Evaluating The Impact of Social Media Sentiment on University Enrollment Decisions Using Machine Learning Classifier

Painem[1,*], Hari Soetanto[2], Achmad Solichin[3], Anju A Nair[4]

[1,2,3]Faculty of Information Technology, Universitas Budi Luhur,Jakarta, Indonesia
[4]Acharya Institute of Technology, Bengaluru, Karnataka 560107, India
*Corresponding email: painem@budiluhur.ac.id

Received: Month xx, xxxx; Revised: Month xx, xxxx; Accepted: Month xx, xxxx.

Abstract: Public sentiment expressed through social media is increasingly recognized as a potential factor influencing higher education enrollment decisions. This study investigates whether sentiments on Twitter regarding Universitas Budi Luhur correlate with the number of new student admissions. To achieve this, tweet data were collected and analyzed using four supervised machine learning algorithms—Support Vector Classifier (SVC), Naïve Bayes, K-Nearest Neighbor (KNN), and Logistic Regression (LR)—combined with two lexicon-based sentiment dictionaries: SentiWord and InSet. Experimental results demonstrate that the SentiWord-based approach consistently outperformed the InSet-based approach across all models, with the SVC-SentiWord combination achieving the highest F1-score of 0.86. Despite the strong performance of these models in classifying sentiment, correlation analysis reveals no statistically significant relationship between Twitter sentiment and actual student enrollment trends. These findings underscore the effectiveness of lexicon-enhanced machine learning in sentiment analysis while raising important questions about the real-world impact of online sentiment on university admissions. From a practical perspective, the findings of this study provide valuable insights for university stakeholders, particularly in strategic decision-making related to promotion and reputation management. The developed sentiment analysis model can be utilized to monitor public perception in real time, identify emerging issues that may affect institutional image, and design more effective communication strategies to enhance the university's attractiveness and credibility among prospective students.

Keywords: social media, sentiment analysis, student enrollment, machine learning, support vector classifier

1. Introduction

Universitas Budi Luhur (UBL) is a private higher education institution in Indonesia committed to delivering quality education and fostering innovation. However, similar to many private universities nationwide, UBL has faced a declining trend in new student admissions over the past several years. This pattern aligns with national-level concerns reported by the Directorate General of Higher Education (DIKTI), which indicate that private universities—constituting more than 95% of all higher education institutions in Indonesia—are experiencing increasing pressure due to fluctuating enrollment numbers and a stagnating Gross Enrollment Ratio (GER) of around 31–32% [1]. These indicators highlight the urgent need for private institutions to strengthen their competitiveness, maintain financial sustainability, and improve public trust in order to remain viable in an increasingly competitive educational landscape.

Universitas Budi Luhur (UBL) is a private higher education institution in Indonesia committed to delivering quality education and fostering innovation. Like many universities, UBL places great emphasis on increasing the number of new student admissions each academic year as a measure of institutional growth and sustainability. However, over the past five years, UBL has experienced a concerning downward trend in student enrollment, particularly in the regular admission track (see Figure 1). This trend not only affects institutional revenue but also poses challenges to maintaining competitiveness and public trust in the university's reputation.

In the era of digital communication, public perception has become a critical factor influencing students' decision-making processes. Social media platforms, particularly Twitter, have emerged as real-time channels where prospective students express their opinions about universities, academic programs, campus facilities, and institutional reputation. Such sentiment expressions may shape or signal enrollment behaviors, making them a potential early indicator of changes in student interest. However, despite the growing relevance of social media sentiment, universities rarely utilize these digital signals in a systematic and data-driven manner to support strategic enrollment management.

Understanding the root causes behind the decreasing in admissions has become an urgent strategic priority. While various internal and external factors could be influencing this trend—such as program relevance, tuition costs, or changes in demographics—public perception, especially as voiced through social media, is an increasingly influential factor in shaping the decisions of prospective students. In the digital age, platforms like Twitter serve as a real-time barometer of public sentiment, enabling institutions to capture and respond to public opinion at scale.

This research hypothesizes that public sentiment expressed on Twitter toward UBL may have a measurable impact on new student admissions. To test this hypothesis, we propose a data-driven approach that leverages Machine Learning (ML) techniques for sentiment analysis. Specifically, we evaluate the performance of several classification algorithms—Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbor (KNN), and Logistic Regression—using two sentiment feature extraction approaches: a SentiWord and InSet lexicon dictionary. These models classify social media posts into positive, negative, or neutral sentiments. The resulting sentiment trends are then statistically correlated with actual enrollment data to examine whether a significant relationship exists.

By combining natural language processing with predictive modeling, this study aims to offer both practical and academic contributions. Practically, it provides university stakeholders with valuable insights into how social media sentiment may shape enrollment patterns, supporting more targeted communication and marketing strategies. Academically, the study contributes to the growing literature on the application of machine learning in educational contexts—particularly in sentiment analysis for institutional decision-making. Moreover, this methodology can serve as a model for other educational institutions facing similar enrollment challenges in the digital era.

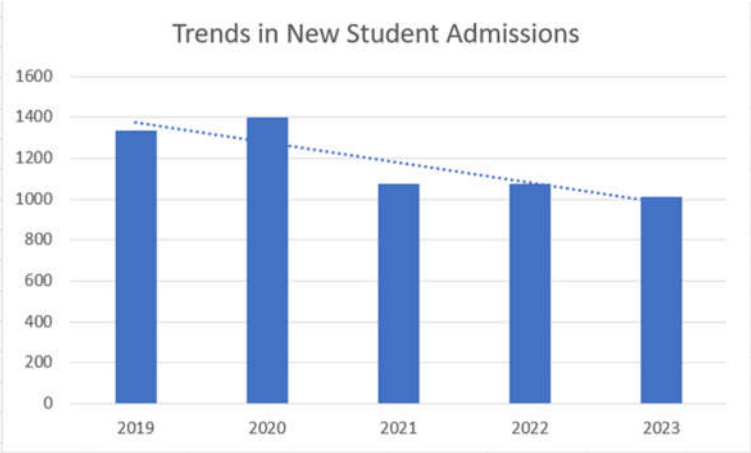

Figure 1: Trends in new student admissions Caption of figure.

Existing studies in Indonesia have widely applied machine learning (ML) techniques to sentiment analysis across various domains, including education, public policy, and institutional reputation. Recent studies have demonstrated that public sentiment expressed via social media platforms can significantly influence university perception and decision-making among prospective students [2]. In a similar vein, sentiment analysis on Twitter has been proposed as an alternative indicator for evaluating institutional reputation [3]. Within the Indonesian context, Widyayulianto et al. [4] applied deep learning techniques to assess Twitter sentiment about Universitas Pertamina, highlighting the practical potential of machine learning for understanding public opinion toward higher education institutions [4]. Table 1 presents recent research relevant to various methods of sentiment analysis of social media data.

Table 1: Recent Research related to Sentiment Analysis Methods using a Machine Learning approach

Method	Pros	Cons	Recent Research
Support Vector Machine (SVM)	High accuracy	Requires quite large data	[5], [6], [7], [8], [9], [10]
Naive Bayes	Simple and easy to implement	Lower accuracy compared to SVM	[11], [12], [13], [14], [15]
Decision Tree	Can handle complex data	Lower accuracy compared to SVM and Naive Bayes	[12], [16], [17]
Random Forest	High accuracy	Requires quite large data	[8], [18], [19], [20]
Neural Network	Very high accuracy	Huge amounts of data and long training times are required	[21], [22]

In addition to the methods used, the success of sentiment analysis also depends on how features are extracted from the text data. Various methods of text feature extraction are widely used in various research. Table 2 presents some popular text feature extraction methods among researchers.

Table 2: Text Feature Extraction Method on Social Media

Method	Pros	Cons	Recent Research
Bag of Words (BoW)	Simple and easy to implement	Does not consider the context of the word	[23], [24]
N-gram	Consider the context of the word	Still has the same weaknesses as BoW	[25], [26]
TF-IDF	Take into account the number of times a word appears and the number of times a word appears throughout the document	Still has the same weaknesses as BoW and n-gram	[23], [26], [27]
Word Embedding	Consider the meaning of words	Requires quite large data	[26]
Deep Learning	Can produce complex and meaningful features	Huge amounts of data and long training times are required	[21], [28]

The selection of the best feature extraction method for sentiment analysis depends on several considerations. For simple data and small data volumes, simple feature extraction methods, such as BoW, can provide sufficiently high accuracy. These methods are also relatively easy to implement and do not require

long training times. For complex data and large volumes of data, more complex feature extraction methods, such as deep learning, can provide higher accuracy. However, these methods also require longer training times.

Nevertheless, the majority of these studies focus only on classifying sentiments or describing sentiment trends. **Very few studies attempt to link sentiment data to real institutional outcomes, such as actual student enrollment numbers**. As a result, there remains a critical research gap: understanding whether public sentiment on social media correlates with, or potentially influences, student admissions in Indonesian higher education institutions—particularly within the private sector, where enrollment challenges are most pressing.

To address this gap, this study investigates whether public sentiment expressed on Twitter toward Universitas Budi Luhur has a measurable relationship with new student enrollment trends. Using four supervised machine learning algorithms—Support Vector Classifier (SVC), Naïve Bayes, K-Nearest Neighbor (KNN), and Logistic Regression—and two lexicon-based sentiment dictionaries (SentiWord and InSet), the study analyzes sentiment patterns and correlates them with actual monthly enrollment data.

The novelty of this research lies in its integration of machine learning–based sentiment analysis with institutional enrollment metrics, offering not only sentiment classification but also empirical evidence of whether online sentiment reflects real-world admission outcomes. This approach moves beyond descriptive analysis by positioning sentiment as a potential analytical tool for strategic decision-making in higher education.

By providing insights into the relationship between public sentiment and student enrollment, this study contributes both theoretically and practically. For researchers, it enhances the understanding of sentiment-based behavioral indicators in educational contexts. For university stakeholders, it offers actionable knowledge for monitoring public perception, refining communication strategies, and strengthening institutional competitiveness amid declining enrollment trends.

2. Research Method

2.1 Research Methodology

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to guide the research process. CRISP-DM provides a structured approach for carrying out data mining projects, encompassing six key phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Figure 2 presents the stages of the CRISP-DM methodology.

1. **Business Understanding**: In this initial phase, the research aims and objectives are defined in alignment with the university's goals and requirements. The focus is on understanding the business context and determining how data mining can address the research problem of analyzing the impact of social media sentiment on new student admissions at Universitas Budi Luhur. The primary objective of this study is to examine whether public sentiment expressed on social media—specifically Twitter—has a measurable relationship with new student enrollment at Universitas Budi Luhur (UBL).
2. **Data Understanding**: This phase involves collecting, exploring, and initial data preparation. Data sources, including Twitter social media data, are identified and gathered. Exploratory data analysis techniques are applied to gain insights into the characteristics of the data and to identify potential challenges and opportunities for analysis.
3. **Data Preparation**: Data preprocessing tasks, such as cleaning, integration, and transformation, are performed in this phase to ensure the data is suitable for analysis. Text data from social media sources are preprocessed to remove noise, tokenize, and convert into a structured format for further analysis.
4. **Modeling**: Machine Learning models, including Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbor (KNN), and Logistic Regression, are applied to the prepared data to build predictive models of sentiment analysis. The models are trained on labeled data and tuned for optimal performance.
5. **Evaluation**: The performance of the developed models is evaluated using appropriate metrics, such as F1-score, accuracy, and precision-recall curves. The models are assessed for their ability to accurately classify sentiment in social media data and their suitability for predicting new student admissions trends.

6. **Deployment**: In this final phase, the findings and insights obtained from the analysis are presented to stakeholders, including university administrators and marketing teams. Recommendations for strategic decision-making and future actions are provided based on the results of the sentiment analysis.

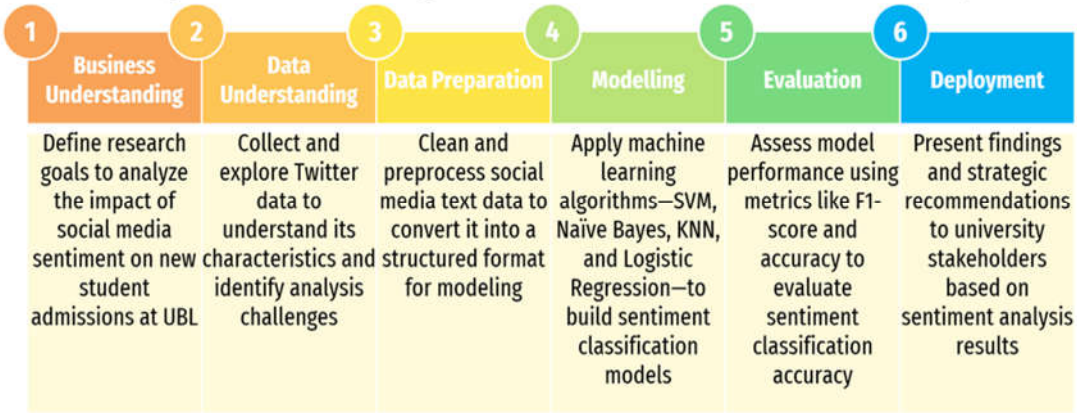

Figure 2: Research methodology [29]

2.2 Dataset

The dataset for this research was meticulously constructed from Twitter, a prominent social media platform, with data collected over the period from January 2022 to December 2022. The data collection process targeted tweets containing specific keywords associated with "Universitas Budi Luhur" and "Budi Luhur".

The dataset for this research was meticulously constructed from Twitter, a prominent social media platform, with data collected over the period from January 2022 to December 2022. The data collection process targeted tweets containing specific keywords associated with "Universitas Budi Luhur" and "Budi Luhur".

1. **Data Source**: Twitter was selected as the primary data source due to its extensive user base and the abundance of user-generated content relevant to our research objectives.
2. **Data Collection Period**: The data collection was conducted over the course of one year, spanning from January 2022 to December 2022. This duration was chosen to capture a comprehensive snapshot of public sentiment over an extended period.
3. **Keywords Used**: The keywords employed for data collection were "Universitas Budi Luhur" and "Budi Luhur". These keywords were carefully chosen to target discussions and mentions specifically related to Budi Luhur University.
4. **Data Acquisition**: The data acquisition process yielded a total of 1686 tweets containing the keyword "Budi Luhur" and 253 tweets containing the keyword "Universitas Budi Luhur". This resulted in a combined total of 1939 tweets.
5. **Data Cleansing**: Subsequently, the collected dataset underwent a rigorous cleansing process to enhance its quality and relevance. This cleansing process involved several steps, including the removal of duplicate tweets, elimination of retweets, and filtering out of non-English tweets.

After the cleansing process, the dataset was refined to 1526 tweets, ensuring that only relevant and high-quality data were retained for further analysis. Table 3 provides a summary of the data acquisition process, including the number of tweets obtained for each keyword and the resulting number of tweets after the cleansing process.

Table 3 : Summary of Data Acquisition and Cleansing

#	Keyword	Total Tweets Obtained	Cleansed Tweets
1.	“Budi Luhur”	1686	1301
2.	“Universitas Budi Luhur”	253	225
	TOTAL	1939	1526

It is evident that the majority of tweets were collected using the keyword "Budi Luhur", contributing 1686 tweets to the dataset. After cleansing, 1301 tweets remained for analysis. Additionally, 253 tweets were obtained using the keyword "Universitas Budi Luhur", with 225 tweets remaining after cleansing.

The cleansing process was crucial in ensuring the quality and relevance of the dataset, as it eliminated duplicate tweets, retweets, and non-Indonesian content. The resulting dataset of 1526 tweets serves as the foundation for subsequent analysis and modeling efforts, providing a focused and reliable dataset for sentiment analysis regarding Budi Luhur University on Twitter during the specified time frame. Figure 3 presents preprocessing steps :

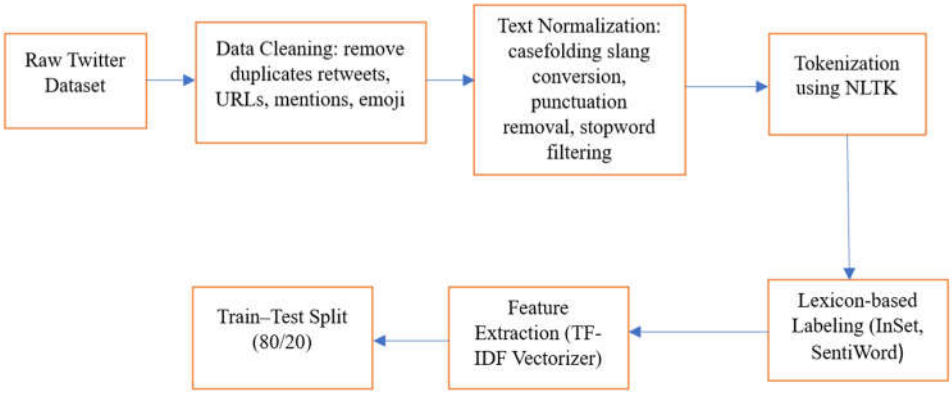

Figure 3: Preprocessing steps

Based on Figure 3, this is the preprocessing steps which consists of : cleaning: Removal of URLs, hashtags, mentions, emojis, duplicates, and retweets. Case Folding: Converting all tokens to lowercase. Stopword Removal: Using NLTK's Indonesian stopword list. Tokenization: Splitting sentences into tokens using NLTK tokenizer. Normalization: Handling slang/abbreviations common in Indonesian tweets. Lexicon-based Labeling: Two lexicons used: InSet and SentiWord and Sentiment scores computed using equations (1)–(3). Feature Extraction: TF-IDF vectorizer used to transform text into numerical features. Dataset Splitting: 80% for training; 20% for testing.

2.3 Data Labeling

The text data labeling process using the lexicon dictionary approach involves the use of a list of words that have been labeled with sentiment categories (such as positive, negative, or neutral) to determine the sentiment of the text. Labeling is performed using the lexicon method. Determination is made on text data in the form of sentences containing words from the lexicon dictionary, which consists of negative and positive words. Words identified in the lexicon dictionary are scored based on the number of occurrences in each text or sentence.

$$S_{positive} = \sum_{i \in t}^{n} (positif\ score_{\ i}) \qquad (1)$$

$$S_{negative} = \sum_{i \in t}^{n} (\ negative\ score_{\ i}) \qquad (2)$$

The text labeling process using the lexicon dictionary is based on Equation 1 and Equation 2. According to these equations, Spositive represents the weight of the sentence obtained by summing the scores of positive opinion words (Snegative), and Snegative is the weight of the sentence obtained by summing the scores of negative opinion words. From the equation, the sentiment value in one sentence is obtained, and then Equation

3 is derived to determine the sentiment orientation by comparing the total values of positive, negative, and neutral.

$$Sentence_{sentiment} = \begin{Bmatrix} positive\ if\ S_{positive} > S_{negative} \\ neutral\ if\ S_{positive} = S_{negative} \\ negative\ if\ S_{positive} > S_{negative} \end{Bmatrix} \quad (3)$$

If a text contains more positive words than negative words, the text data will be labeled as having a positive sentiment. If a text contains fewer positive words than negative words, the text data will be labeled as having a negative sentiment. If a text contains an equal number of positive and negative words, the text data will be labeled as having a neutral sentiment. This description outlines the labeling criteria based on the comparison of positive and negative word counts within a text. It clarifies that the sentiment label (positive, negative, or neutral) is determined by the relative frequency of positive and negative words in the text.

For the labeling process, this research utilized two lexicon dictionaries, namely the InSet dictionary and the SentiWord dictionary. InSet is a lexicon dictionary containing words labeled with sentiment categories, such as positive, negative, or neutral. This dictionary is specifically designed for sentiment analysis tasks and is tailored to capture sentiment in Indonesian language text data. It provides a comprehensive list of words along with their sentiment labels, facilitating the sentiment analysis process.

Meanwhile, the SentiWord is a lexicon dictionary used to label sentiment in Indonesian text. This dictionary contains a list of words annotated with corresponding sentiment values, such as positive, negative, or neutral. SentiWord enables researchers and practitioners in sentiment analysis to identify the sentiment contained within Indonesian language texts. Table 4 and Table 5 present statistics on words in the two dictionaries along with their weighting.

Table 4: InSet Lexicon Weighting

Positive		**Negative**	
Score	**Amount**	**Score**	**amount**
1	546	-1	263
2	660	-2	578
3	1353	-3	2573
4	870	-4	1980
5	180	-5	1215

Table 5: SentiWord Lexicon Weighting

	Positive		**Negative**
Score	**Amount**	**Score**	**Amount**
1	55	-1	134
2	30	-2	232
3	123	-3	315
4	268	-4	462
5	41	-5	69

3. Results

This research compares four Machine Learning methods: K-Nearest Neighbors (KNN), Logistic Regression, Multinomial Naive Bayes, and Support Vector Classifier. Each of these methods is applied to the

dataset labeled with the InSet and SentiWord lexicon dictionaries. KNN is a non-parametric classification algorithm that assigns a class label to a data point based on the majority class among its k nearest neighbors. It is simple to implement and suitable for datasets with small to moderate sizes. Logistic Regression is a statistical model used for binary classification tasks. It estimates the probability that a given input belongs to a certain class using a logistic function. It is widely used due to its simplicity and interpretability.

Multinomial Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with strong independence assumptions between features. It is particularly suited for text classification tasks, such as sentiment analysis. SVC is a supervised learning algorithm that finds the optimal hyperplane in a high-dimensional space to separate data points into different classes. It is effective in handling high-dimensional data and is often used in sentiment analysis tasks.

In this research, each of these Machine Learning methods is trained and evaluated using the dataset labeled with both the InSet and SentiWord lexicon dictionaries. By comparing the performance of these methods on the labeled dataset, the research aims to identify the most effective approach for sentiment analysis in the context of Universitas Budi Luhur mentions and discussions in Indonesian languages.

3.1 Logistic Regression

The logistic regression method is a machine learning technique for classification tasks. Its steps include: first, data collection consisting of features and class labels. Second, data preprocessing such as normalization or standardization of features, handling missing values, and splitting data into training and testing sets. Third, initialization of model parameters, typically with random or zero values. Fourth, model training using optimization algorithms such as gradient descent to optimize the loss function, which in logistic regression usually employs the log-loss function. Fifth, model evaluation using metrics such as accuracy, precision, recall, or area under the ROC curve (AUC-ROC) depending on the application context. And sixth, the use of the trained model to make predictions on new data or implementation in relevant applications.

Table 6: Modeling results using the Logistic Regression method

Lexicon Dictionary	Accuracy	Precision	Recall	F1-score
Sentiword	0.83	0.85	0.83	0.81
InSet	0.81	0.83	0.81	0.80

From the data used in this research, the results of modeling with the Logistic Regression method can be seen in Table 6. Based on Table 6, this method yields similar results to the performance evaluation using the Sentiword and InSet lexicons. In both lexicons, Logistic Regression produces high accuracy, with Sentiword slightly higher (0.83) than InSet (0.81). This indicates that the Logistic Regression model is capable of classifying data well using both lexicons. Additionally, in both Sentiword and InSet, the Logistic Regression model demonstrates high precision, meaning the model tends to reduce false positives. Then, Sentiword's recall is slightly higher than InSet. And also, Sentiword's F1-score (0.81) is higher than InSet (0.80), indicating a better balance between precision and recall when using Sentiword with the Logistic Regression model. Therefore, overall, the Logistic Regression model provides satisfactory and effective results in classification using both evaluated lexicons.

3.2 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) method is a learning algorithm that utilizes the distance between data to classify or regress new instances. Its steps include: first, selecting the parameter k, which is the number of nearest neighbors to be used for classifying the new instance. Second, calculating the distance between the new instance and every data point in the training set using a specific distance metric such as Euclidean distance or Manhattan distance. Third, selecting the k nearest neighbors based on these distances. Fourth, for classification, choosing the most common class among these neighbors, while for regression, calculating the average or

median of the target values of these neighbors. And fifth, assigning the class or regression value of the new instance based on the majority class or value.

Table 7: KNN method modeling results

Lexicon Dictionary	Accuracy	Precision	Recall	F1-score
Sentiword	0.81	0.83	0.81	0.80
InSet	0.79	0.82	0.79	0.77

Based on Table 7, the SentiWord method exhibits the highest accuracy (0.81) among the two lexicon dictionaries compared, followed by InSet with an accuracy of 0.79. The SentiWord also demonstrates slightly higher precision (0.83) compared to InSet (0.82), indicating that SentiWord tends to produce fewer false positives. Nevertheless, SentiWord boasts a higher recall (0.81) than InSet (0.79), signifying that SentiWord is generally better at identifying positive cases overall. Overall, the F1-score of SentiWord (0.80) is slightly higher than InSet (0.77), indicating a better balance between precision and recall with SentiWord. Therefore, based on this performance evaluation, the SentiWord method may be superior in the given context.

In summary, while InSet exhibits marginally higher precision, SentiWord offers a better balance between precision and recall, ultimately leading to a higher F1-score. This suggests that SentiWord is more effective overall in correctly identifying positive sentiment instances, making it a preferable choice for sentiment analysis tasks within the specified context.

3.3 Multinomial Naive Bayes

The Multinomial Naive Bayes method is a machine learning technique commonly used for text classification and categorical data. Its steps include: first, data collection of text and preprocessing, including steps such as tokenization (splitting text into words or tokens), punctuation removal, and normalization (e.g., converting all letters to lowercase). Second, building a model based on the Naive Bayes assumption, where it is assumed that each feature (word in the case of text classification) is independent of each other with respect to its class. Third, calculating prior probabilities for each class based on the frequency of class occurrences in the training data. Fourth, calculating conditional probabilities of each word in each class, which is the probability of a word appearing in a certain class divided by the total number of words in that class. Fifth, when test data is input, the model uses prior probabilities and conditional probabilities to predict the class of the test data. And sixth, evaluating the model's performance using metrics such as accuracy, precision, recall, or F1-score, depending on the application context.

Table 8: Modeling results of the Multinomial Naive Bayes method

Lexicon Dictionary	Accuracy	Precision	Recall	F1-score
Sentiword	0.85	0.85	0.85	0.84
InSet	0.82	0.83	0.82	0.82

From the data used in this research, the results of modeling with the Multinomial Naive Bayes method can be seen in Table 8. Based on Table 8, the model performance is similar between the two evaluated lexicons. In both Sentiword and InSet, the Multinomial Naive Bayes model demonstrates high accuracy, with Sentiword having a slight advantage (0.85) compared to InSet (0.82). The precision obtained from both is also quite high, with almost the same values for Sentiword (0.85) and InSet (0.83), indicating the model's ability to minimize false positives. Although Sentiword's recall (0.85) is slightly higher than InSet (0.82), Sentiword's F1-score (0.84) is also slightly higher than InSet (0.82), indicating a better balance between precision and recall when using Sentiword with the Multinomial Naive Bayes model. Therefore, overall, the Multinomial Naive Bayes model provides consistent and effective results in classification using both evaluated lexicons.

3.4 Support Vector Classifier (SVC)

The Support Vector Classifier (SVC), also known as the Support Vector Machine (SVM) in classification cases, is a machine learning technique used for data classification. Its steps include: first, data collection consisting of features and class labels. Second, preprocessing data such as normalization or standardization of features, handling missing values, and splitting data into training and testing sets. Third, selecting an appropriate kernel to map the data into a higher-dimensional feature space. Commonly used kernels include linear, polynomial, radial basis function (RBF), or sigmoid. Fourth, training the model by finding the best hyperplane that separates instances from different classes in the feature space with maximum margin and/or handling class imbalance cases by adjusting class weights. Fifth, evaluating the model using metrics such as accuracy, precision, recall, or area under the ROC curve (AUC-ROC) depending on the application context. And sixth, using the trained model to make predictions on new data or implementing it in relevant applications.

Table 9: Modeling results of the Support Vector Classifier (SVC) method

Lexicon Dictionary	Accuracy	Precision	Recall	F1-score
Sentiword	0.86	0.86	0.86	0.86
InSet	0.84	0.85	0.84	0.84

Based on Table 9, the modeling results using the Support Vector Classifier (SVC) method show good performance for both evaluated lexicons. SVC provides high accuracy, with a value of 0.86 for Sentiword and 0.84 for InSet, as well as reasonably high precision, namely 0.86 for Sentiword and 0.85 for InSet, indicating the model's ability to reduce false positives. Although Sentiword's recall is slightly higher than InSet, resulting in a higher F1-score for Sentiword (0.86) compared to InSet (0.84), indicating a better balance between precision and recall when using Sentiword with the SVC model. Therefore, overall, the SVC model delivers excellent results in classification using both evaluated lexicons, with superior performance especially when used with the Sentiword lexicon.

4. Discussion

Table 10 presents the comparison results of the evaluated methods including Logistic Regression, K-Nearest Neighbors (KNN), Multinomial Naïve Bayes, and Support Vector Classifier (SVC). The higher the accuracy value, the better the model predicts sentiment. In this experiment, SVC using SentiWord has the highest accuracy (0.86), while KNN using InSet has the lowest accuracy (0.79). High precision indicates that the model has few false positives. In this table, SVC using SentiWord has the highest precision (0.86). High recall indicates that the model has few false negatives. Then, SVC using SentiWord also has the highest recall (0.86).

Table 10 : Comparison of Logistic Regression, KNN, Multinomial Naive Bayes and SVC methods

Lexicon	Method	Accuracy	Precision	Recall	F1-score
SentiWord	Logistic Regression	0.83	0.85	0.83	0.81
	KNN	0.81	0.83	0.81	0.80
	Multinomial Naïve Bayes	0.85	0.85	0.85	0.84
	Support Vector Classifier	**0.86**	**0.86**	**0.86**	**0.86**
InSet	Logistic Regression	0.81	0.83	0.81	0.80
	KNN	0.79	0.82	0.79	0.77
	Multinomial Naïve Bayes	0.82	0.83	0.82	0.82
	Support Vector Classifier	0.84	0.85	0.84	0.84

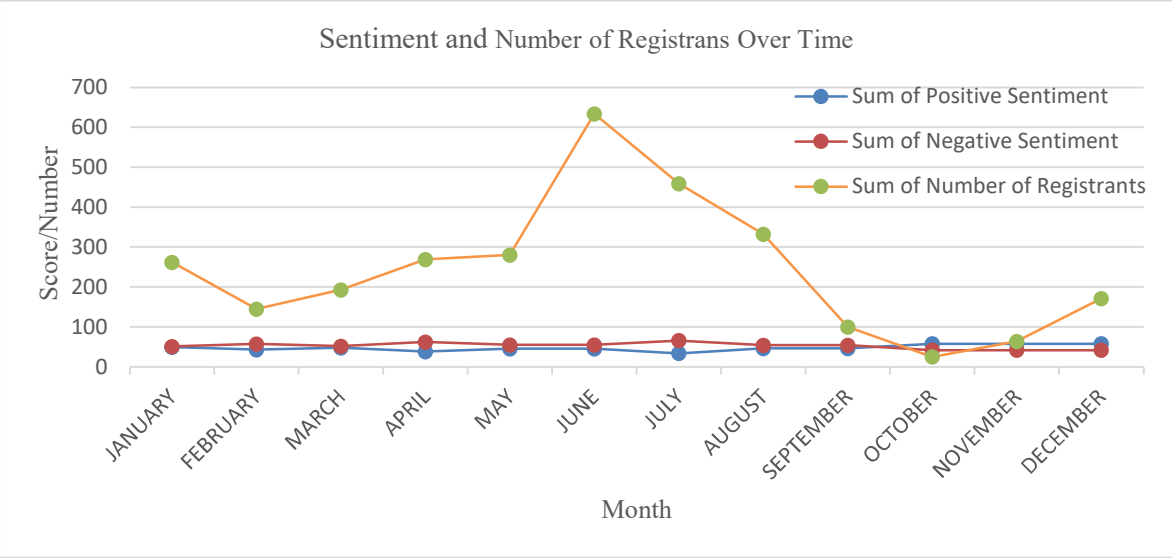

Figure 4: Relationship between positive and negative sentiment and the number of student enrollment

From Table 10, it is evident that in the evaluation of model performance using various machine learning methods and different lexicons, the Support Vector Classifier (SVC) model exhibits the most consistent and superior performance overall. Both with the SentiWord and InSet lexicons, the SVC model achieves high accuracy (0.86 and 0.84), high precision (0.86 and 0.85), sufficiently high recall (0.86 and 0.84), and good F1-score (0.86 and 0.84). This demonstrates SVC's capability to produce consistent and effective results in classification under various conditions and using different lexicons. However, the Logistic Regression model on the SentiWord lexicon also yields good results, with high accuracy, precision, recall, and F1-score. Additionally, Multinomial Naïve Bayes also demonstrates good performance, especially in terms of precision and F1-score. Nevertheless, overall, SVC stands out as the most consistent and superior choice in this evaluation.

Figure 4 presents a time series comparison between public sentiment on social media and the number of new student registrants at Universitas Budi Luhur throughout the year. The number of registrants exhibits significant fluctuation, with a sharp increase in June followed by a steep decline from July to September. In contrast, both positive and negative sentiment levels remain relatively stable across all months, showing only minor variations within a narrow range of scores (approximately 40 to 80). This suggests that public sentiment, as expressed on Twitter, remained generally constant regardless of changes in admission numbers.

Despite the strong model performance, the correlation analysis demonstrates **no statistically significant relationship** between monthly Twitter sentiment and new student enrollment at Universitas Budi Luhur. This outcome diverges from several studies conducted in international settings, where sentiment has been shown to influence educational decision-making. Several factors may explain the insignificance of the correlation: (1) university enrollment decisions in Indonesia are influenced more strongly by structural and institutional variables—such as tuition fees, program relevance, accreditation, and scholarship availability—than by general online sentiment; (2) Twitter users represent a broad public audience rather than prospective students specifically; (3) sentiment fluctuations do not necessarily align with the academic admission cycle; and (4) the sentiment dataset is relatively small and limited to a single platform. These factors collectively reduce the likelihood that Twitter sentiment alone would meaningfully correlate with actual enrollment outcomes.

The observed pattern provides empirical support for the correlation analysis conducted in this study, which found no statistically significant relationship between social media sentiment and student enrollment figures. Notably, the highest registration month (June) did not coincide with any substantial rise in positive

sentiment, nor did the decline in registrants align with a surge in negative sentiment. These findings indicate that while sentiment analysis can effectively capture public perception, it may not be a strong predictor of actual enrollment behavior. Other variables—such as academic schedules, marketing campaigns, scholarship offers, or changes in admission policies—are likely to have a more pronounced influence on prospective students' decisions. Therefore, sentiment analysis should be viewed as a complementary tool rather than a primary indicator in predicting university admissions trends.

5. Conclusion

Based on the experimental results using both SentiWord and InSet lexicon dictionaries, the Support Vector Classifier (SVC) achieved the highest performance among all models tested, with F1-scores of 0.86 for SentiWord and 0.84 for InSet. These findings suggest that SVC is a strong candidate for sentiment analysis tasks where a balance between precision and recall is critical. While Logistic Regression also showed competitive results, SVC consistently outperformed across both lexicons, particularly when paired with the more robust SentiWord dictionary.

However, despite the successful classification of sentiment from Twitter data, the correlation analysis revealed no statistically significant relationship between public sentiment on social media and the number of new student admissions at Universitas Budi Luhur. This indicates that social media sentiment, although measurable and classifiable, does not directly influence university enrollment trends. Consequently, institutions should interpret online sentiment as a complementary indicator rather than a decisive factor in admissions forecasting. When selecting machine learning models and sentiment lexicons, considerations such as domain relevance, scalability, and computational resources remain important, but they should be aligned with realistic expectations regarding the actual impact of public sentiment on enrollment behavior.

This study acknowledges several limitations: the dataset is relatively small; it relies solely on Twitter data; the analysis is based on a single institutional case; and lexicon-based labeling may not fully capture linguistic nuances such as sarcasm. Future research may expand the dataset across multiple platforms, integrate deep-learning–based sentiment representations, and compare results across several universities to enhance generalizability.

In summary, while Twitter sentiment shows no significant correlation with student enrollment, the sentiment model remains useful for monitoring public perception and supporting strategic communication. The findings underscore the importance of combining sentiment metrics with broader institutional data when developing predictive or diagnostic tools for enrollment management.

Acknowledgments

Acknowledgement is only addressed to funders or donors and object of research. Acknowledgement can also be expressed to those who helped carry out the research.

References

[1] W. Wattanapornprom, N. Tongta, N. Jaisamak, P. Lakhan, P. Dilokpatpongsa, and W. Susutti, “Enhanced Sentiment Detection in Thai University Admissions Using Complement Naive Bayes,” in *2024 28th International Computer Science and Engineering Conference (ICSEC)*, IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/ICSEC62781.2024.10770715.

[2] R. Al Bashaireh, V. Sabeeh, and M. Zohdy, “Towards a New Indicator for Evaluating Universities Based on Twitter Sentiment Analysis,” in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, Dec. 2019, pp. 1398–1404. doi: 10.1109/CSCI49370.2019.00261.

[3] M. R. Widyayulianto, M. Susanty, and A. Irawan, "Sentiment Analysis Terhadap Tulisan Mengenai Universitas Pertamina Di Media Sosial Twitter," *PETIR*, vol. 15, no. 2, pp. 276–286, Nov. 2022, doi: 10.33322/petir.v15i2.1197.

[4] S. Pandya and P. Mehta, "A Review On Sentiment Analysis Methodologies, Practices And Applications," *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, vol. 9, p. 2, 2020, [Online]. Available: www.ijstr.org

[5] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, "Sentiment analysis and classification of Indian farmers' protest using twitter data," *International Journal of Information Management Data Insights*, vol. 1, no. 2, Nov. 2021, doi: 10.1016/j.jjimei.2021.100019.

[6] B. Al sari *et al.*, "Sentiment analysis for cruises in Saudi Arabia on social media platforms using machine learning algorithms," *J Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00568-5.

[7] N. Leelawat *et al.*, "Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning," *Heliyon*, vol. 8, no. 10, Oct. 2022, doi: 10.1016/j.heliyon.2022.e10894.

[8] G. F. Pramudi, G. Firmansyah, B. Tjahjono, and A. M. Widodo, "Analysis of School Community Sentiment towards Personal Data Protection Law Using Support Vector Machine (SVM) Method," *Asian Journal of Social and Humanities*, vol. 1, no. 12, pp. 1256–1275, Sep. 2023, doi: 10.59888/ajosh.v1i12.121.

[9] O. Chamorro-Atalaya *et al.*, "Supervised learning using support vector machine applied to sentiment analysis of teacher performance satisfaction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, p. 516, Oct. 2022, doi: 10.11591/ijeecs.v28.i1.pp516-524.

[10] M. Agus Arianto and A. Solichin, "Analisis Sentimen Motogp Mandalika Pada Twitter Menggunakan Metode Naïve Bayes," *Jurnal TICOM: Technology of Information and Communication*, vol. 11, no. 1, 2022, [Online]. Available: https://t.co/XyNW7StiWQ

[11] N. Kewsuwun and S. Kajornkasirat, "A sentiment analysis model of agritech startup on Facebook comments using naive Bayes classifier," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 3, pp. 2829–2838, Jun. 2022, doi: 10.11591/ijece.v12i3.pp2829-2838.

[12] K. Zerrouki, R. M. Hamou, and A. Rahmoun, "Sentiment Analysis of Tweets Using Naïve Bayes, KNN, and Decision Tree," *... Sentiment Analysis Across ...*, 2022, [Online]. Available: https://www.igi-global.com/chapter/sentiment-analysis-of-tweets-using-nave-bayes-knn-and-decision-tree/308507

[13] A. A. Dwisanny and S. Supatmi, "Twitter Opinion Sentiment Analysis Based on New Student Admission Zoning Issues Using the Naïve Bayes and TensorFlow Methods," in *2023 9th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, IEEE, Dec. 2023, pp. 1–8. doi: 10.1109/ICSPIS59665.2023.10402745.

[14] A. Ariefah, Widodo, and M. Nugraheni, "Sentiment Analysis for Curriculum of Independent Learning Based on Naïve Bayes with Laplace Estimator," in *2023 International Conference on Information Technology Research and Innovation (ICITRI)*, IEEE, Aug. 2023, pp. 157–161. doi: 10.1109/ICITRI59340.2023.10249320.

[15] A. Naresh and P. Venkata Krishna, "An efficient approach for sentiment analysis using machine learning algorithm," *Evol Intell*, vol. 14, no. 2, pp. 725–731, 2021, doi: 10.1007/s12065-020-00429-1.

[16] Indri Tri Julianto, D. Kurniadi, and B. B. Balilo Jr, "ENHANCING SENTIMENT ANALYSIS WITH CHATBOTS: A COMPARATIVE STUDY OF TEXT PRE-PROCESSING," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 6, pp. 1419–1430, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.1448.

[17] I. B. Prakoso, D. Richasdy, and M. D. Purbolaksono, "Sentiment Analysis of Telkom University as the Best BPU in Indonesia Using the Random Forest Method," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 4, p. 2050, Oct. 2022, doi: 10.30865/mib.v6i4.4567.

[18] H. Junianto, R. E. Saputro, B. A. Kusuma, D. Intan, and S. Saputra, "COMPARISON OF LOGISTIC REGRESSION AND RANDOM FOREST IN SENTIMENT ANALYSIS OF DISDUKCAPIL APPLICATION REVIEWS," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 6, 2024, doi: 10.52436/1.jutif.2024.5.6.1802.

[19] P. Metode *et al.*, "Application Random Forest Method for Sentiment Analysis in Jamsostek Mobile Review," *Jurnal Informatika dan Teknologi Informasi*, vol. 20, no. 1, pp. 116–127, 2023, doi: 10.31515/telematika.v20i1.8868.

[20] U. D. Gandhi, P. Malarvizhi Kumar, G. Chandra Babu, and G. Karthick, "Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM)," *Wirel Pers Commun*, 2021, doi: 10.1007/s11277-021-08580-3.

[21] A. Kumar and S. Kumar, "Sentiment Analysis for Enhancing Student Engagement and Learning Outcomes," in *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)*, IEEE, May 2024, pp. 132–137. doi: 10.1109/INNOCOMP63224.2024.00030.

[22] V. K, P. Samuel, B. V. Krishna, and M. J, "Exploration of sentiment analysis in twitter propaganda: a deep dive," *Multimed Tools Appl*, 2023, doi: 10.1007/s11042-023-17383-6.

[23] N. Khatri, S. Sharma, T. Mukherjee, and A. Dadhich, "Analysis and Design of Twitter Sentiment Analysis Using Improved Machine Learning Approach", [Online]. Available: http://onlineengineeringeducation.com

[24] A. Dey, M. Jenamani, and J. J. Thakkar, "Senti-N-Gram: An n-gram lexicon for sentiment analysis," *Expert Syst Appl*, vol. 103, pp. 92–105, 2018, doi: https://doi.org/10.1016/j.eswa.2018.03.004.

[25] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, "The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation," Apr. 01, 2018, *Association for Computing Machinery*. doi: 10.1145/3185045.

[26] K. K. Putri and E. B. Setiawan, "Jurnal Teknik Informatika (JUTIF) Depression Detection in Indonesian X Social Media Text using Convolutional Neural Networks and Long Short-Term Memory with TF-IDF and FastText Methods," vol. 6, no. 2, 2025, doi: 10.52436/1.jutif.2025.6.2.4206.

[27] Y. Liu, J. Lu, J. Yang, and F. Mao, "Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax," *Mathematical Biosciences and Engineering*, vol. 17, no. 6, pp. 7819–7837, 2020, doi: 10.3934/mbe.2020398.

[28] P. Chapman, "CRISP-DM 1.0: Step-by-step data mining guide," 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:59777418