Predictive Model for Classification of Student Learning Behaviour in E-Learning using English-Malay Code-Mixing Corpus at Institute of Higher Learning

Bavani Raja Pandian*, Azlinda Abdul Aziz[2], Kamsiah Mohamed[3]

[1]Universiti Selangor
Bavaniraja86@yahoo.com
[2]Universiti Selangor
Azlinda@unisel.edu.my
[3]Universiti Selangor
kamsh@unisel.edu.my

Abstract: The evolving landscape of technology in education emphasizes the importance of predicting and comprehending student learning for personalized interventions and effective educational strategies. A qualitative design was employed for predictive model development and evaluation, including code-mixing corpus creation, annotation, linguistic analysis, and model assessment in Malaysia. The predictive model development involved data collection from social media and Learning Management System platforms, resulting in a dataset of 398,539 entries. Data pre-processing included removing duplicates, symbols, and irrelevant sentences, leading to a refined dataset of 78,988 entries. Manual and auto annotation methods were employed for data labeling, with the Support Vector Classifier achieving the highest accuracy of 88%. Model development utilized NLP algorithms with TF-IDF and Word2Vec embeddings. The study achieved high inter-annotator agreement (kappa value of 0.928) and presented comprehensive evaluations across various models, datasets, and metrics. This study utilized a comprehensive approach, collecting data from diverse e-learning platforms and employing rigorous pre-processing techniques. The model evaluation results show that LSTM (W2V) non-augmented obtained a high accuracy of 100%, LSTM (TF-IDF) non-augmented 97% Recall, and lastly, SVC (TF-IDF) non-augmented obtained a high Precision of 99%. In conclusion, the LSTM W2V model can classify the student learning behavior in this research.

Keywords: Predictive Model, Classification, Student Learning Behavior, E-learning, English-Malay Code-Mixing Corpus, Institute of Higher Learning

1. Introduction

Technology integration in education has significantly transformed in recent years, with e-learning platforms emerging as a pivotal component of the educational landscape. Qiu et al. (2022) assert that as Institutes of Higher Learning (IHLs) increasingly adopt online learning environments, understanding and predicting student learning behavior in these settings have become critical for enhancing the effectiveness of educational interventions and ensuring personalized learning experiences.

According to Romadhona et al. (2022), the existing code-mixing corpus utilized English, Malay, and Chinese. However, these languages cannot be employed in the present study as they are precisely centered on the two predominant languages spoken in Malaysia, Malay and English. Yan et al. (2019) found that, in the domain of predictive modeling in education, many predictive models were developed to focus on aspects other than student learning behavior, such as learning performance prediction, dropout rate prediction, and passing rate prediction. These prediction models could provide indirect insights into student learning behavior because they were not developed to analyze or evaluate it directly. Therefore, this research proposed a novel prediction model to classify student learning conduct on e-learning platforms. This enables a more thorough analysis of their behavior and participation in the educational process.

Besides that, in this study, a new code-mixing corpus, which consists of Malay and English code-mixing, is proposed by referring to student phycology state in e-learning platforms. Moreover, this study selects unstructured text data from two popular e-learning platforms, social media, and LMS, to train the predictive model and corpus creation. In addition to machine learning and deep learning algorithms, the BERT and pre-train models will be utilized to create the predictive model and select which classifier can achieve high accuracy. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained natural language processing (NLP) model developed by Google. The aims and Objectives of this research are to analyze the current student learning model at the Institute of Higher Learning, to create an unstructured Predictive model using code mixing corpus in e-learning, and to validate the unstructured predictive model using code mixing corpus in e-learning.

2. Code Mixing and Predictive Model Development Process

2.1 Methodology

This qualitative study employed a comprehensive approach to create and assess a predictive model.

2.2 Data Collection and Preprocessing

Figure 1 illustrates the development of the code-mixing corpus and predictive model, which will progress through several stages.

2.3 Code Mixing Corpus and Predictive Model Development

The following process is word vectorization or embedding, which includes techniques such as Word2Vec and TF-IDF. Meanwhile, for transformer models such as BERT, the corpus is created to train the tokenizer, which will convert the text inputs into machine-readable representations for the BERT model. After the model development, the following process is model training, where the dataset is split into an 80/20 ratio, where 80% of the dataset will be used to train the model, And 20% of the dataset will be used to test the model.

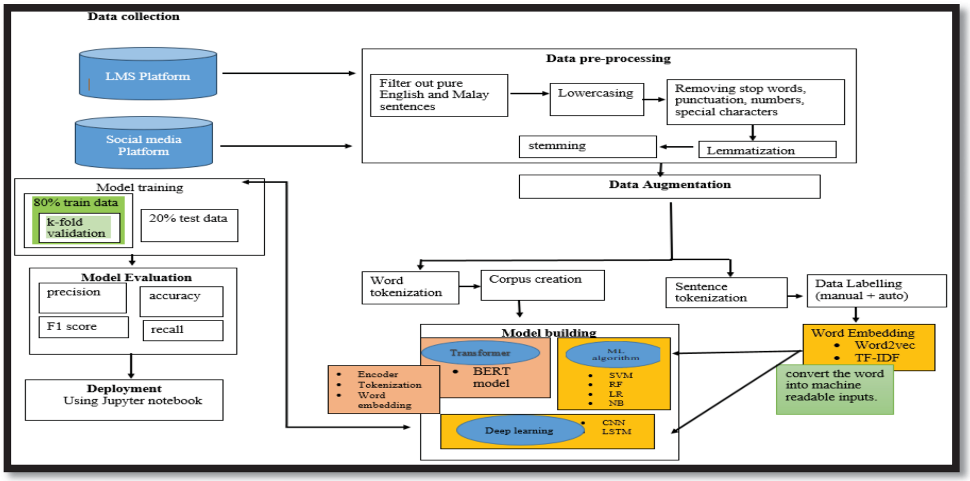


Figure 1. Process of Code-Mixing Corpus and Predictive Model Development

2.4 Content Analysis Process

In this study, code-mixed text data from six e-learning platforms were categorized into four student learning behavior types (activist, pragmatist, theorist, and reflector) by a team of psychologists and counselors, with inter-annotator agreement assessed using Cohen's kappa Shava et al. (2021).

2.5 Sample and Population

This study employed purposive sampling, deliberately selecting 5 Clinical Psychologists and five counselors from private universities, hospitals, and healthcare centers in Malaysia based on their expertise.

2.6 Inter-Annotator Agreement

Cohen's kappa (κ) serves as a crucial metric for assessing inter-annotator agreement in nominal data, contributing to both annotations' reliability and validity (Doewes et al. (2023).

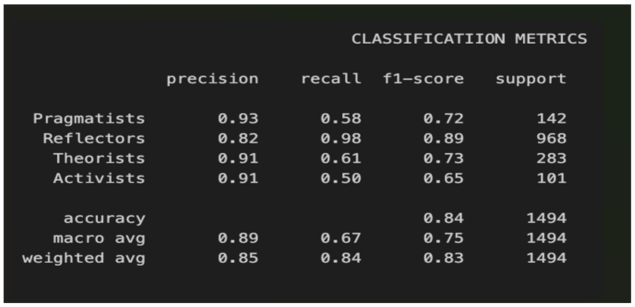
3. Results and Discussion

3.1 Data Labelling

The study used manual and auto-annotation methods to label datasets, employing classifiers for evaluation (see Table 1). Figure 2 presents the classification report for the SVC model.

Table 1. Accuracy of auto-labeling models

Classifiers	Accuracy	Classifiers	Accuracy
Support Vector classifier (SVC)	88%	Random Forest (RF)	65%
Linear Regression (LR)	68%	Conventional Neural Network (CNN)	72%
Naïve Bayes (NB)	65%	Long short-term Memory (LSTM)	75%



```
                          CLASSIFICATIION METRICS

              precision    recall  f1-score   support

 Pragmatists       0.93      0.58      0.72       142
  Reflectors       0.82      0.98      0.89       968
   Theorists       0.91      0.61      0.73       283
   Activists       0.91      0.50      0.65       101

    accuracy                           0.84      1494
   macro avg       0.89      0.67      0.75      1494
weighted avg       0.85      0.84      0.83      1494
```

Figure 2. SVC classification report

3.2 Result and Analysis of Model Development

Natural Language Processing (NLP) models were developed using TF-IDF and Word2Vec embeddings, employing an 80/20 data split for training/testing. Machine learning classifiers (Naive Bayes, SVC, Random Forest, Logistic Regression) utilized 5-fold validation, while Deep Learning and BERT models employed epochs. The evaluation included precision, accuracy, F1 score, and recall, with deployment in a Jupyter notebook. In conclusion, the study thoroughly analyzes code-mixing detection, leveraging a combination of traditional machine learning and deep learning models. The diverse set of evaluation metrics and datasets contributes to a comprehensive understanding of model performance, offering valuable insights for future research and applications in multilingual environments. Results show the Intermodel comparison table; based on that, it can be concluded that LSTM (W2V) non-augmented obtained a high F1 score which is 97%, LSTM (TF-IDF) and LSTM (W2V) non augmented was obtained high accuracy 100%, LSTM (TF-IDF) non augmented 97% Recall and lastly SVC (TF-IDF) non augmented was obtained high Precision 99%. In conclusion, the LSTM W2V model can classify the student learning behavior in this research.

4. Discussion

The study employed NLP algorithms, utilizing TF-IDF and Word2Vec for language code mixing prediction. Various classifiers, including Naive Bayes, SVC, Random Forest, and Logistic Regression, underwent 5-fold validation. Results indicated that LSTM (Word2Vec) achieved a high F1 score of 97% on non-augmented data, providing insights for future research in multilingual settings.

5. Conclusion

This research comprehensively analyses code-mixing detection models, showcasing the importance of diverse techniques and classifiers. The findings offer valuable insights for future endeavors in multilingual environments, emphasizing the potential for effective code-mixing detection in educational contexts.

6. References

Qureshi, M. A., Khaskheli, A., Qureshi, J. A., Raza, S. A., & Yousufi, S. Q. (2023). Factors affecting students' learning performance through collaborative learning and engagement. *Interactive Learning Environments*, *31*(4), 2371–2391. https://doi.org/10.1080/10494820.2021.1884886

Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B., & Chen, P. Kuo. (2022). Predicting students' performance in e-learning using learning process and behavior data. *Scientific Reports 2022 12:1*, *12*(1), 1–15. https://doi.org/10.1038/s41598-021-03867-8

Romadhona, N., Lu, S., … B. L.-P. of the 29th, & 2022, undefined. (n.d.). BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset. *Aclanthology.OrgNP Romadhona, SE Lu, BH Lu, RTH TsaiProceedings of the 29th International Conference on Computational, 2022•aclanthology.Org*, 4418–4428. Retrieved 10 December 2023, from https://aclanthology.org/2022.coling-1.389/

Yan, N., & Au, O. T.-S. (2019). Online learning behavior analysis based on machine learning. *Asian Association of Open Universities Journal*, *14*(2), 97–106. https://doi.org/10.1108/aaouj-08-2019-0029