Online-Journals.org

iTDAF IETI Transactions on Data Analysis and Forecasting

iTDAF | eISSN: 2959-0442 | Vol. 3 No. 3 (2025) | OPEN ACCESS

https://doi.org/10.3991/itdaf.v3i3.56453

PAPER

Sentiment Analysis and Topic Modelling for Academic Integrity in the Era of AI

Yovie Adhisti Mulyono, Oscar Karnalim (✉)

Maranatha Christian University, Bandung, Indonesia

oscar.karnalim@it.maranatha.edu

ABSTRACT

This study explores the sentiments and discussion topics of X/Twitter users regarding academic integrity in the era of artificial intelligence (AI). The approach incorporates sentiment analysis and topic modelling to reveal the public perspective on academic integrity issues, including plagiarism, online exams, and AI usage. Our study aims to provide a framework for exploring topics and findings related to the trend of academic integrity in the era of AI. In sentiment classification, Naive Bayes, support vector machine (SVM), and Random Forest algorithms are combined with vectorization techniques such as Count Vectorizer, Word Level TF-IDF, N-Gram TF-IDF, and Character Level TF-IDF. The results show that Naive Bayes with Count Vectorizer provides the best performance on imbalanced data. For the topic modelling, NMF proved to be the most effective in generating specific topics, such as plagiarism and AI detection, with the highest coherence scores. This study also examines the crucial role of each preprocessing step in enhancing data quality, which significantly impacts classification and topic modelling performance. The findings are expected to provide new insights into sentiment analysis and a deeper understanding of academic integrity issues in the era of artificial intelligence.

KEYWORDS

academic integrity, sentiment analysis, SMOTE, topic modelling

1 INTRODUCTION

Integrity is a concept that consistently upholds moral principles and values, regardless of the situation faced [1]. In an academic context, integrity not only includes avoiding dishonesty and plagiarism but also involves a commitment to doing the right thing and feeling proud of honest achievements [2]. A learning environment that upholds academic integrity allows students to develop strong critical, creative, and ethical thinking skills [3].

The development of artificial intelligence (AI) has a significant impact on academia, especially with the emergence of Generative AI (GenAI) such as ChatGPT

Mulyono, Y. A., Karnalim, O. (2025). Sentiment Analysis and Topic Modelling for Academic Integrity in the Era of AI. *IETI Transactions on Data Analysis and Forecasting (iTDAF)*, 3(3), pp. 61–73. https://doi.org/10.3991/itdaf.v3i3.56453

Article submitted 2025-05-06. Revision uploaded 2025-06-28. Final acceptance 2025-07-08.

and Gemini [4]. GenAI allows for a more personalized and adaptive learning approach, making it easier for students to access educational resources according to their needs [5]. However, this convenience also raises concerns about changes in students' academic behavior, such as the increased potential for plagiarism and the use of AI to complete assignments without critically thinking about the solutions [6], [7]. They might obtain high scores without really achieving the learning objectives.

It is important to analyze public perception of academic integrity in the era of AI to understand how society views this phenomenon. X/Twitter was chosen as the data source because of its popularity in Indonesia and its ability to reflect public opinion in real time [8]. This study aims to identify positive, negative, and neutral sentiments related to academic integrity and uncover the main themes of public concerns in the era of AI by applying sentiment analysis and topic modelling techniques.

Sentiment analysis collects public perspectives about a particular topic from the internet [9]. Each data entry is mapped into several sentiment categories (typically positive, neutral, and negative). The analysis itself commonly relies on machine learning algorithms such as Naïve Bayes [10], support vector machine [11], deep neural networks [12], and long short-term memory (LSTM) [13]. Some of them employ further preprocessing such as word embedding [14], term weighting [15], emotion categorization [16], and adverb analysis [17].

Many sentiment analysis studies limit the study to X/Twitter [18], a social media platform mainly consisting of short text (140 characters). The performance of sentiment analysis varies across various topics, and thus, experiments are necessary on each topic. There are studies covering COVID-19 [19] and its vaccines [20]. Some other studies focus on the perspective towards a particular protest [21] or product [12].

Topic modelling aims to discover trends from a set of text documents [22]. It involves machine learning algorithms, including latent dirichlet allocation (LDA) [23] and latent semantic analysis (LSA) [24]. Topic modelling studies cover a broad range of topics, starting from financial bureau data [25] and healthcare [26] to ChatGPT [27]. Similar to the context of sentiment analysis, topic modelling sometimes focuses on X/Twitter as the data source [28].

This study includes the selection of effective data preprocessing stages as a crucial step to ensure the accuracy of the sentiment analysis. The evaluation of the performance of classification algorithms such as Naïve Bayes, support vector machine (SVM), and Random Forest is expected to optimize the process. Each algorithm has its advantages in recognizing patterns and understanding the context. Last but not least, this study will also evaluate the accuracy of topic modelling in identifying discussion topics related to academic integrity in the era of artificial intelligence.

By combining machine learning and topic modelling techniques such as Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and BERTopic, this study is expected to provide a comprehensive picture of sentiment and issues that are developing in society. The results of this analysis are also likely to provide in-depth insights that form the basis for formulating more focused academic policies for maintaining integrity values in the era of artificial intelligence.

Compared to existing studies, ours is the first to combine multiple sentiment analysis classifiers (i.e., Naïve Bayes, SVM, and Random Forest) and topic modelling algorithms (i.e., BERTopic, NMF, and LDA) to capture the diverse perspectives on academic integrity in the era of AI. The study provides a framework for understanding the current trends in academic integrity in the era of AI, and the findings can be useful for readers interested in this topic, which becomes increasingly important as AI significantly impacts academia. Existing studies focus solely on academic integrity, without considering the influence of artificial intelligence.

2 METHOD

This study will analyze sentiment towards academic integrity in the AI era using English tweets from X/Twitter. The research stages include data collection, labeling, preprocessing, and balancing with SMOTE. Classification models are built using Naïve Bayes, SVM, and Random Forest, while key topics are identified through LDA, NMF, and BERTopic.

This study used 2,875 English-language tweets collected through the API X/Twitter from September 2023 to September 2024. The data sample consisted of 40 positive tweets, 265 negative tweets, 2,418 neutral tweets, 36 positive promotion tweets or tweets promoting AI tools, and 116 negative promotion tweets or tweets promoting assignment/exam completion services. Data were collected using the keywords plagiarism, collusion, contract cheating, exam cheating, research fraud, and AI misuse.

2.1 Data preprocessing

Our data preprocessing involves three steps: data crawling, data collection, and data transformation. X/Twitter posts are preprocessed to a table in which the columns refer to unique words, each row refers to a post, and each cell reports a word frequency in a post. The features are word occurrences, while the target class is the sentiment of the post.

Data crawling is the process of automatically retrieving data from various sources that provide APIs to access information widely. In this study, users need an API key, an API Secret Key, an Access Token, and an Access Token Secret from X/Twitter. The crawling data is stored in a CSV file and will be used for training and evaluating sentiment analysis models and topic modelling.

Data collection was carried out using tweet-harvest, a Node.js-based tool that utilizes the X/Twitter API. Users can set search parameters, including search keywords, the number of tweets to be collected, and the CSV file name. During the tweet collection process, the tool may run beyond the query limit allowed by X/Twitter and cause an error. If this happens, the user must wait about 10 minutes before continuing the crawling process. The queries were a combination of AI with academic integrity, plagiarism, collusion, contract cheating, exam cheating, and research fraud.

The collected tweets were then manually labeled according to relevant sentiments and categories. Positive labels for tweets that support the use of AI, especially in the context of academic integrity violations. Negative for disagreement with AI in matters that undermine academic integrity. Neutral for tweets without a clear opinion on academic integrity issues. Additionally, there are two further categories: 1) Promotion Negative, which is for promoting assignment or exam completion services, and 2) Promotion Positive, which is for advertising academic-related AI tools. There are a total of 2872 tweets: 40 positive, 265 negative, 2418 neutral, 36 promotions positive, and the rest are promotion negative.

Data transformation is the initial step in sentiment analysis, preparing the collected data for machine learning models. The stages include data cleaning, normalization, tokenization, stop word removal, and stemming. These processes aim to make the data cleaner and more consistent, thereby increasing the accuracy of the analysis.

Data cleaning eliminates irrelevant elements such as mentions, hashtags, URLs, symbols, emoticons, and punctuation, as well as duplicate data or inconsistent formats. In this process, case folding is also performed to convert all letters

to lowercase, thereby maintaining text consistency more effectively. Normalization changes abbreviations or non-standard words into standard forms, such as “btw” to “by the way.” It is based on an English thesaurus. Tokenization breaks text into tokens (words/phrases) for further analysis. Stopword removal discards all words that do not provide significant meaning, such as “I,” “a,” “by,” and so on. Lemmatization changes words into their basic forms, but more semantically, considering the context, for example, “better” to “good.”

2.2 Entiment analysis

The sentiment classification process begins by dividing the data into multiple folds to maintain a balanced class distribution in each iteration. The text is then converted into a numeric representation through vectorization, and the most relevant features are also selected to improve the model’s performance. If necessary, SMOTE [29] is applied to handle the imbalance of the training data. After that, the machine learning model is trained and tested and then evaluated with various assessment metrics. This process is repeated for each fold, and the result is obtained by averaging the evaluation metrics across all iterations.

At first, the data is split using Stratified K-Fold Cross-Validation (n_splits = 5) to ensure a balanced class distribution in each fold. Of the 2872 tweets, 80% are used for training and 20% for testing.

Secondly, three classification models are employed. Naïve Bayes (MultinomialNB) [30] is the first model. It allows predictions based on word probabilities in documents. Naïve Bayes is selected due to its ability to handle multidimensional and missing data. SVM [31] is the second model, chosen due to its suitability for datasets with fairly clear class separation and high dimensionality. The model is combined with a linear kernel to separate classes with a maximum margin. The kernel was chosen because it is more efficient and can avoid overfitting when compared to polynomial, RBF, or sigmoid kernels. Random Forest [32] is the third model, selected due to its robustness and ability to effectively understand data patterns. As an ensemble model, it combines 100 decision trees (n_estimators = 100) to improve prediction accuracy and stability. This number of trees was chosen to balance accuracy and computational efficiency, without slowing down the training process. All machine learning models employ default parameters according to their corresponding theories.

Third, text data will be then vectorized (i.e., converted) into numeric representation using several vectorization methods, namely Count Vectorizer, which counts the frequency of words in a document; Word Level TF-IDF, which weights words based on their importance in the document with max_features = 5000 for efficiency, N-Gram TF-IDF, which analyzes unigrams and bigrams using ngram_range = (1,2) to capture word context, and Character Level TF-IDF, which will analyze patterns at the n-gram character level (2–3 characters), such as spelling errors or word prefixes/suffixes.

Fourth, data balancing is employed. Each model will be tested with and without SMOTE (Synthetic Minority Oversampling Technique) to handle class imbalance by generating synthetic samples for the minority class. SMOTE is very important in imbalanced datasets, such as in this study, where the Neutral class is much more dominant. With SMOTE (random_state = 42), the class distribution becomes more balanced, so the model can better recognize the minority class more accurately.

Fifth, the features will be filtered and selected. The Chi-Square method selects 500 words with the highest Chi-Square value based on the feature relationship with the class label. This feature selection enhances the focus on distinguishing the most relevant words or n-grams for each class.

Finally, the sentiment classification models will be evaluated. It is performed using the accuracy metrics to measure the proportion of correct predictions, precision to assess the accuracy of the classification, recall to measure the model's ability to recognize all examples in a particular class, and error rate, which indicates the level of prediction error croft [33]. The results from each fold are then averaged to determine the model's final overall performance.

2.3 Topic modelling

After performing sentiment classification, the study continued with training and evaluation using LDA [34], NMF [35], and BERTopic [34], both with and without SMOTE, to find topics in the lemmatized text. LDA is selected due to its ability to identify abstract topics. NMF is chosen due to its ability to interpret effectively. BERTopic is selected due to its ability to handle short text. For comparison, all topic modelling algorithms employ their default parameter setting.

Topic quality will be assessed based on the coherence score, while the majority vote will be used to determine the dominant category based on the original data label. The dataset used consists of 2,297 documents in the training data, with 7,067 TF-IDF vectorized features.

Our topic modelling has several steps. First, the data is divided into smaller sets. StratifiedKFold is used again to split the text data and sentiment labels into five balanced folds. This is done to ensure that each class is represented in each of the folds. Each fold will take turns becoming test data, while the rest is used for training.

Second, topic modelling models will be generated. This study uses three main algorithms for topic modelling: LDA, NMF, and BERTopic. All employ default parameters. LDA assumes documents are a mixture of several hidden topics, with topic distribution calculated using transform() and dominant topics taken with .argmax(axis = 1). NMF decomposes the data into two non-negative matrices representing document-topic and word-topic relationships, with the main topic determined by majority vote. Meanwhile, BERTopic uses SBERT embeddings to group texts based on semantic similarity, generating a number of topics that are then reduced to five through clustering and topic reduction processes. These three methods are used to identify patterns in the data and evaluate the most relevant topics in the text collection.

Third, the data is then balanced with SMOTE. We need to improve the proportion of the minority class before retraining the model. The training process follows the same steps as without SMOTE, with the additional function display_topic_details_smote to ensure that the topic distribution still reflects the original data.

Fourth, the topic details will be displayed. The display_topic_details and display_topic_details_smote functions are used to analyze the topic modelling results from LDA and NMF by displaying and storing the main words and representative documents of each topic. Both of these functions calculate the topic distribution of documents using model.transform(), with the main words identified by the largest weights in model.components_. The difference is that display_topic_details works with the original data, while display_topic_details_smote handles data that has been resampled with SMOTE to ensure that representative documents remain from the original data.

Fifth, the majority_vote_topic function will determine the relationship between topics and sentiment categories by receiving a list of predicted topics and document categories. This function creates a topic_label_mapping dictionary to group labels by topic, then uses Counter to count the frequency of labels in each topic. The label with the highest frequency will be selected as the result of the majority vote then generate a dictionary that maps each topic to the dominant label.

Finally, the generated topic models will be evaluated by displaying the ten main words and representative documents for each topic and calculating the coherence score to assess the semantic relatedness between words. This coherence score is extracted based on the probability distribution (LDA), component matrix weights (NMF), or clustering results (BERTopic). It is calculated using Gensim's CoherenceModel with the c_v method, where higher values indicate more coherent topics. In addition, the distribution of categories in each cluster is also analyzed using the majority vote from topic_results, with the calculate_category_percentages function to calculate the proportion of categories based on the model results and the original labels. For example, if a cluster contains 85 neutral tweets, 10 positive tweets, and 5 negative tweets, then the majority category is determined to be neutral, with a percentage of 85%. This process is applied to all clusters, and the average percentage of the majority category is calculated to assess the suitability of the topic to the original data distribution.

3 RESULTS AND DISCUSSION

3.1 Sentiment analysis without SMOTE

This analysis aims to identify the combination of models and vectorization methods that produces the best performance in terms of accuracy, precision, recall, and error rate. According to Table 1, the Naïve Bayes model performs best with Count Vectorizer, producing high accuracy, precision, and recall. Further observation shows that Count Vectorizer, which is effective in the dominant class (Neutral), fails to recognize minority classes such as Positive and Promotion Positive. The TF-IDF method (Word Level, N-Gram, and Character Level) actually lowers the effectiveness in the minority class, indicating the model's difficulty in handling data imbalance.

Table 1. Naïve Bayes results for various vectorizer methods before SMOTE

Vectorization	Accuracy	Precision	Recall	Error Rate
Count Vectorizer	0.89	0.86	0.89	0.10
Word Level TF-IDF	0.86	0.77	0.86	0.13
N-Gram TF-IDF	0.84	0.75	0.84	0.15
Character Level TF-IDF	0.84	0.78	0.84	0.15

The SVM model without SMOTE tends to be biased towards the majority class (Neutral), which causes low effectiveness on minority classes such as Negative (0.2), Positive (0.04), and Promotion Positive (0.6), although accuracy remains high. Count Vectorizer gives the best results due to simpler and denser features, while Word Level TF-IDF shows the worst performance due to low word weights on minority classes. The general performance of various vectorizers on SVM can be seen in Table 2.

Table 2. SVM results for various vectorizers methods before SMOTE

Vectorization	Accuracy	Precision	Recall	Error Rate
Count Vectorizer	0.88	0.85	0.88	0.11
Word Level TD-IDF	0.88	0.84	0.88	0.11
N-Gram TD-IDF	0.88	0.87	0.88	0.11
Character Level TD-IDF	0.87	0.81	0.87	0.12

The Random Forest model shows a similar pattern to Naïve Bayes and SVM, with the Neutral class performing the best while the minority classes are difficult to classify. Count Vectorizer performs the best, producing simple features that help the model recognize patterns without being distracted by rare words. In contrast, Word-Level TD-IDF performs the worst due to the low weights on the minority class. The overall effectiveness of the Random Forest model can be seen in Table 3.

Table 3. Random Forest results for various vectorizers methods before SMOTE

Vectorization	Accuracy	Precision	Recall	Error Rate
Count Vectorizer	0.88	0.85	0.88	0.11
Word Level TD-IDF	0.87	0.84	0.87	0.12
N-Gram TD-IDF	0.85	0.83	0.85	0.14
Character Level TD-IDF	0.88	0.85	0.88	0.11

3.2 Sentiment analysis with SMOTE

In this section, we analyze the effect of applying the SMOTE technique to handle the class imbalance problem in the dataset in the context of sentiment classification. After applying SMOTE, the Naïve Bayes model performs worse, as seen in Table 4. It still has difficulty distinguishing the Positive and Promotion Positive classes, especially with the TF-IDF methods, which disturbs its probability distribution. Count Vectorizer gives the best results because it is under the probabilistic assumption of Naïve Bayes, while TF-IDF tends to decrease performance due to uneven word weights.

Table 4. Naïve Bayes results for various vectorizer methods after SMOTE

Vectorization	Accuracy	Precision	Recall	Error Rate
Count Vectorizer	0.80	0.86	0.80	0.19
Word Level TF-IDF	0.69	0.85	0.69	0.30
N-Gram TF-IDF	0.68	0.86	0.68	0.31
Character Level TF-IDF	0.69	0.86	0.69	0.30

This phenomenon also occurs on the SVM model. Performance is reduced with SMOTE (refer to Table 5) since it is still difficult to distinguish the Positive and Promotion Positive classes. Word Level TF-IDF performs best by helping SVM identify patterns between classes. At the same time, Count Vectorizer is less effective, and N-Gram TF-IDF is at risk of causing overfitting due to its many features.

Table 5. SVM results for various vectorizer methods after SMOTE

Vectorization	Accuracy	Precision	Recall	Error Rate
Count Vectorizer	0.77	0.84	0.77	0.22
Word Level TF-IDF	0.80	0.87	0.80	0.19
N-Gram TF-IDF	0.75	0.88	0.75	0.24
Character Level TF-IDF	0.77	0.87	0.77	0.22

SMOTE does not really affect the performance of the Random Forest model, though it might help recognize minority classes (see Table 6). Character Level TF-IDF and Count Vectorizer provide the best results. Character Level TF-IDF is effective in capturing spelling variations, while Count Vectorizer helps build simpler decision rules, resulting in the highest accuracy.

Table 6. Random Forest results for various vectorizers methods after SMOTE

Vectorization	Accuracy	Precision	Recall	Error Rate
Count Vectorizer	0.88	0.85	0.88	0.11
Word Level TD-IDF	0.87	0.84	0.87	0.12
N-Gram TD-IDF	0.85	0.83	0.85	0.14
Character Level TD-IDF	0.88	0.85	0.88	0.11

3.3 Topic modelling without SMOTE

This section discusses the training and testing process of topic modeling models on original data without modification or balancing. It aims to evaluate the ability of LDA, NMF, and BERTopic models to identify and group topics based on the relatedness of words in the data. The training and testing results include coherence score analysis and interpretation of the resulting topics.

The results of topic modeling using LDA show that the main topics in the dataset are related to fraud, plagiarism, politics, and academic problems, with a dominance of the neutral class. However, the low coherence score (0.34) indicates that the model has difficulty identifying relationships between words due to data imbalance and noise, so the resulting topics are less structured.

The NMF model generates more focused and structured topics related to fraud, plagiarism, dishonesty related to AI, and politics than LDA. With a higher coherence score (0.68), the topics generated by NMF are more semantically connected but still dominated by the neutral class, indicating limitations in capturing sentiment variations without data balancing.

BERTopic results identified key topics such as plagiarism, academic cheating, political collusion, and misuse of AI technology, illustrating various forms of dishonesty and manipulation. Although the coherence score is relatively acceptable (0.67), the model struggles to distinguish sentiment due to the dominance of neutral categories and ambiguous contexts.

3.4 Topic modelling with SMOTE

The results of LDA topic modeling with SMOTE identified major themes such as plagiarism, fraud, integrity, and academic cheating. However, the resulting topics

were still similar to the model without SMOTE. Although SMOTE helped with the data distribution, the resulting coherence score remained low (0.34). This suggests that the model struggles to capture strong semantic relationships in a noisy dataset.

Applying SMOTE to NMF topic modelling improves topic diversity and balance of sentiment distribution. The model identifies the main themes of Promotion Positive (AI usage), Promotion Negative (AI misuse), and Positive (exam integrity) classes more specifically than without SMOTE. In addition, a higher coherence score (0.81) indicates stronger semantic relatedness, allowing the model to recognize sentiment patterns more accurately.

The results of BERTopic with SMOTE demonstrate a more balanced data distribution, facilitating the model's ability to capture topic patterns with particular sentiments, although most remain neutral. The coherence score of several topics has increased (0.55 on average), but the diversity of contexts in the original data still causes some topics to be less connected. The main themes generated include plagiarism, academic cheating, political collusion, and misuse of AI technology.

3.5 The effectiveness of data preprocessing

Data preprocessing steps are evaluated based on their effectiveness in producing accurate classification and topic modelling of tweet data. For each preprocessing step, we compare the performance of a particular model with and without it. A preprocessing step is considered effective if it improves performance. This analysis employs Naïve Bayes Count Vectorizer for sentiment analysis and NMF for topic modelling. All are conducted without SMOTE.

The text cleaning process does not really affect the performance of Naïve Bayes sentiment classification. It results in comparable accuracy, precision, recall, and error rate. For NMF topic modelling, however, the coherence score increases (from 0.47 to 0.68). By removing noise such as "@," "RT," and URLs, the model can easily recognize patterns and produce more structured and specific topics.

The normalization process improves sentiment classification performance and topic modelling quality, as evidenced by improvements in accuracy (0.88 to 0.89), precision (0.85 to 0.86), recall (0.88 to 0.89), and coherence score (0.66 to 0.68). It also lowers the error rate (0.85 to 0.10). By converting word variations such as "dm" to "direct message," the model more effectively recognizes patterns, produces more accurate predictions, and forms more structured and focused topics.

By breaking text into smaller word units, tokenization improves the accuracy of Naïve Bayes sentiment classification (0.84 to 0.89), as well as the precision (0.77 to 0.86) and the recall (0.84 to 0.89). It also reduces the error rate (0.15 to 0.1). Tokenization also positively affects NMF topic modelling (coherence score 0.64 to 0.68). Tokenization allows the model to recognize sentiment patterns more effectively and form more structured topics that align with the context of academic integrity in the AI era.

Stopword removal improves the performance of Naïve Bayes sentiment classification and NMF topic modelling with better accuracy (0.88 to 0.89), recall (0.88 to 0.89), error rate (0.11 to 0.10), and coherence score (0.61 to 0.68). By removing common yet meaningless words, the model focuses more on relevant words and aims for higher performance.

Lemmatization positively affects the performance of Naïve Bayes sentiment analysis by having higher accuracy (0.88 to 0.89), higher precision (0.84 to 0.86), higher recall (0.88 to 0.89), and a lower error rate (0.11 to 0.1). It ignores several

word variations, helping the model focus more on semantics. It is also helpful for NMF topic modelling, increasing the coherence score from 0.63 to 0.68.

3.6 Discussions

Naïve Bayes, combined with a Count Vectorizer, demonstrated the best performance on the dataset without the use of SMOTE, primarily due to its simplicity and effectiveness in handling textual data. The Count Vectorizer converts text into a matrix of token counts, which aligns well with the Naïve Bayes assumption of feature independence and its probabilistic approach to word occurrence. The absence of SMOTE also preserved the natural distribution of the classes, allowing Naïve Bayes to perform optimally without introducing synthetic data that might have disrupted the underlying word distributions.

After applying SMOTE to address class imbalance, the combination of Random Forest with Character-Level TF-IDF yielded the best performance, highlighting its strength in capturing complex data patterns. Character-level TF-IDF is effective at identifying subtle textual variations, such as misspellings or morphological changes, which enhances feature representation for classification. When paired with Random Forest, a powerful ensemble learning method capable of handling high-dimensional and noisy data, the model benefited from improved generalization and robustness. The synthetic samples introduced by SMOTE provided a more balanced training set, enabling Random Forest to learn richer decision boundaries, which ultimately led to improved classification performance on the imbalanced dataset.

In topic modelling, NMF achieved the highest coherence scores (ranging from 0.714 to 0.870), outperforming both LDA and BERTopic by generating more distinct and meaningful topics. The model was particularly effective at identifying clear thematic groupings. This indicates that NMF was better suited to capturing the underlying structure of the text data, resulting in topics that were both coherent and highly relevant to the subject matter.

The findings of this study have important implications for educators, policymakers, and AI developers. For educators, the identified sentiments and topics highlight growing concerns among students and the public about academic integrity in the era of AI, emphasizing the need to update academic policies, assessment methods, and digital literacy training. Educators should focus more on higher-order thinking. Policymakers can utilize the framework and the findings to develop more effective regulations and ethical guidelines. Students are more exposed to GenAI, and thus regulations are needed to ensure the best learning experiences. For AI developers, the framework can be further expanded to different topics with more diverse algorithms and preprocessing to capture a better overview of the trends.

4 CONCLUSIONS AND FUTURE WORK

This study uses sentiment analysis models and topic modelling to analyze the sentiment and discussion topics of X/Twitter users related to academic integrity in the AI era. Naïve Bayes with a Count Vectorizer showed the best performance on the data without SMOTE, due to its simplicity in handling text and word distribution. After applying SMOTE, the combination of Random Forest with Character Level TF-IDF proved to be the best, due to its ability to handle more complex

data variations. In topic modelling, NMF outperformed LDA and BERTopic with the highest coherence score (0.714–0.870), producing clearer and more relevant topics, such as "AI, plagiarism, and content detection," "cheat, contract, and online exam," and "plagiarism in research and academic assignments."

Our study has a number of limitations that can be addressed in future work. First, it is focused on the topic of academic integrity on the X/Twitter platform. The findings might be different if the source is changed. There is a need to replicate the study on other social media platforms. Second, while our selected models, their tunings, and their preprocessing steps are considered appropriate, we acknowledge that other models with different tunings and preprocessing steps might be helpful to consider.

5 REFERENCES

[1] L. W. J. C. Huberts, "Integrity: What it is and why it is important," *Public Integrity,* vol. 20, pp. S18–S32, 2018. https://doi.org/10.1080/10999922.2018.1477404

[2] G. M. Currie, "Academic integrity and artificial intelligence: Is ChatGPT hype, hero or heresy?" *Semin. Nucl. Med.,* vol. 53, no. 5, pp. 719–730, 2023. https://doi.org/10.1053/j.semnuclmed.2023.04.008

[3] T. Lancaster, "Academic integrity for computer science instructors," in *Higher Education Computer Science*, J. Carter, M. O'Grady, and C. Rosen, Eds., 2018, pp. 59–71. https://doi.org/10.1007/978-3-319-98590-9_5

[4] T. Corbin, P. Dawson, and D. Liu, "Talk is cheap: Why structural assessment changes are needed for a time of GenAI," *Assess. Eval. High. Educ.,* pp. 1–11, 2025. https://doi.org/10.1080/02602938.2025.2503964

[5] O. Karnalim, M. Ayub, and K. Kusbiantoro, "Perspective of AI chatbots in K-12 education," in *2024 IEEE International Conference on Advanced Learning Technologies (ICALT),* 2024, pp. 239–241. https://doi.org/10.1109/ICALT61570.2024.00076

[6] O. Karnalim, H. Toba, and M. C. Johan, "Detecting AI assisted submissions in introductory programming via code anomaly," *Educ. Inf. Technol.*, vol. 29, pp. 16841–16866, 2024. https://doi.org/10.1007/s10639-024-12520-6

[7] M. A. Haque and S. Li, "Exploring chatgpt and its impact on society," *AI and Ethics*, pp. 1–13, 2014.

[8] J. F. Raisa, M. Ulfat, A. Al Mueed, and S. M. S. Reza, "A review on Twitter sentiment analysis approaches," in *2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 – Proceedings*, 2021, pp. 375–379. https://doi.org/10.1109/ICICT4SD50815.2021.9396915

[9] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review,* vol. 55, pp. 5731–5780, 2022. https://doi.org/10.1007/s10462-022-10144-1

[10] P. Gaur, S. Vashistha, and P. Jha, "Twitter sentiment analysis using naive bayes-based machine learning technique," in *Sentiment Analysis and Deep Learning, Advances in Intelligent Systems and Computing,* S. Shakya, K. L. Du, and K. Ntalianis, Eds., vol. 1432, Springer, Singapore, 2023, pp. 367–376. https://doi.org/10.1007/978-981-19-5443-6_27

[11] N. Yadav, O. Kudale, A. Rao, S. Gupta, and A. Shitole, "Twitter sentiment analysis using supervised machine learning," in *Intelligent Data Communication Technologies and Internet of Things*, in Lecture Notes on Data Engineering and Communications Technologies, J. Hemanth, R. Bestak, and J. IZ. Chen, Eds., vol. 57, Springer, Singapore, 2021, pp. 631–642. https://doi.org/10.1007/978-981-15-9509-7_51

[12] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," *Concurr. Comput.,* vol. 33, no. 23, p. e5909, 2021. https://doi.org/10.1002/cpe.5909

[13] V. Yadav, P. Verma, and V. Katiyar, "Long short-term memory (LSTM) model for sentiment analysis in social data for e-commerce products reviews in Hindi languages," *International Journal of Information Technology,* vol. 15, pp. 759–772, 2023. https://doi.org/10.1007/s41870-022-01010-y

[14] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: A review," *Artificial Intelligence Review,* vol. 56, pp. 10345–10425, 2023. https://doi.org/10.1007/s10462-023-10419-1

[15] H. Zhao, Z. Liu, X. Yao, and Q. Yang, "A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach," *Inf. Process. Manag.,* vol. 58, no. 5, p. 102656, 2021. https://doi.org/10.1016/j.ipm.2021.102656

[16] P. Kumar and M. Vardhan, "PWEBSA: Twitter sentiment analysis by combining Plutchik wheel of emotion and word embedding," *International Journal of Information Technology,* vol. 14, pp. 69–77, 2022. https://doi.org/10.1007/s41870-021-00767-y

[17] S. Haider, M. Tanvir Afzal, M. Asif, H. Maurer, A. Ahmad, and A. Abuarqoub, "Impact analysis of adverbs for sentiment classification on Twitter product reviews," *Concurr. Comput.,* vol. 33, no. 4, p. e4956, 2021. https://doi.org/10.1002/cpe.4956

[18] Y. Wang, J. Guo, C. Yuan, and B. Li, "Sentiment analysis of Twitter data," *Applied Sciences,* vol. 12, no. 22, p. 11775, 2022. https://doi.org/10.3390/app122211775

[19] 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. https://doi.org/10.1109/ACCESS.2023.3242234

[20] Z. Bokaee Nezhad and M. A. Deihimi, "Twitter sentiment analysis from Iran about COVID 19 vaccine," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* vol. 16, no. 1, p. 102367, 2022. https://doi.org/10.1016/j.dsx.2021.102367

[21] 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, p. 100019, 2021. https://doi.org/10.1016/j.jjimei.2021.100019

[22] B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki, and H. M. Abdulwahab, "Short text topic modelling approaches in the context of big data: Taxonomy, survey, and analysis," *Artificial Intelligence Review*, vol. 56, pp. 5133–5260, 2022. https://doi.org/10.1007/s10462-022-10254-w

[23] U. Chauhan and A. Shah, "Topic modeling using latent Dirichlet allocation: A survey," *ACM Comput. Surv.,* vol. 54, no. 7, pp. 1–35, 2022. https://doi.org/10.1145/3462478

[24] B. Rupendra Reddy, D. Sai Tharun Reddy, M. C. Sandeep Preetham, A. H. N. Rajasekhar, and R. Subramani, "Comparative study analysis on news articles categorization using LSA and NMF Approaches," in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT),* 2022, pp. 1–6. https://doi.org/10.1109/ICCCNT54827.2022.9984466

[25] S. Vasudeva Raju, B. Kumar Bolla, D. K. Nayak, and K. H. Jyothsna, "Topic modelling on consumer financial protection Bureau data: An approach using BERT based embeddings," in *2022 IEEE 7th International Conference for Convergence in Technology (I2CT),* 2022, pp. 1–6. https://doi.org/10.1109/I2CT54291.2022.9824873

[26] K. Rajendra Prasad, M. Mohammed, and R. M. Noorullah, "Visual topic models for healthcare data clustering," *Evol. Intell.,* vol. 14, pp. 545–562, 2021. https://doi.org/10.1007/s12065-019-00300-y

[27] O. D. Okey, E. U. Udo, R. L. Rosa, D. Z. Rodríguez, and J. H. Kleinschmidt, "Investigating ChatGPT and cybersecurity: A perspective on topic modeling and sentiment analysis," *Comput. Secur.*, vol. 135, p. 103476, 2023. https://doi.org/10.1016/j.cose.2023.103476

[28] L. Corti, M. Zanetti, G. Tricella, and M. Bonati, "Social media analysis of Twitter tweets related to ASD in 2019–2020 with particular attention to COVID-19: Topic modelling and sentiment analysis," *J. Big. Data.*, vol. 9, 2022. https://doi.org/10.1186/s40537-022-00666-4

[29] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for handling imbalanced data problem," in *2021 6th International Conference on Informatics and Computing (ICIC),* 2021, pp. 1–8. https://doi.org/10.1109/ICIC54025.2021.9632912

[30] A. Bhavani and B. Santhosh Kumar, "A review of state art of text classification algorithms," in *Proceedings – 5th International Conference on Computing Methodologies and Communication (ICCMC),* 2021, pp. 1484–1490. https://doi.org/10.1109/ICCMC51019.2021.9418262

[31] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput. Oper. Res.*, vol. 152, p. 106131, 2023. https://doi.org/10.1016/j.cor.2022.106131

[32] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," *Int J. Data Sci. Anal.*, vol. 20, pp. 197–211, 2024. https://doi.org/10.1007/s41060-024-00509-w

[33] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice.* Reading: Addison-Wesley, 2010.

[34] D. Murthy *et al.*, "Categorizing E-cigarette-related tweets using BERT topic modeling," *Emerging Trends in Drugs, Addictions, and Health*, vol. 4, p. 100160, 2024. https://doi.org/10.1016/j.etdah.2024.100160

[35] R. Egger and J. Yu, "A topic modeling comparison between LDA," *NMF, Top2Vec, and BERTopic to Demystify Twitter Posts, Frontiers in Sociology,* vol. 7, p. 886498, 2022. https://doi.org/10.3389/fsoc.2022.886498

6 AUTHORS

Yovie Adhisti Mulyono is a fresh graduate of the Bachelor of Informatics Engineering, Maranatha Christian University, Indonesia. Her interests include data and sentiment analysis (E-mail: 2172005@maranatha.ac.id).

Oscar Karnalim is an Associate Professor from the Faculty of Smart Technology and Engineering, Maranatha Christian University, Indonesia. His interests include data analytics, academic integrity, and learning technologies (E-mail: oscar.karnalim@it.maranatha.edu).