ISSN 2085-4552

Sentiment Analysis of University X Students: Comparing Naive Bayes and BERT Approaches

Jonathan David[1], Kie Van Ivanky Saputra[1], Andry Manodotua Panjaitan[2], Feliks Victor Parningotan Samosir[3]
[1] Mathematics Department, Faculty of Science and Technology (FaST), Universitas Pelita Harapan, Tangerang, Banten 15811, Indonesia
[2] Industrial Engineering Department, Faculty of Science and Technology (FaST), Universitas Pelita Harapan, Tangerang, Banten 15811, Indonesia
[3] Informatics Department, Faculty of Information Technology (FIT), Universitas Pelita Harapan, Tangerang, Banten 15811, Indonesia
[1] joda48614@gmail.com, [2] kie.saputra@uph.edu, [3] andry.panjaitan@uph.edu, [4] feliks.parningotan@uph.edu

Accepted 12 August 2025
Approved 06 January 2026

***Abstract*— Student satisfaction with university facilities and services requires in-depth analysis to ensure improvements in unsatisfactory facilities or services while maintaining those that meet expectations. This study aims to analyze sentiment in student satisfaction surveys using Natural Language Processing (NLP) methods. Survey data collected from 2022 to 2024 were analyzed using two main approaches: Naive Bayes (NB) with n-grams ($n = 1, 2, 3$) employing feature extraction methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW), and Bidirectional Encoder Representations from Transformers (BERT). The analysis reveals that BERT achieves higher sentiment prediction accuracy than NB, with an F1-score of 0.777 compared to NB's 0.676 (a difference of 0.101), though this improvement margin is not statistically significant. This study also identified keywords for both positive and negative sentiments. These keywords were then analyzed across 11 categories of facilities and services to provide focused insights into aspects that need to be maintained or improved. This study concludes that sentiment analysis provides significant contributions to universities in evaluating and enhancing the quality of facilities and services according to student preferences.**

***Index Terms*— Student Satisfaction; Sentiment Analysis; NLP; NB; BERT; n-gram; TF-IDF; BoW; University Facilities and Services.**

I. INTRODUCTION

Higher education institutions play a pivotal role in cultivating students' soft and hard skills, as well as their competitiveness, by offering a range of facilities and services. When adequate facilities and services are in place, students are empowered to fully actualize their personal potential through various opportunities. The Student Satisfaction Inventory (SSI) is a widely employed instrument in assessing student satisfaction with the array of facilities and services provided by universities. Student satisfaction with university facilities and services is a critical factor that can influence their overall performance and experience [1].

The SSI has been developed as an instrument specifically designed to measure student satisfaction with various aspects of campus life. The SSI was developed by Ruffalo Noel Levitz, an educational consulting and satisfaction measurement tool development organization. The SSI covers various aspects that students consider important, such as the enrollment process, teaching quality, facilities, campus environment, security, effectiveness of academic advising, and others [2].

The analysis of student comments is a complex process, particularly when dealing with substantial quantities of qualitative data. The diversity in the backgrounds and disciplines of students contributes to the complexity of the task, as each student expresses their thoughts and ideas in a unique manner, which introduces subjectivity into the data [3]. This diversity poses a significant challenge in standardizing the analysis and ensuring the validity and reliability of the results [4].

Sentiment analysis employing the neural network (NN) approach processes sentences that fall into the category of unstructured data [5]. NN is applied to process and analyze data using two main approaches: supervised and unsupervised learning. In the supervised learning approach, NN is built and trained using labeled data to classify sentences into positive or negative sentiment categories. The unsupervised learning approach attempts to classify data without the need for labels [6].

The sentiment analysis research will be conducted using a supervised learning approach, as labeled data has been collected in this research. There are various supervised learning approach methods for sentiment analysis, such as Naive Bayes (NB), Support Vector Machine (SVM), Long Short-Term Memory Networks (LSTM), Bidirectional Encoder Representations from Transformers (BERT), and many more. Therefore, this research will compare and analyze the NB and BERT methods on labeled data collected by SSI University X from 2022 to 2024. The NB and BERT methods are applied by finding the best parameters to achieve the highest level of accuracy in performing sentiment analysis. By using data from SSI University X from 2022 to 2024, it is expected that the results and analysis obtained in the research can be in line with reality.

II. THEORY

A. Sentiment Analysis

Sentiment analysis is the process of extracting and assessing the emotional tone of text messages to understand human opinion or behavior. Sentences are analyzed and processed by separating the words to determine whether the sentiment is positive, neutral, or negative. The benefit of sentiment analysis is that it helps in understanding other people's views on a phenomenon [7].

B. Text Preprocessing

Text preprocessing constitutes a critical step in the analysis process, with the objective of averting substantial deterioration in its performance [8]. Text preprocessing is divided into several stages, such as data cleaning, case folding, tokenizing, and stop words removal. The stages involved in this process are described as follows:

1) Data Cleaning: In this stage, the data is cleaned by removing characters such as symbols. Additionally, punctuations and numbers are also removed. The objective of this step is to minimize disruptions in the classification results [9].
2) Case Folding: This stage involves converting text into a uniform format, specifically by converting all text to lowercase [9].
3) Tokenizing: Sentences are broken down into individual words, known as tokens [9].
4) Stop Words Removal: Stop words are words that occur with high frequency but possess minimal semantic significance. These irrelevant common words are identified, flagged, and removed from the text, resulting in a cleaner text corpus [9].

C. Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is a combined method of TF and IDF that produces a combined weight for each term in each document [10]. The formula for calculating TF-IDF is as follows:

$$TFIDF(t_i, D_j) = TF(t_i, D_j)IDF(t_i), \quad (1)$$

$$TF(t_i, D_j) = f_{i,j}, \quad (2)$$

$$IDF(t_i) = \log_{10}\left(\frac{N}{df(t_i)}\right), \quad (3)$$

where $TF(t_i, D_j)$ is the TF of term t_i in document D_j, $IDF(t_i)$ is the IDF of term t_i, t_i is the i-th term, D_j is the j-th document, and $f_{i,j}$ is the number of occurrences of term t_i in document D_j. Index i ranges from 1 to V and index j ranges from 1 to N, where V is the number of unique vocabularies in a set of documents and N the total documents.

D. Bag of Words (BoW)

The Bag-of-Words (BoW) method quantifies word frequencies in a document by disregarding word order. It constructs a dictionary of unique words from a document set and represents each document as a vector, where each element corresponds to a word's frequency. Although BoW ignores word order, it effectively captures topic prevalence and sentiment patterns across documents [11].

E. n-gram

The n-gram method captures word order by analyzing the frequency of consecutive word sequences (defined by n). Unlike BoW, which tracks single words, n-grams generate a dictionary of unique word combinations. Each document is then represented as a vector, where elements indicate the count of these n-grams. This approach preserves contextual relationships between words, offering richer linguistic insights [12].

F. Naive Bayes (NB)

NB classification is a classification method that is both simple and efficient. It is known for its ease of implementation. NB classification is based on Bayes' theorem, where the term "naïve" refers to the assumption that the features in the dataset are mutually independent [13]. The formula for calculating NB is as follows:

$$P(Y = y_j | X = x_i) = \frac{P(X=x_i|Y=y_j) \cdot P(Y=y_j)}{P(X=x_i)}, \quad (4)$$

where:

- x_i: feature vector of sample i, $i \in \{1, 2, \dots, n\}$
- y_j: notation of class j, $j \in \{1, 2, \dots, n\}$
- $P(Y = y_j | X = x_i)$: the probability of sample x_i given a variable that belongs to class y_j.

G. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a pretrained model for English that has been trained on specialized datasets. The BERT model

has been trained on BookCorpus and English Wikipedia, which contains 11,038 unpublished books. Therefore, the BERT model benefits from its pretraining on a large-scale corpus, enabling it to extract richer linguistic patterns and deeper contextual representations compared to traditional models [14].

BERT's architecture builds on the Transformer model, employing stacked encoder layers to process language. Each encoder integrates a multi-head self-attention mechanism that analyzes all tokens in a sequence bidirectionally, capturing nuanced contextual relationships. This is followed by a feed-forward neural network, which applies non-linear transformations to further refine each token's representation. Depending on the variant, BERT uses 12 (Base) or 24 (Large) such layers, enabling it to generate deep, context-aware embeddings. This design makes BERT exceptionally effective for diverse NLP tasks, including sentiment analysis, question answering, and named entity recognition [14].

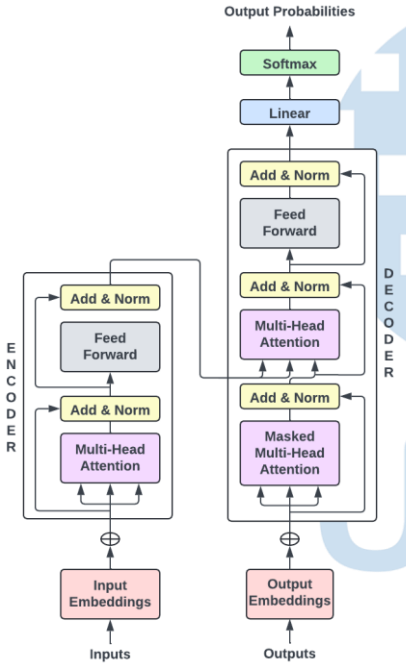

Fig. 1. BERT Architecture

H. Indo-BERT

The Indo-BERT model is distinguished from the standard BERT model in that it is also a pretrained model on the Indonesian language corpus. This signifies that the Indo-BERT model has been trained on a specialized Indonesian dataset, encompassing diverse sources such as online news, social media, Wikipedia, online articles, and recorded video subtitles. Evidently, the Indo-BERT model is replete with Indonesian-specific information and exhibits remarkable capacity to effectively learn from other data sources [15].

I. Hyperparameter Tuning

Hyperparameter tuning is defined as the process of identifying the optimal combination of parameters for a machine learning model. The objective of this process is to ascertain the most effective hyperparameter combination to enhance performance and mitigate the risk of overfitting and underfitting [16]. In this study, for naive bayes model, α as smoothing will be optimized using grid search to ensure that the class-conditional probability value does not equal zero, as this could result in the posterior probability value also being zero. For BERT, the following parameters will be optimized using the grid search method:

- **Learning Rate (LR):** 1×10-5, 2×10-5, 3×10-5, 4×10-5, 5×10-5 because the BERT model requires a low LR. The BERT method is a pre-trained model, and if a low LR value is used, the pre-trained information may be lost and the model may become unstable.
- **Epochs**: 3, 4, 5, 6, 7, because the BERT model is a model that is already rich in information. If you use an epoch value that is too high, it will overfit.
- **Batch Size (BS):** 8, 16, 32 because BERT has high memory requirements. Using multiple BS values helped achieve stable accuracy while maintaining reasonable memory usage and computation time.

J. Confusion Matrix

The Confusion Matrix is a method for calculating the accuracy of a classification model [17]. It is presented as a table showing the number of correct and incorrect classifications for the test data. Then, the accuracy and F1 score can be calculated from the confusion matrix. Accuracy represents the percentage of correctly classified tuples in the test data [18]. It is calculated with the formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \quad (5)$$

The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. Recall is the rate at which positive tuples are correctly identified, and precision is the percentage of tuples labeled positive that are actually positive [18]. It is calculated with the formula:

$$F1score = \frac{2 \cdot precision \cdot recall}{precision+recall}, \quad (6)$$

$$precision = \frac{TP}{TP+FP}, \quad (7)$$

$$recall = \frac{TP}{TP+FN}, \quad (8)$$

where:

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

III. METHOD

This study involves several methodologies and processes, as outlined below:

1) Data Collection

The data used is a set of text documents containing comments made by students at X University about various facilities and services at X University. Facilities and services such as Career Center (CC), Registrar Office (RO), Finance (FIN), Library (LIB), Sports (OR), General Affair (GA), Student Life (SL), Information Technology Service Desk (ITSD), Wifi, Mobile App (APP), and Learning Experience (STUDY). In the data collection of comment text, sentiment information from the comment is available, sentiment can be 1 (positive), 0 (negative). The data collected were 27,659 comments from the X University Student Satisfaction Survey in 2022 to 2024.

TABLE I. SENTIMENT DATA

Sentiment	Comment Count
Positive	22475
Negative	5184

TABLE II. CATEGORY DATA

Category	Comment Count
GA	4084
STUDY	4285
LIB	3541
SL	3551
APP	3267
RO	2816
OR	1637
WIFI	1259
FIN	1386
ITSD	1332
CC	501

TABLE III. LANGUAGE DATA

Language	Comment Count
Indonesia	27406
English	253

2) Data Preprocessing

At this stage of the process, which is referred to as "text preprocessing," a series of critical steps must be taken. Initially, a data cleansing procedure is executed to eliminate non-alphanumeric characters, punctuation marks, and numerals. This is done to avert any potential disruptions in the ensuing classification results. Secondly, case folding involves the conversion of all words and sentences to lowercase, thereby ensuring a uniform text format and eliminating any capitalized words or sentences. The text is then segmented into smaller parts, or tokens, through a process known as tokenization. Finally, in the step of stopword removal, words that are frequently used and have minimal impact on the sentence's meaning, such as "and," "or," and "which," are eliminated. The removal of English stop words will be executed through the utilization of the *NLTK* library, while the removal of Indonesian stop words will be accomplished via the employment of the *re* module.

3) Data Undersampling

The undersampling process is implemented exclusively for Indonesian data due to the significant imbalance in the amount of data for each class category following pre-processing. Undersampling is a necessary procedure to ensure data balance, thereby enhancing the efficacy and performance of the model. In the absence of undersampling, the model's predictions are likely to be influenced by the majority class, as the majority class typically has a higher volume of data. In scenarios involving multiple categories for analysis, it is imperative to ensure that the proportion of each category is balanced. This approach facilitates more effective and equitable learning by the model.

4) Feature Extraction

Feature extraction constitutes a pivotal step following data pre-processing. This process entails the conversion of text or token data into numerical representations, thereby facilitating machine learning processes. Given the limitation of machines and models in comprehending text directly, but rather, their capacity to interpret numerical values, feature extraction emerges as a crucial step. This enables the subsequent interpretation of data by the machine or model, facilitating more profound prediction and analysis capabilities. The methodologies employed encompass n-grams for the Naive Bayes model and word embeddings for the BERT model.

5) Data Splitting

The process of data partitioning is executed in a random manner, involving the allocation of 70% of the data for training, 17.5% for validation, and 12.5% for testing. The data partitioning ratio of 70:17.5:12.5 is employed due to the substantial memory requirements and extended execution time of the BERT model. Subsequent to the undersampling process, the data is segmented into eight non-overlapping datasets, with one dataset allocated for prediction and the remaining datasets utilized for training and validation. It is anticipated that the model will demonstrate the capacity to predict with precision during the testing

phase, contingent on the successful identification of optimal parameters during the training phase.

6) Model Implementation

In this stage, a model is created using data that has undergone the text preprocessing steps. The output from these preprocessing steps is then processed using the Naive Bayes and BERT algorithm. The training of both algorithms is trained to produce the best possible hyperparameter combination for the task at hand. The training process for each model is to be executed independently, given that each model possesses unique steps and characteristics.

7) Keyword Extraction

Keyword extraction will be conducted subsequent to the training and testing process. This procedure will adhere to the same protocol as feature importance, wherein features/words exhibiting the most significant influence on specific class categories will be identified. In the context of NB modeling, keyword extraction can be achieved by calculating the log probability for each feature/word given a class. This will then be incorporated into a vector containing log probabilities for all features. BERT modeling is equipped with an inherent attention mechanism, wherein the attention score is determined during the model training process.

The process of keyword extraction will prioritize the identification of the aspect, disregarding other linguistic elements such as adjectives, verbs, and other non-essential components. To facilitate the sorting of words, an external library will be employed. The library utilized for this purpose is *Spacy* [19] for English and *Stanza* [20] for Indonesian.

8) Evaluation Testing

The trained NB and BERT models have the capacity to utilize the optimal parameters to predict other data. The most effective parameters obtained during the training process are stored in variables. Following the preparation of other data or testing data, the model can be directly applied to predict using the parameters that have been obtained in the training process. Following the execution of the prediction process and the subsequent acquisition of the prediction results, the performance of the two trained models will undergo evaluation. The evaluation of both models will be conducted employing the F1-score metric. The utilization of the F1-score metric is predicated on its capacity to facilitate a fair evaluation, even in scenarios where data is imbalanced.

IV. Results and Discussion

A. Data

Data was collected from 2022 to 2024 based on two languages, Indonesian and English. The data used consisted of answers to open-ended questions in the survey. The answers do not represent a direct assessment or evaluation of the facility/service; rather, they reflect personal opinions expressed in free text format. The data collection process also encompassed 11 distinct keyword categories, namely General Affair (GA), Sports (OR), Registrar Office (RO), Library (LIB), Career Center (CC), Student Life (SL), Learning Experience (Study), WiFi, Mobile App (APP), IT Service Desk (ITSD), and Finance (FIN).

TABLE IV. Data Example

Bahasa	Kategori	Komentar	Sentimen
id	GA	kampus sangat bersih dan tertata dengan rapih.	1
id	GA	Kurang sentuhan hijau di kampus semanggi	0
id	OR	Lengkap dan mudah untuk diakses	1
en	LIB	more hand sanitizers posted please	1
id	WIFI	Ditingkatkan lagi kualitas Wifi nya	0

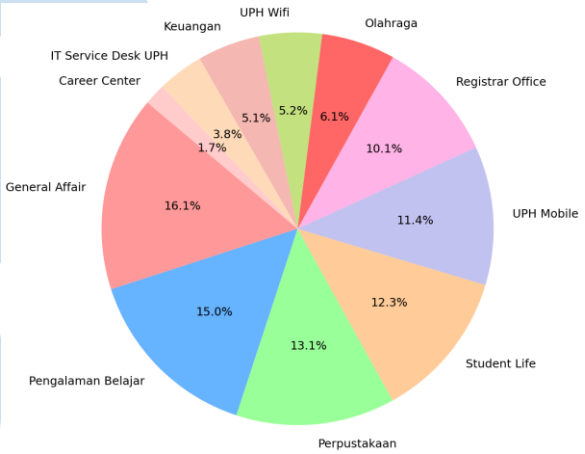

Fig. 2. Category Distribution in the Dataset

B. Text Preprocessing

The data cleaning stage entails the removal of empty sentiments resulting from errors, as illustrated in Table 4. This results in enhanced organization and richer information content in the comment data compared to previous iterations. During the data cleaning process, irrelevant words are systematically removed, which can lead to the removal of comments due to the presence of empty sentiments. These comments are subsequently removed from the data prior to further processing. It is noteworthy that the data cleaning process is meticulously tailored for both the Indonesian and English data sets. Subsequent to the cleansing process, the individual data sets are then seamlessly integrated.

TABLE V. DATA EXAMPLE

Before Preprocessing	After Preprocessing
Layanan dan fasilitas sudah sangat baik	layanan fasilitas sangat baik
The connection sometimes is bad.	connection sometimes bad

C. Undersampling Data

Prior to the integration of the NB and BERT models, an undersampling technique is employed to address the class imbalance present in Indonesian data. Class imbalance arises when the dataset exhibits a significant disparity in the proportion of data between classes, with a disproportionate number of instances belonging to class 1 or class 2, as depicted in Figure 4.2. The undersampling method involves the random selection of data points from the majority and minority classes, thereby ensuring a balanced distribution of data. The effectiveness of this method is evident in Tables 4.6 and 4.7, which illustrate the impact of undersampling on the Indonesian data. However, for the English data, the undersampling process is not employed due to its already substantial and balanced nature.

TABLE VI. DATA COUNT

Indonesian Data Comment Count		
Class	**Before Undersampling**	**After Undersampling**
Positive	15583	2001
Negative	4217	1739
English Data Comment Count		
Positive	147	
Negative	90	

TABLE VII. CATEGORY PROPORTION AFTER UNDERSAMPLING

Indonesian Data Comment Count				
	Before		**After**	
Category	**Negative**	**Positive**	**Negative**	**Positive**
GA	563	2634	170	170
OR	264	935	170	170
RO	329	1676	170	170
LIB	366	2225	170	170
CC	60	284	60	280
SL	268	2173	170	170
STUDY	593	2362	170	170
WIFI	719	301	170	170
APP	667	1581	170	170
ITSD	149	606	149	191
FIN	237	782	170	170

D. Feature Extraction

The subsequent stage of the process involves the extraction or transformation of features from words into numbers. This stage follows the undersampling process. The feature extraction of the NB model will be evaluated through two distinct methods: TF-IDF and BoW, as illustrated in Tables 8 and 9. Subsequently, the feature extraction of the BERT model will be executed using word embeddings.

TABLE VIII. TF-IDF RESULT

bersih	sangat	…	baca	…	kampus	sentimen
0.3807	0.2145	…	0	…	0.3702	1
0	0.2162	…	0	…	0	1
0	0	…	0	…	0	1
…	…	…	…	…	…	…
0	0	…	0.3308	…	0	0

TABLE IX. BOW RESULT

bersih	sangat	…	baca	…	kampus	sentimen
1	1	…	0	…	1	1
0	1	…	0	…	0	1
0	0	…	0	…	0	1
…	…	…	…	…	…	…
0	0	…	1	…	0	0

E. Model Training

In the training phase, the GSCV process is executed on both Indonesian and English data, and the optimal hyperparameter combinations obtained can vary between the two datasets. A total of 18 hyperparameter experiments have been selected for the GSCV test of the NB model, including 1e-6, 1e-5, 1e-4, 0.001, 0.01, 0.0 5, 0.1, 0.5, 1, 2, 5, 10, 20, 50, 100, 200, 500, and 1000. The determination of the optimal hyper-parameters for both datasets is achieved by selecting the highest F1-score, as illustrated in Table 10.

TABLE X. TOP 5 PERFORMING HYPERPARAMETER NB

F1-score							
TF-IDF 1-gram				**BoW 1-gram**			
α	ID	α	EN	α	ID	α	EN
5	0.6921	2	0.7639	5	0.6921	2	0.7639
10	0.6902	5	0.7639	10	0.6902	5	0.7639
2	0.6896	500	0.7639	2	0.6896	500	0.7639
0.5	0.6887	200	0.7639	0.5	0.6887	200	0.7639
0.1	0.6870	1	0.7639	0.1	0.6870	1	0.7639

A total of 125 hyperparameter combination experiments have been selected for the BERT model GSCV test. The BERT model was tested using three different hyperparameters: LR, epochs, and BS. The determination of the optimal hyperparameters for both datasets is achieved the same way like NB, by selecting the highest F1-score, as illustrated in Table 11.

TABLE XI. TOP PERFORMING HYPERPARAMETER BERT

Data	LR	Epoch	BS	F1 Train	F1 Test
DataID 1	$2 \cdot 10^{-5}$	4	32	0.8892	0.7856
DataID 2	$3 \cdot 10^{-5}$	5	32	0.8753	0.8278
DataID 3	$3 \cdot 10^{-5}$	5	8	0.8466	0.8361
DataID 4	$2 \cdot 10^{-5}$	5	16	0.9084	0.8047
DataID 5	$\mathbf{1 \cdot 10^{-5}}$	**5**	**16**	**0.9146**	**0.8491**
DataID 6	$5 \cdot 10^{-5}$	5	32	0.8214	0.8064
DataID 7	$3 \cdot 10^{-5}$	6	8	0.8903	0.8076
DataEN	$\mathbf{1 \cdot 10^{-5}}$	**3**	**8**	**0.8611**	**0.8308**
DataEN	$4 \cdot 10^{-5}$	3	32	0.7938	0.8136
DataEN	$5 \cdot 10^{-5}$	3	16	0.8780	0.8286

F. *Keyword Extraction*

Keyword extraction is the process of identifying words or tokens that exert the greatest influence on sentiment prediction in a given method. This is conducted subsequent to sentiment prediction.Each method employs distinct techniques to identify keywords that impact sentiment prediction.In this instance, the NB method utilizes log probability, while the BERT method employs a feature from its own model, namely attention and attention score. The objective of keyword extraction is to identify efficiently and quickly which facilities/services have been rated as satisfied and dissatisfied in each category.

The NB method of keyword extraction involves the calculation of the log probability of each word or token. The log probability value obtained for a word or token indicates its importance in sentiment prediction. That is, the higher the log probability value, the more significant the word or token is to sentiment prediction. Conversely, the lower the log probability value, the less relevant the word or token is to sentiment prediction. The results and visualization of the NB method of keyword extraction of GA category can be observed in Figure 2 and Figure 3. Other categories and English versions of the keyword extraction are available at https://github.com/j0daaaa/TA_SentimentAnalysis_NLP.

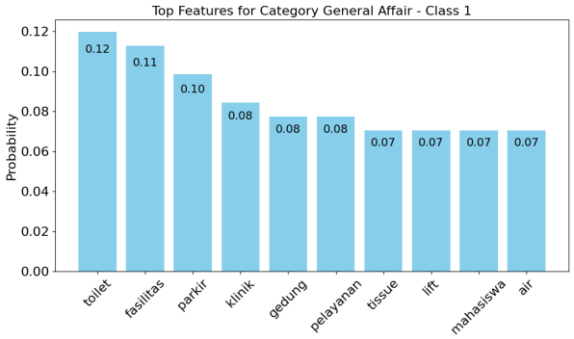

Fig. 3. GA Category Positive Keywords NB 1-gram

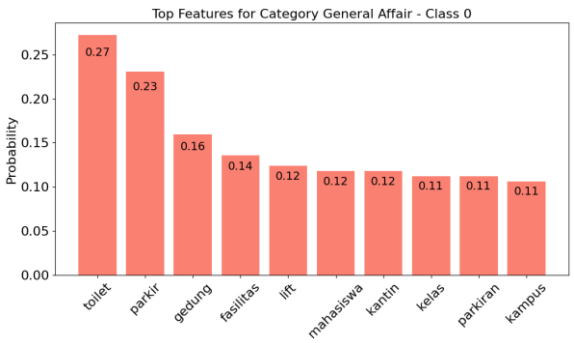

Fig. 4. GA Category Negative Keywords NB 1-gram

The BERT method for keyword extraction entails the extraction of the attention score feature for each word or token in the comment, utilizing the BERT model. Subsequently, the value of the word or token is extracted in its entirety, and the attention score value for a word or token is totaled. The aggregate attention score of a word or token toward positive or negative sentiment is then obtained. This calculation is analogous to the calculation of log probability in the NB method, in that the greater the aggregate attention score value, the more important the word/token is to sentiment prediction. The visualization results of the BERT method keyword extraction of the GA category are presented in Figure 4 and Figure 5.

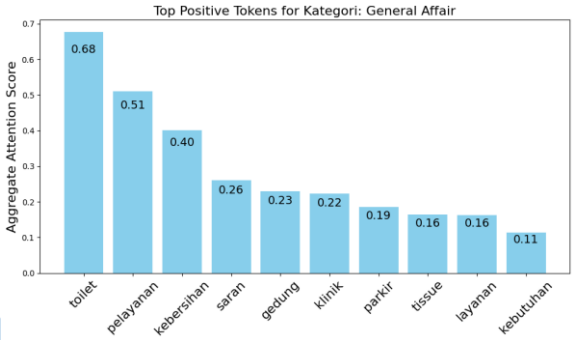

Fig. 5. GA Category Positive Keywords BERT

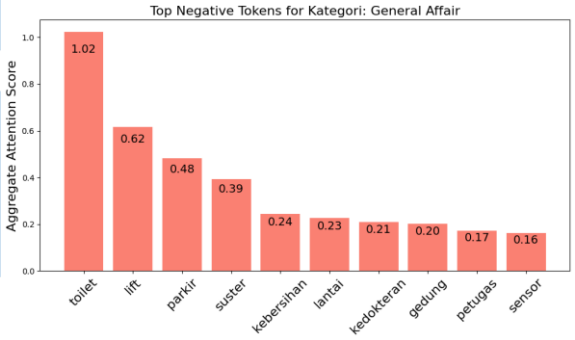

Fig. 6. GA Category Positive Keywords NB BERT

G. *Model Evaluation*

A total of 508 testing data sets were utilized for sentiment prediction, comprising 53 instances of GA data, 38 instances of OR data, 45 instances of RO data, 44 instances of LIB data, 54 instances of CC data, 46 instances of SL data, 52 instances of STUDY data, 49 instances of WIFI data, 31 instances of APP data, and 57 instances of ITSD data, along with 39 instances of FIN data. Predictions have been made using the TF-IDF method with n = 1, 2, 3, the BoW method with n = 1, 2, 3, and the BERT method. A comprehensive summary of the results obtained from all methods employed is provided in Table 12.

TABLE XII. MODEL SUMMARY

Model	CM		F1-score
TF-IDF 1-gram	27	95	0.689977
	39	347	
TF-IDF 2-gram	34	88	0.670839
	59	327	

Model	CM		F1-score
TF-IDF 3-gram	12	110	0.665953
	11	375	
BoW 1-gram	27	95	0.689977
	39	347	
BoW 2-gram	34	88	0.670839
	59	327	
BoW 3-gram	12	110	0.665953
	11	375	
BERT	**83**	**39**	**0.776978**
	116	**270**	

As shown in Table 12, BERT's superior performance over NB remains consistent across all tested values of n, reinforcing its robustness in sentiment analysis tasks. This superiority can be attributed to the BERT model's capacity to effectively handle complex language patterns, a capability that is inherently limited in the NB model due to its assumption of feature independence. The NB model demonstrates greater result instability as n-gram levels increase, leading to diminishing reliability in conclusions. In contrast to the NB model, the BERT model demonstrates a notable enhancement in accuracy. This enhancement can be attributed to its ability to learn complex patterns, understand word context bidirectionally, and effectively handle language elements such as negation or sarcasm.

Furthermore, BERT's pre-training on a substantial and varied text corpus enhances its adaptability and efficacy in sentiment analysis, a capability that is lacking in NB. Deep learning models such as BERT have more accurate predictive performance than machine learning models such as NB. This statement is supported by prior research conducted by Braig et al.[21], where it was found that deep learning models such as BERT or RoBERTa achieve higher predictive accuracy compared to machine learning models such as logistic regression, multinomial naive bayes, and others.

The comments in the suggestion column constitute responses to open-ended inquiries. This constitutes a factor that influences the model's comprehension of the context to be acquired. In the development of the BERT model, there was a decline in the F1-score accuracy of approximately 0.07. This decline is presumably attributable to the characteristic nature of comments, which manifest as open-ended responses.

V. CONCLUSION

The performance effectiveness of the Naive Bayes (NB) and BERT models in sentiment analysis of student satisfaction surveys with mixed and nonstandard languages demonstrates that BERT is superior in capturing sentiment. Following the training of both models and the identification of optimal parameters, BERT attained a prediction accuracy of 0.776978, marginally exceeding the accuracy of 0.689977 achieved by NB with 1-gram, 0.670839 with 2-gram, and 0.665953 with 3-gram. The NB method utilizes the n-gram approach ($n = 1, 2, 3$) with TF-IDF and BoW representations to capture patterns in the data. The primary advantage of BERT lies in its capacity to understand complex language contexts, thereby making it a more reliable choice for sentiment analysis.

The keywords derived from sentiment analysis of student satisfaction surveys, encompassing both positive and negative sentiments, offer a comprehensive representation of students' perceptions regarding various facilities or services. However, the BERT method has been found to outperform the NB method in terms of keyword accuracy. This is primarily due to the presence of equal positive and negative keywords in the NB method, which hinders the ability to draw definitive conclusions. In the context of positive sentiments, keywords such as ”kebersihan”, “layanan”, “court”, “fast respon”, and “pelayanan sangat baik”, reflecting student satisfaction with the facility or service. Conversely, in the case of negative sentiments, keywords such as "toilet", "sinyal", "errors", "kelas karyawan", and "mohon teliti menginput" indicate student dissatisfaction with certain facilities or services. The results of these keywords can be used as evaluation material for the university to identify facilities or services that need to be maintained or improved to increase overall student satisfaction.

Despite BERT’s superior performance, its practical adoption faces challenges. The model’s reliance on large annotated datasets for fine-tuning may limit scalability in resource-constrained scenarios, and its pretraining biases could affect generalizability across domains (e.g., informal text or low-resource languages). Running BERT demands expensive hardware, limiting its use in real-world systems. These constraints suggest that simpler models like Naive Bayes remain viable for tasks where interpretability or efficiency outweighs marginal gains in accuracy.

REFERENCES

[1] B. B. Agubosim, M. M. Arshad, S. N. Alias, and A. Mousavi, “Job satisfaction and job performance among university staff in Nigeria,” International Journal of Academic Research in Progressive Education and Development, 12 (2): 2620-2631. DOI: 10.6007/IJARPED/viz-i2, vol. 17669, 2023.

[2] R. N. Levitz, “National student satisfaction and priorities report,” Cedar Rapids, Iowa: Ruffalo Noel Levitz. Retrieve from RuffaloNL. com/Benchmark, 2017.

[3] B. L. McCombs, “The learner-centered model: From the vision to the future,” Interdisciplinary applications of the person-centered approach, pp. 83–113, 2013.

[4] M. Y.-P. Peng and C. C. Chen, “The effect of instructor’s learning modes on deep approach to student learning and learning outcomes,” Educational sciences: theory & practice, vol. 19, no. 3, 2019.

[5] P. Chaovalit and L. Zhou, “Movie review mining: A comparison between supervised and unsupervised classification approaches,” in Proceedings of the 38th annual Hawaii international conference on system sciences, 2005, pp. 112c–112c.

[6] K. Dave, S. Lawrence, and D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519–528.

[7] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, “Random forest and support vector machine based hybrid approach to sentiment analysis,” Procedia Comput Sci, vol. 127, pp. 511–520, 2018.

[8] M. Syarifuddinn, “Analisis sentimen opini publik terhadap efek PSBB pada twitter dengan algoritma decision tree, knn, dan na\"\ive bayes,” INTI Nusa Mandiri, vol. 15, no. 1, pp. 87–94, 2020.

[9] D. Darwis, N. Siskawati, and Z. Abidin, “Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional,” Jurnal Tekno Kompak, vol. 15, no. 1, pp. 131–145, 2021.

[10] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008. [Online]. Available: https://nlp.stanford.edu/IR-book/

[11] K. Juluru, H.-H. Shih, K. N. Keshava Murthy, and P. Elnajjar, “Bag-of-words technique in natural language processing: a primer for radiologists,” RadioGraphics, vol. 41, no. 5, pp. 1420–1426, 2021.

[12] W. B. Cavnar, J. M. Trenkle, and others, “N-gram-based text categorization,” in Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, 1994, p. 14.

[13] S. Raschka, “Naive bayes and text classification i-introduction and theory,” arXiv preprint arXiv:1410.5329, 2014.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” CoRR, vol. abs/1810.04805, 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[15] B. Wilie et al., “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020.

[16] R. Ahuja, K. Vats, C. Pahuja, T. Ahuja, and C. Gupta, “Pragmatic Analysis of Classification Techniques based on Hyper-parameter Tuning for Sentiment Analysis,” 2020.

[17] A. M. Zuhdi, E. Utami, and S. Raharjo, “Analisis sentiment twitter terhadap capres Indonesia 2019 dengan metode K-NN,” Jurnal Informa: Jurnal Penelitian dan Pengabdian Masyarakat, vol. 5, no. 2, pp. 1–7, 2019.

[18] M. Y. Aldean, P. Paradise, and N. A. S. Nugraha, “Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac),” Journal of Informatics Information System Software Engineering and Applications (INISTA), vol. 4, no. 2, pp. 64–72, 2022.

[19] Explosion, “spaCy: Industrial-strength Natural Language Processing in Python,” 2020. [Online]. Available: https://spacy.io

[20] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages,” arXiv preprint arXiv:2003.07082, 2020.

[21] N. Braig, A. Benz, S. Voth, J. Breitenbach, and R. Buettner, “Machine learning techniques for sentiment analysis of COVID-19-related twitter data,” *IEEE Access*, vol. 11, pp. 14778–14803, 2023