Data-Driven Analysis of Senior High Students' Sentiments and Strand Selection Using Machine Learning

James E. Rosales[1], Cristopher C. Abalorio[2]
Department of Education[1]
Caraga State University - Main Campus, Butuan City, Philippines[1,2]
james.rosales2@carsu.edu.ph

Abstract – Senior high school students' sentiments play a vital role in shaping educational strategies that promote mental health, well-being, and learner protection as mandated by the Department of Education's child-friendly school system. This study investigates the relationship between students' sentiments and their chosen academic strand using machine learning algorithms. A total of 580 students participated, providing feedback through a guided support session conducted after the first quarter of the school year. The sentiments that were collected, which are code-mixing data, were preprocessed and vectorized using Term Frequency–Inverse Document Frequency (TF-IDF). A Support Vector Machine (SVM) classifier with four-fold cross-validation was employed to categorize sentiments into positive, neutral, and negative classes. The model achieved an average accuracy of 72.59% with a standard deviation of 3.91%, indicating a statistically significant relationship between students' sentiments and their chosen strands. Findings suggest the need for a dedicated technical working group to monitor and interpret sentiment trends per strand, enabling timely interventions. Future research is encouraged to explore alternative machine learning algorithms, such as Multinomial Naive Bayes, Logistic Regression, and Gradient Boosting (LightGBM), as well as advanced techniques like semantic embeddings and sarcasm detection, to improve classification accuracy. The results of this study contribute to data-driven decision-making and provide actionable insights for educational policy frameworks aimed at enhancing student welfare and learning outcomes.

Keywords – Cross-Validation, Senior High Students, Sentiment, Strand, Support Vector Machine

1 Introduction

The Department of Education (DepEd) has institutionalized a program to promote learners' rights and protection, which monitors their behavior and psychological well-being. However, the aforementioned resolution lacks in terms of recognizing the learners' mental well-being. Nowadays, mental health is at its core to acknowledgement, especially in basic education. Suicidal cases are increasing in the country. Moreover, DepEd is routing its resolution and mental health agenda to other agencies, hoping that an automated and systematic behavioral analysis can be implemented, which will easily identify learners' negativity, allowing for immediate action. The Department of Science and Technology believes that behavioral analysis will be enhanced through machine learning algorithms.

Behavioral analysis is not complex to configure, as numerous machine learning algorithms are used to identify the tone of sentiments and classify them into positive, negative, and neutral categories [1], [2]. Machine learning algorithms are helpful in describing learners' behavior easily [3].

Moreover, four machine learning algorithms genuinely stand out in terms of text classification when the data is in code-mixed. These are the Multinomial Naive Bayes, Logistic Regression, Gradient Boosting (LightGBM), and Support Vector Machine (SVM). All these machine learning algorithms have been tested in other studies, which produce efficient and accurate data-driven predictions [4], [5].

Several agencies still need to adapt to the information systems harmonization, as per the Department of Information and Communications Technology (DICT), which is on the verge of implementing the said harmonization across all agencies. This harmonization includes the data-driven analysis using machine learning algorithms that detect patterns and visualize data predictions [6].

2 Related Literature

Comparative study on binary classification for learners' sentiments prediction evaluates four machine learning algorithms Multinomial Naive Bayes, Logistic Regression, Gradient Boosting (LightGBM), and Support Vector Machine (SVM) highlighting their strengths and limitations; however, it does not dive advanced optimization techniques such as Genetic Algorithms (GA) for feature selection, potentially limiting model performance, and applies Random Oversampling only to the validation dataset, until the training data becomes balanced, which could impact use-case applicability. Similarly, Cruz et al. (2021) analyzed sentiments using a machine learning algorithm for preparing learners in basic education for their tertiary courses. Moreover, the Support Vector Machine achieves an accuracy of 89.96%, and the study recommends it [9]. Vorecol (2024) states that sentiment analysis, utilizing machine learning algorithms, is crucial in the basic education setting as it supports equity and inclusion issues [10]. Other studies have emphasized the importance of learners' mental and psychological health, especially those from the senior high school department. DepEd also highlights the need to strengthen the promotion of learners' rights and protection, with a focus on immediate intervention [7]. A study by Zhang and Zheng (2021) found that machine learning algorithms, such as Support Vector Machine (SVM), Multinomial Naive Bayes, Logistic Regression, and Gradient Boosting, are particularly applicable to sentiment analysis, especially when the primary objective is text classification. Furthermore, the vectorization of the Term Frequency–Inverse Document Frequency (TF-IDF) is significant because it facilitates the representation of textual code-mixing data. A study by Li et al. (2022) highlights the advantages and disadvantages of TF-IDF in handling code-mixed data, thereby addressing the redundancy issues associated with the vectorization [8], [9].

3 Methods

This section provides a detailed process for this study. It focuses on comparing the standard performance of four classifiers – Multinomial Naive Bayes, Logistic

Regression, Gradient Boosting (LightGBM), and Support Vector Machine (SVM) - under a single feature selection technique, namely Term Frequency–Inverse Document Frequency (TF-IDF). This approach aims to enhance the predictive performance of the models and provide a robust comparison.

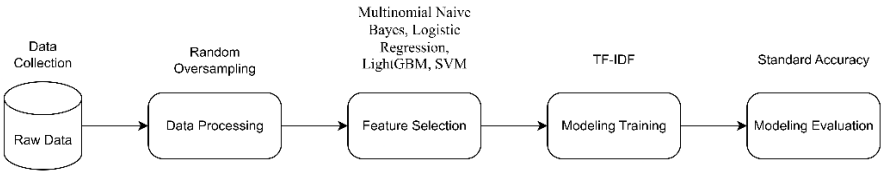

Fig. 1. Framework of the Study

3.1 Data Collection

For data collection, researchers surveyed 580 senior high school students in Taligaman National High School. Researchers have obtained consent from each respondent in accordance with the ethical standards. After the survey, the collected data were imported into the model for data preprocessing and cleaning. Researchers also verified the data before feeding it into the well-structured model for training. Using an actual primary data source aligns with the research's need for high-quality, authentic data that accurately represents the data-driven sentiment analysis.

3.2 Preprocessing and Balancing

The preprocessing phase includes collected data which is in code-mixed text from students in Taligaman National High School. Data goes with cleaning to remove duplication and noises including special characters and symbols. Tokenization was also implemented to make sure that data is in the form of individual token. By the function of the TF-IDF, text was vectorized into a numerical representation showcasing sensitive words minimizing common and insignificant text.

3.3 Feature Selection

The initial process involves implementing TF-IDF, which transforms the raw data into a numerical representation. The objective of this phase is to identify the significant features from the data for sentiment classification. Moreover, it reduces the insignificant text found in the raw data. The data was assessed through four-fold cross-validation.

3.4 Model Training and Evaluation

The model-building process involves a structured approach to classifier selection, training, and evaluation to ensure standard performance on the sentiment dataset. Classifier Selection is the initial step, where suitable classifiers, such as Support Vector Machine (SVM), Multinomial Naive Bayes, Logistic Regression, and Gradient Boosting, are chosen based on the problem's requirements and the data characteristics. Model Training follows, where the selected classifiers are trained on a balanced version

of the dataset, incorporating relevant features to improve the model's generalizability and predictive power. Finally, Model Evaluation is conducted using a held-out test set, employing standard accuracy to gauge each model's effectiveness; this metric provides a comprehensive assessment of model performance, enabling the identification of the most reliable classifier for analyzing sentiments.

3.5 Model Selection and Deployment

The model selection process involves identifying the best-performing model based on key performance metric, ensuring the chosen model aligns closely with the desired outcomes and accuracy requirements. Once the optimal model is selected, it undergoes a refitting procedure where it is retrained on the entire training dataset, utilizing the most effective subset of features identified during model tuning. This refitting step ensures that the model fully leverages all available training data for improved generalization and robustness. Finally, the refitted model is saved for future use or deployment, allowing seamless application in real-world scenarios or subsequent analyses.

4 Results & Discussion

In addressing the problem of senior high school sentiments by performing a random oversampling and preparing a balanced dataset for future predictive modeling. It begins with loading the dataset and conducting Exploratory Data Analysis (EDA) to understand its structure, including identifying missing values, generating summary statistics, and visualizing the original distribution of the target variable, `Sentiment_Tone`. The EDA reveals that the dataset is imbalanced, with more instances of positive sentiments rather than negative sentiments, which could have impacted the standard performance of machine learning classifiers.

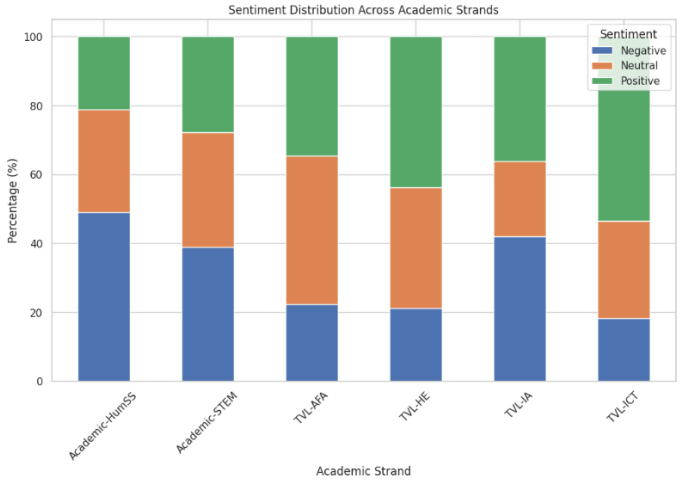

Fig. 2. Sentiments Distribution

Technical-vocational (TVL) strands like ICT, IA, HE, and AFA have a predominantly higher positive sentiment. Academic-HumSS strand exhibits the highest negative sentiment. Academic-STEM is more balanced but still leans negative.

Table 1. Accuracy Standard Deviation Results for a Feature Selection Method and Classifiers

Feature Selection Method	Accuracy (%)	Std. Dev (%)
Support Vector Machine	72.59	3.14
Multinomial Naïve Bayes	70.17	3.46
Logistic Regression	70.86	3.91
Gradient Boosting (LightGBM)	48.45	6.23

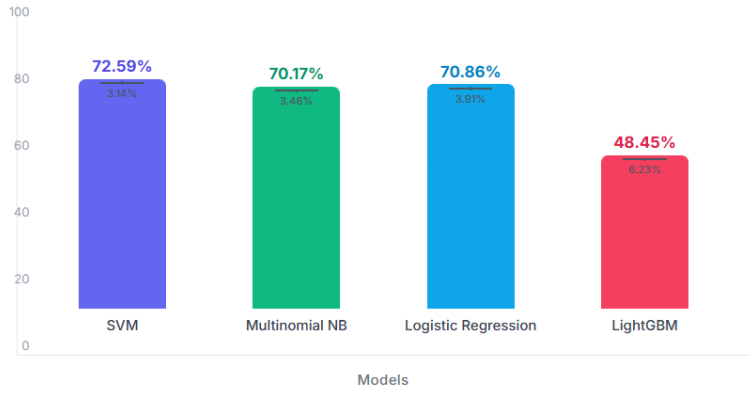

Fig. 3. Accuracy of Different Classifiers

The figure illustrates the test accuracy of four classifiers - Multinomial Naive Bayes, Logistic Regression, Gradient Boosting (LightGBM), and Support Vector Machine (SVM) - when applied to four-fold cross-validation and TF-IDF vectorization. It shows that SVM significantly outperforms the other classifiers, achieving the highest accuracy. This result reflects its ability to handle complex feature interactions effectively, which is a key strength of ensemble learning methods. In contrast, LightGBM demonstrates the lowest accuracy, indicating its limitations in capturing non-linear relationships within the data. Multinomial Naïve Bayes, while performing better than Logistic Regression, achieves moderate accuracy, reflecting its capability to handle high-dimensional data. The findings emphasize that SVM is the most effective model for sentiment prediction using TF-IDF vectorization.

5 Conclusion

The study concludes that Support Vector Machine (SVM) with four-fold cross-validation and TF-IDF is the most effective classifier for sentiment prediction with 72.59% standard accuracy. It consistently achieved the highest standard accuracy, demonstrating its ability to balance the tone of sentiments. This makes it a robust and reliable model for both fairness and sentiments' tone compared to Multinomial Naïve Bayes 70.17% standard accuracy indicating its weakness. On the other hand, logistic regression 70.86% standard accuracy and gradient boosting (LightGBM) got only

48.45% which is the lowest standard accuracy among all machine learning algorithms. These results indicate that linear classifiers such as SVM, Multinomial NB, and Logistic Regression are more reliable for code-mixed sentiment analysis tasks in educational datasets. The relatively high standard deviations highlight the importance of further model tuning and robust validation, but SVM's overall performance suggests it should be prioritized for practical sentiment monitoring in schools.

References

[1] Madasu, A., & Sivasankar, E. (2020). Efficient feature selection techniques for sentiment analysis. arXiv preprint arXiv:1911.00288. https://arxiv.org/abs/1911.00288

[2] Ahmad, S. R., Yusop, N. M. M., Asri, A. M., & Amran, M. F. M. (2021). A review of feature selection algorithms in sentiment analysis for drug reviews. International Journal of Advanced Computer Science and Applications, 12(12). https://doi.org/10.14569/IJACSA.2021.0121217

[3] Mujawar, S. S. ., & Bhaladhare, P. R. . (2023). Effective Feature Selection Methods for User Sentiment Analysis using Machine Learning. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3s), 37–45. https://doi.org/10.17762/ijritcc.v11i3s.6153

[4] Sharma, A., & Dey, S. (2012). A comparative study of feature selection and machine learning techniques for sentiment analysis. In Proceedings of the 2012 ACM Research in Applied Computation Symposium (RACS '12) (pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/2401603.2401605

[5] Pooja, Bhalla R. A Review Paper on the Role of Sentiment Analysis in Quality Education. SN Comput Sci. 2022;3(6):469. doi: 10.1007/s42979-022-01366-9. Epub 2022 Sep 9. PMID: 36106178; PMCID: PMC9462624.

[6] Lee, K. M., & Kim, K. (2025). An analysis of educational satisfaction using sentiment analysis: correlation analysis between Likert-scale evaluation and descriptive feedback. Human Resource Development International, 1–25. https://doi.org/10.1080/13678868.2025.2563353

[7] Baragash, R. S., Aldowah, H., & Umar, I. N. (2022). Students' perceptions of e-learning in Malaysian universities: Sentiment analysis based machine learning approach. Journal of Information Technology Education: Research, 21, 439-463. https://doi.org/10.28945/5024

[8] Saraswathi, N., Sasi Rooba, T., & Chakaravarthi, S. (2023). Improving the accuracy of sentiment analysis using a linguistic rule-based feature selection method in tourism reviews. Measurement: Sensors, 29, 100888. https://doi.org/10.1016/j.measen.2023.100888

[9] P. H. Prastyo, I. Ardiyanto and R. Hidayat, "A Review of Feature Selection Techniques in Sentiment Analysis Using Filter, Wrapper, or Hybrid Methods," 2020 6th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 2020, pp. 1-6, doi: 10.1109/ICST50505.2020.9732885.

[10] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.