Comparing logistic regression and extreme gradient boosting on student arguments

Tri Wahyuningsih, Danny Manongga, Irwan Sembiring
Department of of Computer Science, Universitas Kristen Satya Wacana, Salatiga, Indonesia

Article Info

Article history:

Received Jan 1, 2024
Revised Jan 27, 2024
Accepted Feb 10, 2024

Keywords:

Argumentative competence
Effectiveness identification
Logistic regression
Multi-output classification
TF-IDF vs. CountVectorizer

ABSTRACT

Identifying the effectiveness level and quality of students' arguments poses a challenge for teachers. This is due to the lack of techniques that can accurately assist in identifying the effectiveness and quality of students' arguments. This research aims to develop a model that can identify effectiveness categories in students' arguments. The method employed involves the logistic regression+XGBoost algorithm combined with separate implementations of term frequency-inverse document frequency (TF-IDF) and CountVectorizer. Student argument data were collected and processed using natural language processing techniques. The research results indicate that TF-IDF outperforms in identifying effectiveness classes in student arguments with an accuracy of 66.20%. The multi-output classification yielded an accuracy of 89.32% in the initial testing, which further improved to 92.34% after implementing one-hot encoding. A novel finding in this research is the superiority of TF-IDF as a technique for identifying effectiveness classes in student arguments compared to CountVectorizer. The implications of this research include the development of a model that can assist teachers in identifying the effectiveness level of students' arguments, thereby improving the quality of learning and enhancing students' argumentative competence.

Corresponding Author:

Tri Wahyuningsih
Department of of Computer Science, Universitas Kristen Satya Wacana
Salatiga, Indonesia
Email: 982022001@student.uksw.edu

1. INTRODUCTION

The development of argumentation identification models poses a significant challenge in artificial intelligence development, especially when dealing with the complexity and diversity of human language structures [1]. Argument identification involves not only understanding individual words but also requires the model's capacity to interpret context, capture nuanced meanings, and recognize relationships between parts of a text. While many advancements have been made in natural language processing (NLP) and machine learning, the development of argumentation identification models still faces several critical challenges [2]. One of them is the diversity in how humans present arguments, ranging from linear structures to more concealed and implicit deliveries. Models must be able to capture this depth and complexity to provide satisfactory results. Additionally, in developing argumentation identification models, it is crucial to address biases that may emerge in the training data. These biases can create less accurate or even harmful models when applied in real-world situations. Therefore, a critical aspect of this model development is ensuring that the training data representation includes diversity that reflects the reality of human argumentation without causing distortion or imbalance.

The use of logistic regression algorithms has been employed in some previous studies [2], and based on these studies, it was found that logistic regression performs well. This provides evidence and also raises questions in this study regarding the accuracy of the algorithm's performance. The method used in this study is a combination of the logistic regression and XGBoost algorithms. Both algorithms were chosen because they can measure the effectiveness and quality of student arguments with high accuracy [3]. Logistic regression is used to predict the effectiveness class in student arguments, while XGBoost is used to enhance the prediction results of the logistic regression algorithm [4]. The use of logistic regression algorithms along with XGBoost as a method for evaluating student arguments is an important step in addressing teachers' challenges in measuring the effectiveness and quality of student arguments.

The study compares the efficacy of the term frequency-inverse document frequency (TF-IDF) and CountVectorizer methods in analyzing student arguments, aiming to aid teachers in evaluating argument quality. TF-IDF gauges word occurrence while mitigating the impact of common terms, whereas CountVectorizer converts text into a numeric format for algorithmic processing. Results demonstrate TF-IDF's superiority in discerning argument effectiveness. This research innovates by addressing the challenge of accurately assessing argument quality, offering a model to support teachers in this task where current tools are lacking. Contributions include the development of a model for assessing argument quality, a comparison of TF-IDF and CountVectorizer methods, and enhancements through machine learning techniques like apostrophes and one-hot encoding.

This research endeavors to construct a model aimed at gauging the effectiveness and quality of student arguments, utilizing machine learning techniques to aid educators in a more streamlined and unbiased assessment process. The primary research inquiries revolve around the feasibility of employing the logistic regression+XGBoost algorithm for this task, the comparison between logistic regression+XGBoost+TF-IDF and logistic regression+XGBoost+CountVectorizer algorithms to determine superiority, and the potential enhancements in model performance through text mining optimization techniques like apostrophes and one-hot encoding. Through the development and evaluation of this model, the study aims to provide educators with a more objective and efficient means of evaluating student arguments, ultimately assisting in their instructional endeavors.

2. RELATED RESEARCH

The presence of several levels regarding the quality of students' arguments, as outlined in the previous section, organizes discussions on related quality based on these areas and focuses on the same approach as the research objectives. This research specifically highlights the methods that have different research approaches used in the field of detecting the quality of student arguments. It begins by presenting relevant approaches in the domain of argument quality detection, followed by model development techniques. It then discusses works on the quality of student arguments in general. Finally, this section outlines works related to model development with comparative analysis.

2.1. Related research on the use of text mining to detect text quality

Text mining, as one branch of data mining, utilizes techniques and algorithms to analyze and extract information from textual data [5]–[9]. The use of text mining in identifying text quality is highly beneficial and relevant, considering that most information is currently conveyed through text [10]–[13]. Some previous studies have employed text mining to measure text quality using methods such as sentiment analysis, opinion mining, and text classification [6], [14]–[17]. The results of these studies vary greatly and still have ample room for improvement [18]–[22]. Several criteria are used to measure text quality, such as accuracy, precision, recall, and F1-score. These criteria provide an overview of how well text mining algorithms identify text quality. In identifying text quality, text mining also employs techniques such as bag of words, n-gram, TF-IDF, and word embedding. These techniques are used to transform text into numerical representations that can be processed by machines [8], [10], [17]. Research related to text mining for identifying text quality also takes into account factors such as context, slang, emotion, and sentiment. These factors significantly influence measuring text quality, necessitating better algorithms for measurement. With research related to the use of text mining to identify text quality, it is hoped that better and more accurate methods for measuring text quality can be discovered.

2.2. Related research on argument quality detection

Argument quality detection is a crucial field in communication and learning sciences. Several studies have been conducted to explore how to identify argument quality. However, many of these studies use manual methods that are time-consuming and inefficient. Therefore, there is a need to develop more effective and efficient tools for identifying argument quality [23]–[26]. Some research on argument quality detection uses machine learning algorithms such as logistic regression and XGBoost [24]. These algorithms can identify the effectiveness and quality levels of arguments more accurately and efficiently than manual methods. One interesting study in this field is the development of sentiment analysis models to identify the effectiveness and quality of student arguments. In this research, machine learning algorithms are used to learn and identify patterns in student arguments and determine the effectiveness and quality of the arguments. Despite many studies related to argument quality detection, there are still many challenges to be addressed. One of the biggest challenges is ensuring that the developed tools have high accuracy and can be widely used by teachers and learners. Therefore, research related to argument quality detection is still evolving to address these challenges and provide more effective and efficient solutions.

2.3. Related research on the development of machine learning models

Machine learning models in sentiment measurement system development are an area undergoing significant development. These related studies aim to develop predictive models and clustering in categorizing sentiment in text or paragraphs [27]–[29]. Some studies have been conducted to develop models for educational aspects, such as assessing student responses [20], [30]–[33]. In some studies, models are also compared with other text mining methods such as TF-IDF and CountVectorizer to evaluate their effectiveness in text mining [22]. Previous research results indicate that models have higher performance levels compared to other methods. There is also research focusing on improving the accuracy and effectiveness of models [13], [18], [21]. This research involves the development of better machine learning models and the application of optimization techniques to improve the results of previous models. This study examines previous research [13], [14], [33] both of which used a dataset similar to this study, namely "feedback prize-predicting effective arguments" on Kaggle. Nevertheless, the goals of this research differ significantly from the works of [13], [14], [34]. Ding *et al.* [13] focused on improving the accuracy and reliability of automatic assessment of the quality of student arguments by using multi-task learning, while this research aims to explore the combination of three tasks: automatic span detection, type prediction, and quality prediction. Meanwhile, Ding *et al.* [14] investigates the influence of prompts in identifying student arguments. Furthermore, Wang *et al*. [15] also adopting a combination of logistic regression and XGB models, uses a different dataset, focusing on the detection of credit card fraud risk in the UCI Public Germany dataset in 2018. Table 1 describes the reseach summary used in this study.

Table 1. Research summary

Research	Dataset	Algorithm	Target	Year
Wang *et al.* [15]	UCI public germany “the German credit data set (1994)”	Linear regression+XGBoost	Credit fraud risk detection	2018
Rahman *et al.* [1]	Twitter “sentiment polarity datasets”	N-Gram, TF-IDF, ensemble	Comparison	2020
Dogra *et al.* [18]	Bank report “banking financial news”	DistilBERT	Evaluation	2021
Yu *et al.* [17]	Questionnaire “Chengde Medical College, China”	TF-IDF + GRU neural networks	Model development	2021
Ding *et al.* [14]	Kaggle “feedback prize - predicting effective arguments”	k-means clustering & TF-IDF	Model development and evaluation	2022
Ding *et al*. [13]	Kaggle “feedback prize - predicting effective arguments”	Logistic regression, BERT, multi-task learning (MTL)	Comparison	2023
Recent research	Kaggle “feedback prize - predicting effective arguments”	Logistic regression, Xgboost, TF-IDF, CountVectorizer	Model development and evaluation	2023

3. METHOD

3.1. Student argument dataset

The dataset, sourced from Kaggle as secondary data, contains argumentative essays written by students in grades 6 to 12 in the United States, originating from Georgia State University (GSU). It includes contextual information gathered from students' questionnaire responses on various essay topics, totaling 36,765 entries. Utilizing this dataset, a model was developed to assess the quality of students' arguments, leveraging text and argument types for predictions and providing feedback for improvement. The choice of English for the model's development stems from its widespread usage, particularly in scientific literature and NLP research, benefiting from available resources and preprocessing techniques. However, efforts to extend the model's applicability to other languages are underway to enhance inclusivity and broaden its impact. For more described on Table 2 related dataset description, meanwhile Table 3 described example of dataset.

Table 2. Dataset description

Column	Explanation
discourse_id	Identification of the discussion containing those arguments.
essay_id	Identification of the essay from the tested discussion.
discourse_text	Text from the argument itself.
discourse_type	Types of arguments, such as evidence, rebuttal, counterclaim, lead, concluding statement, position, and claim.
discourse_effectiveness	Numbers representing the types of arguments, such as 1 for ineffective, 2 for adequate, or 3 for effective.

Table 3. Example dataset

discourse_id	essay_id	discourse_text	discourse_type	discourse_effectiveness
c22adee811b6	007ACE74B050	I think that the face is a natural landform because there is no life on Mars that we have descovered yet	Claim	Adequate

This research employs one dependent variable, namely discourse effectiveness, which is divided into three categories: "effective," "adequate," and "ineffective." This dependent variable reflects the extent to which a student's argument or text is considered effective in the context of discourse analysis. The independent variable used in this study is a combination of discourse text and discourse type, with the rationale that discourse type is combined with discourse text to clarify the type of argument. Discourse type encompasses seven categories: lead, position, claim, CounterClaim, rebuttal, evidence, and concluding statement. Each type of discourse contributes uniquely to constructing the structure and content of an argument. In the context of this research, this variable becomes the main focus in identifying patterns and characteristics of discourse types that can influence discourse effectiveness. The comparison diagram of the distribution of each category is shown in Figure 1, and words with the highest frequency in each category are visualized using a word cloud in Figure 2.

Wordcloud is one of the crucial visualization techniques in text mining, particularly for datasets with two classes. It provides a visual representation of the most commonly used words in each class, aiding in the identification of key words that may differentiate the two classes. Therefore, wordclouds can expedite and simplify the data exploration process, helping researchers understand the characteristics of each class more effectively. Additionally, wordclouds assist researchers in selecting the most relevant and important features when constructing a classification model. Consequently, wordclouds contribute to enhancing the quality and accuracy of text mining analysis on datasets with two classes.

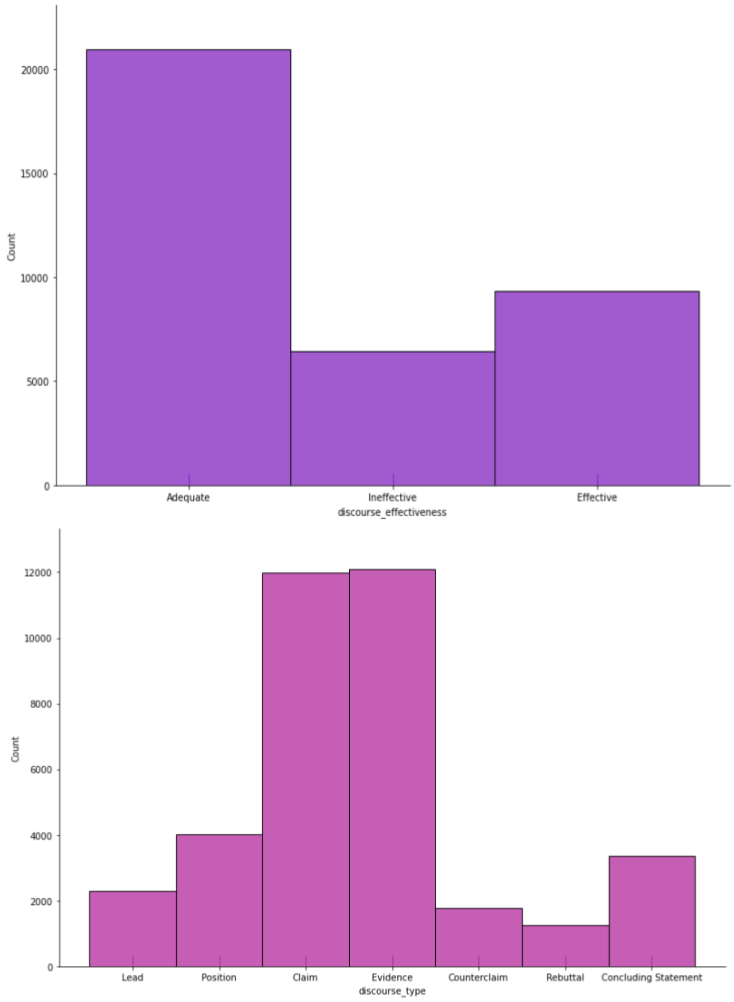

Figure. 1. Class type distribution

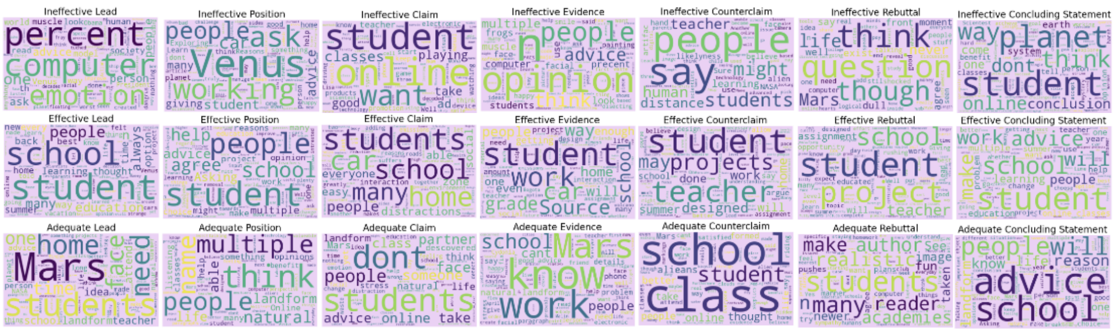

Figure 2. Wordcloud on each type and category

3.2. Research flow

The study employed a systematic research flow, starting with data collection from a Kaggle dataset featuring argumentative essays evaluated by experts for various elements, categorized into argument types and effectiveness classes. Preprocessing involved data cleaning, transformation, integration, and reduction to enhance efficiency, followed by word and text preprocessing utilizing apostrophes and one-hot encoding for better data preparation [30]–[32], [35], [36]. Model evaluation tested three models, including logistic regression combined with XGBoost with TF-IDF and CountVectorizer features to weigh words. The goal was to identify the most accurate model for predicting data classes. Evaluation results validated analysis accuracy through classification, cluster similarity, and topic representativeness, utilizing a multi-output classification model to select the best-performing model and test multi-output predictions. Ultimately, the conclusion stage will present the results of the text mining analysis and summarize the findings of this research in a clear and understandable manner, whether in the form of tables, graphs, or narratives. The steps of this research are illustrated in Figure 3.

Figure 3. Research step

4. RESULTS AND DISCUSSION

4.1. Model development

The comparison of the developed models is focused on identifying answers to the research question. This study investigates how efficiently students' arguments can be detected using:

a) Logistic regression+XGBoost

Combining logistic regression with XGBoost using (1):

$$\hat{y}_{KomA} = \alpha\hat{y}_{LR} + (1-\alpha)\hat{y}_{XGB} \tag{1}$$

Where $\alpha\hat{y}_{LR}$ is the prediction from logistic regression, $\hat{y}_{XGB}$ is the prediction from the XGBoost model, and α is the weight of the specified model.

The weight α will influence the extent to which the model wants to give preference to the predictions from each model. If α=0, the model will only use predictions from the XGBoost model, and if α=1, the model will only use predictions from logistic regression. If α is between 0 and 1, the model will assign different weights to both predictions.

b) Logistic regression+XGBoost+TF-IDF

The combination of logistic regression with XGBoost and TF-IDF involves four equations as (2):

$$\hat{y}_{KomB} = \alpha\hat{y}_{LR} + \beta\hat{y}_{XGB} + (1-a-\beta)\hat{y}_{TFIDF} \tag{2}$$

Where α is the weight assigned to the predictions from the logistic regression model, β is the weight assigned to the predictions from the XGBoost model, and $(1-a-\beta)\hat{y}_{TFIDF}$ is the weight assigned to the predictions from the TF-IDF method.

The weights α, β, and $(1-a-\beta)$ will influence how much the model prefers each prediction source. If α = 1, only predictions from logistic regression will be used, if β=1, only predictions from XGBoost will be used, if $(1-a-\beta)=1$, only predictions from TF-IDF will be used.

c) Logistic regression+XGBoost+CountVectorizer

The combination of logistic regression with XGBoost and CountVectorizer is achieved using (3):

$$\hat{y}_{KomC} = \alpha\hat{y}_{LR} + \beta\hat{y}_{XGB} + (1-a-\beta)\hat{y}_{CV} \tag{3}$$

Where α is the weight assigned to the predictions from the logistic regression model, β is the weight assigned to the predictions from the XGBoost model, and $(1-a-\beta)\hat{y}_{CV}$ is the weight assigned to the predictions from the CountVectorizer method.

The weights α, β, dan (1 - α - β) will influence the extent to which the model prefers each source of predictions. If α=1, then only predictions from logistic regression will be used. If β=1, then only predictions from XGBoost will be used. If $(1-a-\beta)=1$, then only predictions from CountVectorizer will be used.

This research discusses two methods for determining parameters α and β in composite models:

– Manual determination involves optimization techniques like grid or random search across parameter space, offering insights into parameter effects, potentially utilizing complex optimization algorithms.
– Automatic tuning with machine learning utilizes algorithms such as random search or bayesian optimization. While manual determination offers deeper insights into parameter variations, the choice depends on research goals and resource availability. The modeling aims to analyze machine learning algorithm performance across different feature space configurations, as depicted in Table 4, providing clear experimental labels for result differentiation and discussion.

Table 4. The experimental setup is based on the algorithm used

Label experimen	Model
A	Logistic regression+XGBoost
B	Logistic regression+XGBoost+TF-IDF
C	Logistic regression+XGBoost+CountVectorizer

For the purpose of machine learning education, this research utilizes Python, a programming language that supports data analysis and mining using various machine learning algorithms. The objective of employing multiple machine learning algorithms is to examine the consistency of the acquired knowledge. The study conducts three experiments (A, B, and C), employing three classifications for each. The next section presents experimental findings for all label combinations and experimental classifiers, followed by a conclusive discussion on the research findings.

4.2. Model training and testing

In this study, the initial dataset consists of 36,765 annotated argumentative essay data ready for analysis. To optimize computational performance and provide flexibility to model users, the initial dataset is divided into two parts: training data and testing data. This division is done in a 70:30 ratio, where 70% of the total initial data (25,735 data) is used as training data, while 30% (11,030 data) is used as testing data. This separation is carried out automatically and will be tested randomly, aiming to provide flexibility to model users to easily change test data without rearranging training data, thereby facilitating the testing and evaluation process of the model. In this way, the research can test and evaluate the model's performance with previously unseen data, maintaining the integrity of the analysis results and ensuring that the model has good generalization capabilities for new data.

4.3. Model evaluation

Table 5 presents the summary accuracy results of three different experiments using models in this research. Experiment A utilizes a combination of logistic regression and XGBoost models. The accuracy result of this experiment is 58.29%. Precision, measuring the proportion of correctly identified data from a specific class, is 65.05%. Recall, measuring the proportion of actual data from a specific class correctly identified by the model, is 57.93%. The F-measure, combining precision and recall, has a value of 61.28%. Experiment B involves a combined model of logistic regression, XGBoost, and TF-IDF features. This experiment achieves an accuracy of 66.20%. Precision reaches 75.44%, recall is 70.12%, and the F-measure reaches 72.31%. Experiment C also uses a combined logistic regression and XGBoost model, this time with CountVectorizer features. Its accuracy is 66.02%, with precision at 64.11%, recall at 87.32%, and F-measure at 74.12%. The evaluation matrix analysis in the table provides a comprehensive overview of the performance of the three different experiments in this research. Accuracy is a metric measuring the overall alignment of model predictions with actual data. Experiment B with the combined features of logistic regression, XGBoost, and TF-IDF shows the highest accuracy at 66.20%, indicating the model's ability to classify data correctly. Precision, indicating how well the model can correctly identify positive data, is highest in experiment B with a value of 75.44%. Recall, measuring the model's ability to identify all positive data, has the highest value in experiment C at 87.32%. The F-measure, combining precision and recall, shows that experiment B has the highest balance between these two metrics with a value of 72.31%. Overall, experiments B and C stand out with better performance, where experiment C excels in recognizing all positive data (recall), while experiment B demonstrates a good balance between precision and recall. The higher accuracy in both experiments indicates the model's accuracy in classifying the entire dataset.

Table 5. Summary of Model Accuracy Results

	Experiment A (Logistic regression+ XGBoost)	Experiment B (Logistic regression+ XGBoost+TF-IDF)	Experiment C (Logistic regression+ XGBoost+CountVectorizer)
Accuracy (%)	58.29	66.20	66.02
Precision (%)	65.05	75.44	64.11
Recall (%)	57.93	70.12	87.32
F-Measure (%)	61.28	72.31	74.12

4.4. Multi-output classification

Multi-output classification is a text classification method used to categorize text into multiple different classes simultaneously. This method is useful for identifying more than one attribute or category within a text, providing more comprehensive and accurate information. For example, in sentiment analysis research, multi-output classification can be employed to classify a text as positive, negative, or neutral, while also identifying the topics or themes discussed in the text. With multi-output classification, we can analyze text in more detail and obtain more useful information for specific purposes. In this study, the model testing is based on or has a reference in the form of testing data that has been categorized or assigned classes for each argument. The goal of multi-class and multi-output is to use this testing data as a reference for classifying the quality classes based on computational text, resulting in primary classes such as effective, sufficient, or ineffective. The testing data and results used in this study can be seen in Figures 4(a) to 4(c).

In the initial test results, it was found that the model accuracy rate was approximately 89.32%. There were 11.68% prediction errors out of 11,030 tested data. However, after employing the one-hot encoding technique in the second test, it was discovered that the model successfully provided prediction results with an accuracy rate reaching 92.34%. Therefore, it can be concluded that the model used with TF-IDF and one-hot encoding techniques is highly effective in classifying student arguments with a high level of accuracy. In the context of using different data in this model, the integration of additional data allows for highly possible adoption, especially if the dataset is consistent with the same target labels. A model that has been trained with the original dataset can be updated or fine-tuned with additional datasets to enhance its classification capabilities. This approach leverages the diversity of information that may exist in new datasets, enriching the model's understanding of variations in the data. It is expected that the combination of different datasets can help the model improve its generalization and performance. However, it is crucial to emphasize the importance of maintaining consistency in annotations or target labels across the entire dataset so that the model can produce consistent and reliable classification results. By merging different data, it is hoped that the model can achieve better robustness and handle variations that may arise in the context of classification tasks.

Ineffective	Adequate	Effective
1	0	0
0	1	0
0	1	0
0	1	0
…	…	…
0	1	0
0	1	0
0	1	0
0	1	0

(a)

Ineffective	Adequate	Effective
1	0	0
0	1	0
0	1	0
0	1	0
…	…	…
0	1	0
1	1	0
0	1	0
0	1	0

(b)

Ineffective	Adequate	Effective
1	0	0
0	1	0
0	1	0
0	1	0
…	…	…
0	1	0
0	1	0
0	1	0
0	1	0

(c)

Figure 4. Data testing class prediction result: (a) actual data, (b) before using one-hot encoding, and (c) after using one-hot encoding

5. CONCLUSION

The results of developing the model using logistic regression and XGBoost indicate that both methods can be used to predict a text category with fairly high accuracy. The comparison results using TF-IDF and CountVectorizer techniques show that TF-IDF is superior in terms of prediction accuracy. In the initial test, the model achieved an accuracy of approximately 89.32%, but there was a 11.68% prediction error out of 11,030 tested data. In the second test using one-hot encoding technique, the model successfully predicted with an accuracy rate of 92.34%. Therefore, it can be concluded that the model used with TF-IDF and one-hot encoding techniques is highly effective in classifying student arguments with a high level of accuracy. Using TF-IDF, the model can predict a text category more accurately because it considers the importance of a word in a document. This can reduce bias compared to using CountVectorizer, which only counts the occurrences of words in a document. The success of the model development provides a new breakthrough for educators in evaluating and categorizing classes of arguments written by students. This can be further developed by using the data from the model to test patterns or patterns of effectiveness and types of arguments. For future research, the use of other classification algorithms can also be implemented on the same model to test their performance with different classifications.

REFERENCES

[1] S. S. M. M. Rahman, K. B. M. B. Biplob, M. H. Rahman, K. Sarker, and T. Islam, "An investigation and evaluation of N-gram, TF-IDF and ensemble methods in sentiment classification," in *Cyber Security and Computer Science*, Cham: Springer, 2020, pp. 391–402, doi: 10.1007/978-3-030-52856-0_31.

[2] G. N. Harywanto, R. Siautama, A. C. I. Ardison, and D. Suhartono, "Extractive hotel review summarization based on TF/IDF and adjective-noun pairing by considering annual sentiment trends," *Procedia Computer Science*, vol. 179, pp. 558–565, 2021, doi: 10.1016/j.procs.2021.01.040.

[3] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: a comparative study," *Electronics*, vol. 9, no. 3, pp. 1–29, 2020, doi: 10.3390/electronics9030483.

[4] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[5] H. Liu, X. Chen, and X. Liu, "A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis," *IEEE Access*, vol. 10, pp. 32280–32289, 2022, doi: 10.1109/ACCESS.2022.3160172.

[6] M. Kamyab, G. Liu, and M. Adjeisah, “Attention-based CNN and Bi-LSTM model based on TF-IDF and glove word embedding for sentiment analysis,” *Applied Sciences*, vol. 11, no. 23, pp. 1–17, 2021, doi: 10.3390/app112311255.

[7] M. Chiny, M. Chihab, Y. Chihab, and O. Bencharef, “LSTM, VADER and TF-IDF based hybrid sentiment analysis model,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, pp. 265–275, 2021, doi: 10.14569/IJACSA.2021.0120730.

[8] R. S. Patil and S. R. Kolhe, “Supervised classifiers with TF-IDF features for sentiment analysis of Marathi tweets,” *Social Network Analysis and Mining*, vol. 12, no. 1, 2022, doi: 10.1007/s13278-022-00877-w.

[9] S. Fransiska and A. Irham Gufroni, “Sentiment analysis provider by. U on Google Play Store reviews with TF-IDF and support vector machine (SVM) method,” *Scientific Journal of Informatics*, vol. 7, no. 2, pp. 2407–7658, 2020.

[10] A. Mee, E. Homapour, F. Chiclana, and O. Engel, “Sentiment analysis using TF–IDF weighting of UK MPs’ tweets on Brexit,” *Knowledge-Based Systems*, vol. 228, 2021, doi: 10.1016/j.knosys.2021.107238.

[11] F. M. Al-kharboush and M. A. Al-Hagery, “Features extraction effect on the accuracy of sentiment classification using ensemble models,” *International Journal of Science and Research (IJSR)*, vol. 10, no. 3, pp. 228–231, 2021, doi: 10.21275/SR21303123511.

[12] K. S. Kalaivani, S. Uma, and C. S. Kanimozhiselvi, “A review on feature extraction techniques for sentiment classification,” in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 679–683, doi: 10.1109/ICCMC48092.2020.ICCMC-000126.

[13] Y. Ding, M. Bexte, and A. Horbach, “Score it all together: a multi-task learning study on automatic scoring of argumentative essays,” in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 13052–13063, doi: 10.18653/v1/2023.findings-acl.825.

[14] Y. Ding, M. Bexte, and A. Horbach, “Don’t drop the topic-the role of the prompt in argument identification in student writing,” in *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 2022, pp. 124–133, doi: 10.18653/v1/2022.bea-1.17.

[15] M. Wang, J. Yu, and Z. Ji, “Credit fraud risk detection based on XGBoost-LR hybrid model,” in *Proceedings of the International Conference on Electronic Business (ICEB) 2018*, 2018, pp. 336–343.

[16] N. S. M. Nafis and S. Awang, “An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification,” *IEEE Access*, vol. 9, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.

[17] H. Yu, Y. Ji, and Q. Li, “Student sentiment classification model based on GRU neural network and TF-IDF algorithm,” *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 2, pp. 2301–2311, 2021, doi: 10.3233/JIFS-189227.

[18] V. Dogra, A. Singh, S. Verma, Kavita, N. Z. Jhanjhi, and M. N. Talib, “Analyzing DistilBERT for sentiment classification of banking financial news,” *Lecture Notes in Networks and Systems*, vol. 248, pp. 501–510, 2021, doi: 10.1007/978-981-16-3153-5_53.

[19] D. Suhartono and K. Khodirun, “System of information feedback on archive using term frequency-inverse document frequency and vector space model methods,” IJIIS: International Journal of Informatics and Information Systems, vol. 3, no. 1, pp. 36–42, 2020, doi: 10.47738/ijiis.v3i1.6.

[20] J. R. Batmetan and T. Hariguna, “Sentiment Unleashed: Electric Vehicle Incentives Under the Lens of Support Vector Machine and TF-IDF Analysis,” *J. Appl. Data Sci.*, vol. 5, no. 1, pp. 122–132, 2024, doi: 10.47738/jads.v5i1.162.

[21] Riyanto and A. Azis, “Application of the vector machine support method in twitter social media sentiment analysis regarding the covid-19 vaccine issue in Indonesia,” *Journal of Applied Data Sciences*, vol. 2, no. 3, pp. 102–108, 2021, doi: 10.47738/jads.v2i3.40.

[22] S. Qaiser and R. Ali, “Text mining: use of TF-IDF to examine the relevance of words to documents,” *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[23] A. Shankar, C. Jebarajakirthy, and M. Ashaduzzaman, “How do electronic word of mouth practices contribute to mobile banking adoption?,” *Journal of Retailing and Consumer Services*, vol. 52, 2020, doi: 10.1016/j.jretconser.2019.101920.

[24] H. Cho and W. Chiu, “COVID-19 pandemic: consumers’ purchase intention of indoor fitness products during the partial lockdown period in Singapore,” *Asia Pacific Journal of Marketing and Logistics*, vol. 34, no. 10, pp. 2299–2313, 2022, doi: 10.1108/APJML-04-2021-0235.

[25] M. C. G. Davidson, “Does organizational climate add to service quality in hotels?,” *International Journal of Contemporary Hospitality Management*, vol. 15, no. 4, pp. 206–213, 2003, doi: 10.1108/09596110310475658.

[26] H. R. Zeinabadi, “Principal-teacher high-quality exchange indicators and student achievement: testing a model,” *Journal of Educational Administration*, vol. 52, no. 3, pp. 404–420, 2014.

[27] S. Klotz and A. Lindermeir, “Multivariate credit portfolio management using cluster analysis,” *Journal of Risk Finance*, vol. 16, no. 2, pp. 145–163, 2015, doi: 10.1108/JRF-09-2014-0131.

[28] Q. Li, K. H. Cheung, J. You, R. Tong, and A. Mak, “A robust automatic face recognition system for real-time personal identification,” *Sensor Review*, vol. 26, no. 1, pp. 38–44, 2006, doi: 10.1108/02602280610640661.

[29] K. Plangger, M. Montecchi, I. Danatzis, M. Etter, and J. Clement, “Strategic enablement investments: exploring differences in human and technological knowledge transfers to supply chain partners,” *Industrial Marketing Management*, vol. 91, pp. 187–195, 2020, doi: 10.1016/j.indmarman.2020.09.001.

[30] L. S. Riza, A. B. Rachmat, Munir, T. Hidayat, and S. Nazir, “Genomic repeat detection using the Knuth-Morris-Pratt algorithm on R high-performance-computing package,” *International Journal of Advances in Soft Computing and its Applications*, vol. 11, no. 1, pp. 94–111, 2019.

[31] F. Wang, Q. Wang, F. Nie, Z. Li, W. Yu, an9d F. Ren, “A linear multivariate binary decision tree classifier based on K-means splitting,” *Pattern Recognition*, vol. 107, 2020, doi: 10.1016/j.patcog.2020.107521.

[32] Y. Choi, “Finding ‘just right’ books for children: analyzing sentiments in online book reviews,” *Electronic Library*, vol. 37, no. 3, pp. 563–576, 2019, doi: 10.1108/EL-01-2019-0018.

[33] N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, “Automatic short answer grading and feedback using text mining methods,” *Procedia Computer Science*, vol. 169, pp. 726–743, 2020, doi: 10.1016/j.procs.2020.02.171.

[34] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, “A novel text mining approach based on TF-IDF and support vector machine for news classification,” in *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, 2016, pp. 112–116, doi: 10.1109/ICETECH.2016.7569223.

[35] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, “The application of text mining methods in innovation research: current state, evolution patterns, and development priorities,” *R and D Management*, vol. 50, no. 3, pp. 329–351, 2020, doi: 10.1111/radm.12408.

[36] I. A. Orobor and N. O. Obi, “Machine learning pipeline for multi-class text classification,” *International Journal of Engineering Applied Sciences and Technology*, vol. 7, no. 2, pp. 64–69, 2022.

BIOGRAPHIES OF AUTHORS

Tri Wahyuningsih is doctoral student of computer science program at Satya Wacana Christian University. She has a strong interest in information systems management and decided to pursue her doctoral degree in computer science. She has interests and capabilities in data mining and text mining. Before starting her doctoral program, she completed her bachelor and master degrees in informatics engineering. She has working experience as a system analyst and showed excellent achievements during her work. Now, she is focusing on her doctoral studies and working on several research projects in the field of information systems management. She can be contacted at email: 982022001@student.uksw.edu.

Danny Manongga obtained his Bachelor's degree (Drs) from Universitas Kristen Satya Wacana in 1982, followed by a Ph.D. from the University of East Anglia in 1997. His international exposure includes an M.Sc. in IT from Queen Mary College, London, in 1989. Currently serving as a professor and the dean of the Faculty of Information Technology, he brings a wealth of expertise to his administrative and academic roles. Professor Manongga's extensive contributions to research and education underscore his commitment to advancing technology and information studies. He can be contacted at email: danny.manongga@uksw.edu.

Irwan Sembiring earned his Bachelor of Engineering degree in informatics engineering from Universitas Pembangunan Nasional "Veteran" Yogyakarta in 2001, Master of Computer Science in computer science from Gadjah Mada University Yogyakarta in 2004, and Doctorate in computer science from Gadjah Mada University Yogyakarta in 2015. His main research interests are computer network security, and he has done more than 40 publications during his education and teaching career. His research interests include computer network security and computer network designing. Currently, he is active as a lecturer at the Faculty of Information Technology, Satya Wacana Christian University Salatiga. He can be contacted at email: irwan@uksw.edu.