Sentiment Mining on Social Media using Naive Bayes: A Tool for Enhancing Academic Program Decisions

Alphie P. Lavarias
San Carlos College
San Carlos City, Pangasinan, Philippines

Christian Ernes A Caranto
San Carlos College
Binmaley, Pangasinan, Philippines

Junard S. Secretario
San Carlos College
San Carlos City Pangasinan, Philippines

Queen Amchell B. Papa
San Carlos College San Juan,
San Carlos City, Pangasinan, Philippines

Romulo L. Olalia Jr.
San Carlos College
San Carlos City, Pangasinan, Philippines

Maynard Gel F. Carse
San Carlos College
Basista Pangasinan, Philippines

ABSTRACT
This study investigates the application of sentiment analysis to social media posts related to academic programs, utilizing datasets composed of both Filipino and English texts. Employing a Naive Bayes classifier, the system achieved an overall classification accuracy of 78.66%, effectively distinguishing positive, negative, and neutral sentiments within the feedback. The data preprocessing pipeline included thorough cleaning, normalization, tokenization, stopword removal, and lemmatization, all of which contributed to enhanced model performance. These findings demonstrate the practical utility of sentiment analysis as an analytical tool for academic institutions seeking to gauge stakeholder opinions and feedback. By identifying trends in sentiment, educational administrators can make informed decisions to improve program quality and engagement.

General Terms
Pattern Recognition, Data Mining, Educational Technology, Information Retrieval

Keywords
Artificial Intelligence, Academic Program, Educational Institution, Natural Language Processing, Machine Learning Algorithm, Sentiment Analysis, Stemming, Tokenization.

1. INTRODUCTION
The widespread adoption of social media has significantly influenced how individuals communicate, access information, and express opinions, with over 5.04 billion users engaging on these platforms globally as of early 2024—making them valuable sources of real-time public sentiment. For higher education institutions, social media presents an untapped opportunity to gain insights into student perceptions of academic programs, which traditional feedback methods like surveys and interviews often fail to capture due to their limited scope and delayed responses. Sentiment analysis, a branch of Natural Language Processing (NLP), enables the automated classification of opinions into positive, negative, or neutral categories, offering a scalable and timely alternative for understanding public sentiment. This project proposes a sentiment analysis framework designed to evaluate student feedback related to academic program offerings using a static dataset sourced from Kaggle. The textual data undergoes a preprocessing pipeline that includes normalization, tokenization, removal of non-English and non-Filipino content, and exclusion of emoticons and non-textual elements to ensure data consistency. Machine learning techniques—specifically Naive Bayes classification—are then employed to categorize sentiments, and the results are visualized through a custom-built, responsive dashboard using HTML, CSS, JavaScript, and Bootstrap. The project's main objectives are to extract and analyze feedback about academic programs, classify sentiments accordingly, and assess the usefulness of sentiment analysis in guiding institutional decisions. While the system presents a robust approach, it is limited to a static dataset and does not support real-time data or image-based content; it also excludes complex expressions such as sarcasm or symbolic text. Ethical standards are strictly followed by analyzing only publicly available, non-personal data. Despite challenges such as sarcasm detection, data quality variability, and evolving user behavior, this project aims to deliver actionable insights that promote continuous improvement and innovation in academic program offerings within higher education institutions.

2. RELATED LITERATURE
Sentiment analysis, also referred to as opinion mining, is a significant area within natural language processing (NLP) and machine learning, focusing on identifying and extracting subjective information from textual data to understand public emotions and opinions. This analytical approach finds widespread application across various domains such as education, marketing, and customer service, where gauging stakeholder sentiment is critical for informed decision- making [1], [6], [16].

In this study, sentiment analysis is applied to student feedback on university services and academic programs. The dataset used was sourced from Kaggle, a platform that provides diverse structured and unstructured data suitable for academic research. This dataset comprises student reviews that capture perceptions and satisfaction levels regarding institutional offerings, facilities, and overall experiences. By classifying the feedback into positive, negative, or neutral categories, institutions can derive actionable insights for strategic development and continuous improvement [4], [10].

For data storage and handling, MySQL was utilized due to its robustness, scalability, and high performance in managing structured data. According to Šušter and Ranisavljević [14] and Vyas [20], MySQL's ease of integration, secure architecture, and query efficiency make it suitable for research involving large-scale text data storage and retrieval.

The foundational study by Aqlan et al. [1] provides a comprehensive overview of sentiment analysis, encompassing the fundamental techniques, frameworks, and challenges. The methodology employed in this project mirrors this framework, especially through preprocessing stages such as tokenization, stop word removal, normalization, and lemmatization. These steps are essential in refining raw text data into structured inputs suitable for machine learning models [7], [12]. Rajesh and Hiwarkar [12], as well as Gurusamy and Kannan [7], have emphasized the critical importance of preprocessing in improving classification accuracy and reducing noise introduced by informal language elements such as emojis, URLs, and slang. Camilleri [4] further explores the utility of sentiment analysis in higher education, asserting its potential to enhance engagement, align academic offerings with student needs, and improve institutional strategies. Abdul Lasi et al. [10] demonstrate how sentiment analysis can guide improvements not only in business environments but also within educational settings, by interpreting user feedback effectively through NLP.

The implementation of this research project is supported by several technological tools and frameworks. The Massive Online Analysis (MOA) system facilitates stream-based learning and real-time analytics, offering valuable support in large-scale textual data analysis [3]. Python libraries such as NLTK and scikit-learn are employed to execute preprocessing and machine learning tasks efficiently, leveraging their mature ecosystem and proven capabilities [1], [13].

Despite its potential, sentiment analysis faces notable challenges. Ahmed et al. [2] highlight issues with misclassification in multilingual and culturally nuanced datasets, noting that algorithms may fail to interpret the context and tone accurately. Moreover, ethical concerns such as data privacy, algorithmic bias, and responsible use of results must be addressed, especially in educational and public sectors [2], [15].

In terms of classification models, Naive Bayes was selected for this study due to its simplicity, computational efficiency, and proven performance in sentiment classification tasks [6], [19]. Farhan Aftab et al. [6] regard it as a reliable model for text-based sentiment classification, and Vadlamudi et al. [19] reported an 87.80% accuracy rate using the Naive Bayes algorithm in similar applications. The model estimates the likelihood of sentiment classes—positive, negative, or neutral—based on word frequencies and trained data distributions. Its probabilistic nature allows for effective generalization, even with relatively small datasets.

Through these techniques, this research aims to deliver accurate sentiment evaluations of student feedback, contributing to academic quality assurance and student-centered planning. In an increasingly competitive higher education environment, understanding student sentiment becomes a strategic necessity to enhance satisfaction, reputation, and institutional effectiveness [4], [10], [11].

3. PROJECT AND DESIGN METHODOLOGY

This section outlines the complete data pipeline that converts raw social-media feedback into sentiment-based insights for academic decision-making. The process, illustrated in Figure 1, consists of eight sequential stages designed for transparency and reproducibility.

3.1 Pipeline Overview

An End-to-end bilingual sentiment-analysis pipeline.

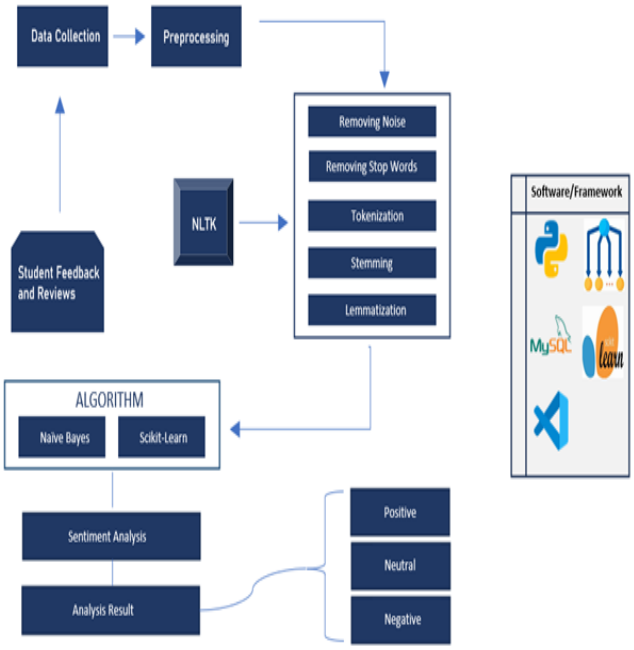

Figure 1. Eight (8) Pipeline Overview

3.2 Materials

Python 3.11, NLTK 3.9, Scikit-learn 1.4, Pandas 2.2, and MySQL 8. Experiments were run on a laptop with an Intel Core i3-10110U CPU, 8 GB RAM, and 256 GB SSD.

3.3 Data Collection

Three publicly available Kaggle datasets—*Datasetprojpowerbi* (1 006 records), *ExeterReviews* (602), and *HarvardUniversityReviews* (4 558)—were downloaded in CSV format. Only posts containing Filipino or English text and referencing academic programs or campus services were retained, yielding a blended corpus of **6 166** labelled entries.

Table 1. Dataset Sources and Record Counts

Dataset Name	Institutions	Record Count
Datasetprojpowerbi	General	1,006
ExeterReviews	University of Exeter	602
HarvardUniversityReviews	Harvard University	4,558
Total:		**6,166**

3.4 Data Pre-processing

The raw social-media text underwent a five-step cleaning pipeline to ensure high-quality input for feature extraction and model training, as summarized below.

- **Normalization** – convert text to lowercase and standardize Unicode.
- **Noise Removal** – strip numerals, punctuation, emojis, hashtags, and URLs via regex.
- **Tokenization** – split sentences into word tokens using NLTK's word_tokenize.
- **Stop-word Removal** – filter out common Filipino and English stop-words via custom lists.
- **Lemmatization** – apply WordNetLemmatizer (English) and Stanza (Filipino) to reduce inflectional forms.

3.5 Feature Extraction

Two vectorization schemes were benchmarked:

- **Bag-of-Words (BoW):** unigram and bigram counts limited to the top 10 000 vocabulary terms.
- **TF-IDF:** term-frequency/inverse-document-frequency weighting over the same vocabulary.

3.6 Model Training

A Multinomial Naïve Bayes (MNB) classifier was fit on 70 % of the data, using grid search to optimize the Laplace smoothing parameter $\alpha \in \{0.1, 0.5, 1.0\}$. Ten-fold cross-validation measured stability, and macro-averaged metrics (accuracy, precision, recall, F1) were logged.

3.7 Evaluation Protocol

To assess the generalization capability of the sentiment analysis model, a 70/30 train-test split was implemented, where 30% of the dataset was reserved exclusively for testing. This portion of the data, which remained unseen during the training phase, was used to evaluate how well the model performs on new and previously unobserved inputs. In addition to this standard evaluation approach, a cross-corpus validation was conducted. This involved training the model on two distinct datasets and testing it on a third, which provided deeper insight into the model's robustness across different data distributions and sources. Various performance metrics were computed, including accuracy, precision, recall, and F1-score. Furthermore, confusion matrices were generated to analyze the distribution of true positives, false positives, true negatives, and false negatives for each sentiment class. Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves were also plotted, offering a more granular diagnostic view into the classification performance and threshold behavior of the model.

3.8 Deployment and Visualization

The final MNB model and vectorizer were encapsulated in a Flask REST API and served to a Bootstrap dashboard that streams sentiment summaries (bar charts and line trends) for stakeholder review.

Figure 2. Distribution of Sentiment Labels in the Dataset (Positive, Negative, Neutral)

4. RESULT AND DISCUSSION

4.1 Dataset Overview

The final blended corpus contains **6 166** labelled posts distributed across three sentiment classes. Class proportions are reasonably balanced, ensuring that no category dominates the learning process and mitigating bias during training.

Table 2. Sentiment Class Distribution

Sentiment Class	Record Count	Proportion
Positive	2163	35.1%
Negative	2011	32.3%
Neutral	1992	32.6%
Total	6166	100%

4.2 Classification Performance

Model performance was evaluated on the 30 % hold-out test set. Table 3 presents the confusion matrix, while Table 4 lists the precision, recall, and F1-scores for each sentiment class.

Table 3. Confusion Matrix of Test-Set Predictions

	Predicted Negative	Predicted Neutral	Predicted Negative
Actual Negative	553	102	79
Actual Neutral	59	589	68
Actual Positive	37	122	579

Rows represent true labels; columns represent model predictions. Diagonal cells denote correct classifications.

Table 4. Classification Report (Macro-Averaged Accuracy = 78.66 %)

Sentiment Class	Precision	Recall	F1-Score	Support
Negative	0.85	0.75	0.80	734
Neutral	0.72	0.82	0.77	716
Positive	0.80	0.78	0.79	738
Macro Avg	0.79	0.79	0.79	2188
Weighted Avg	0.79	0.79	0.79	2188

Accuracy Computation.

True Positives = 553 (Negative) + 589 (Neutral) + 579 (Positive) = 1,721

Total Predictions = 553 + 102 + 79 + 59 + 589 + 68 + 37 + 122 + 579 = 2,188

Accuracy = 1,721 / 2,188 = 78.66 %

Interpretation

Negative class. High precision (0.85) indicates few false positives, but recall (0.75) shows that 25 % of true negatives are missed.

Neutral class. Recall is the highest (0.82), yet precision (0.72) reveals some confusion with neighbouring classes.

Positive class. Balanced precision (0.80) and recall (0.78) demonstrate robust identification of favourable feedback.3.1), use no additional space above the subsection head.

4.2.1 Naive Bayes Formula

$$P(C|X) = \frac{P(C) \cdot P(X|C)}{P(X)}$$

The model predicts the sentiment class CCC maximizing: where X represents features extracted from the preprocessed text (primarily term frequency and inverse document frequency).

4.3 Error Analysis

Misclassifications mainly arise between neutral and negative posts, often when a single message contains mixed sentiments (e.g., “Great prof, pero sobrang mahal ng tuition!”). These kinds of posts present difficulties for traditional models due to sentiment ambiguity and code-switching. Future work will explore sarcasm detection mechanisms and the integration of transformer-based contextual embeddings (e.g., BERT), which can better capture sentiment subtleties, semantic context, and language shifts.

Common misclassifications also include sarcastic or ironic expressions expressed in Filipino-English code-switched phrases such as “Ayos naman pero super hassle!” which reflect user frustration in subtle forms. Addressing these challenges may involve incorporating specialized modules for sarcasm and irony detection, as well as training models on a richer, more diverse dataset to improve generalization. Implementing transformer-based architectures with pre-trained multilingual embeddings can enhance the system’s ability to interpret context-sensitive sentiments and improve overall classification accuracy in future iterations.

5. CONCLUSION AND FUTURE SCOPE

This study demonstrates that a lightweight Multinomial Naïve Bayes model, when paired with carefully designed Filipino–English preprocessing steps, can achieve robust performance—achieving a classification accuracy of approximately 78.66%” for precision and consistency in sentiment classification of academic program feedback. The pipeline successfully transforms unstructured social-media posts into structured insights, enabling administrators to make informed, data-driven decisions regarding curriculum design, campus services, and student engagement strategies.

Beyond accuracy, the practical deployment of the model into a live dashboard reinforces its applicability for real-time institutional monitoring. By tracking trends in student sentiment, university leaders can proactively identify emerging concerns and respond to feedback in a timely and transparent manner.

However, while effective, the current approach is not without limitations. As the complexity and volume of social-media content continue to grow, future improvements will focus on deepening the model’s linguistic and contextual understanding. One major direction involves the integration of transformer-based architectures such as BERT and RoBERTa, which can better handle context-switching, sarcasm, and subtle emotional cues—especially prevalent in multilingual or code-switched environments.

In addition, future iterations of the system could incorporate multimodal analysis, expanding the input space to include emoji-rich posts, memes, and image captions that frequently carry sentiment-laden content. This would enable a more holistic view of student feedback across diverse expression formats.

Finally, the pipeline can be evolved into an always-on microservice, continuously connected to social-media APIs. Such a real-time, scalable deployment would ensure that sentiment data remains fresh and actionable, providing a strategic advantage for institutional planning, student support initiatives, and overall educational quality assurance.

Taken together, these enhancements will allow for more adaptive, nuanced, and comprehensive sentiment monitoring systems that align closely with the evolving needs of both students and academic institutions.

6. ACKNOWLEDGEMENTS

The authors would like to express their profound gratitude to the College of Information and Computing Studies of San Carlos College for providing the academic platform, technical resources, and administrative support necessary to conduct this research. The institution’s commitment to fostering innovation, excellence, and research-driven learning played a vital role in the successful development and completion of this study.

We extend our heartfelt appreciation to the Dean of the College of Information and Computing Studies for their continued encouragement and for cultivating a culture that values research, critical inquiry, and academic integrity. Their support served as a cornerstone of our motivation throughout this endeavor.

The authors are especially indebted to their Capstone Project Adviser, whose unwavering guidance, mentorship, and

insightful feedback greatly enhanced the quality and rigor of this research. Their expertise and dedication not only refined our methodology but also inspired a deeper understanding of the relevance and implications of sentiment analysis in the context of academic program evaluation.

We also acknowledge the collaborative efforts and encouragement of our faculty mentors, classmates, and peers, who provided both moral and technical support during the various phases of this work. Furthermore, we are grateful to the open-data contributors on Kaggle, whose datasets were instrumental to our model development and evaluation.

This study would not have been possible without the collective efforts of the academic community at San Carlos College, to whom we owe our sincerest thanks.

7 REFERENCES

[1] Aqlan, A., Bairam, M., & Naik, R. L. (2019). A Study of Sentiment Analysis: Concepts, Techniques, and Challenges. (pp. 147-162). DOI: 10.1007/978-981-13-6459-4_16.

[2] B. Dhiman, "Ethical Issues and Challenges in Social Media: A Current Scenario: 5 Apr 2023, J.C. Bose University of Science and Technology, YMCA, Faridabad, Haryana, India, March 29, 2023

[3] Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., & Gavaldà, R. (2012). MOA: Massive Online Analysis. The Journal of Machine Learning Research, 11, 1601–1604. CBM.

[4] Camilleri, M.A. (2019). Higher Education Marketing: Opportunities and Challenges in the Digital Era. Academia, 0(16-17), 4-28. DOI:10.26220/aca.3169Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] Damota, M. D. (2019). The Effect of Social Media on Society. New Media and Mass Communication, Vol. 78. Retrieved from www.iiste.org ISSN 2224-3267 (Paper) ISSN 2224-3275 (Online) DOI: 10.7176/NMMC

[6] Farhan Aftab, Sibghat Ullah Bazai, Shah Marjan, & Laila Baloch (2023). A Comprehensive Survey on Sentiment Analysis Techniques. International Journal of Technology, 4(6), 1288. DOI: 10.14716

[7] Gurusamy, V., & Kannan, S. (2014). Preprocessing Techniques for Text Mining. The Journal of Machine Learning Research, 11, 1601–1604. Retrieved from Research Gate.

[8] Iglesias, C. A., & Moreno, A. (2019). Sentiment Analysis for Social Media. Applied Sciences, 9(23), 5037. DOI: 10.3390/app9235037.

[9] Kadhim, A. (2018). An Evaluation of Preprocessing Techniques for Text Classification. International Journal of Computer Science and Information Security, 16(6), 22-32. ResearchGate

[10] Lasi, M.b.A., Hamid, A.B.b.A., Jantan, A.H.b., Goyal, S.B., Tarmidzi, N.N.b. (2024). Improving Digital Marketing Using Sentiment Analysis with Deep LSTM. In: Swaroop, A., Polkowski, Z., Correia, S.D., Virdee, B. (eds) Proceedings of Data Analytics and Management. ICDAM 2023. Lecture Notes in Networks and Systems, vol 785. Springer, Singapore

[11] Petrosyan, A. (2024, May 7). Worldwide digital population 2024. Statista. [https://www.statista.com/statistics/617136/digita l-population-worldwide/]

[12] Rajesh, M. A., & Hiwarkar, T. (2023). Exploring Preprocessing Techniques for Natural Language Text: A Comprehensive Study Using Python Code. International Journal of Engineering Technology and Management Sciences, 7(5), DOI:10.46647/ijetms.2023.v07i05.047. Semantic Scholar.

[13] Ramasubramanian, K., Singh, A. (2019). Machine Learning Model Evaluation. In: Machine Learning Using R. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4215- 5_7

[14] Šušter, I., & Ranisavljević, T. (2023). Optimization of MySQL database. Journal of Process Management and New Technologies, DOI:10.5937/jouproman2301141qCorpus ID: 259754662.Semantic Scholar.

[15] Taherdoost, H. (2021). Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Collection Technique for Academic

[16] Tan KL, Lee CP, Lim KM. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. Applied Sciences. 2023; 13(7):4550.

[17] Journal of Innovative Research in Applied Sciences and Engineering, 4(4), 2456-8910. DOI: 10.29027/IJIRASE.v4.i4.2020.735-742.

[18] Uçar, U., Nour, M. K., Sindi, H. F., & Polat, K. (2020). The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets. *Mathematical Problems in Engineering, 2020*(1), 1-17. DOI: 10.1155/2020/2836236

[19] Vadlamudi, P. S., Gunasekaran, M., Nagalakshmi, T. J., & Saveetha University. (2023). An Analysis of the Effectiveness of the Naive Bayes Algorithm and the Support Vector Machine for Detecting Fake News on Social Media. In 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) (pp. DOI:10.5937/jouproman2301141qCorpus ID: 259754662)

[20] Vyas, K. (2023). 8 Major Advantages of Using MySQL. DataMation.