International Journal of Electrical and Computer Engineering (IJECE)
Vol. 15, No. 3, June 2025, pp. 2977~2985
ISSN: 2088-8708, DOI: 10.11591/ijece.v15i3.pp2977-2985

Classification of customer complaints on social media for e-commerce in Indonesia

Achmad Rizki Aditama, Alfan Farizki Wicaksono
Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Article Info

Article history:

Received Jun 13, 2024
Revised Dec 28, 2024
Accepted Jan 16, 2025

Keywords:

Bidirectional encoder representations from transformers
E-commerce
Machine learning
Social media
Text classification

ABSTRACT

The e-commerce industry in Indonesia has experienced rapid growth, especially during the COVID-19 pandemic, which accelerated the shift to online platforms. The market is expected to grow by 105.5% from 2025 to 2030 due to increased internet and smartphone use. As e-commerce expands, companies must improve how they handle customer complaints to build trust and loyalty. Social media is a crucial channel for customer interactions, but it also includes non-complaint messages like positive comments, general questions, and spams that need to be filtered out. This research proposes a machine learning model to automatically classify social media interactions into complaints and non-complaints, focusing on Indonesian-language content. The modeling process utilized 10,600 data points collected from social media X. The best model, a bidirectional encoder representation from transformers (BERT) based classifier, achieved an F1-score of 98.3%. The McNemar test revealed significant performance differences between several models, with the BERT-based model outperforming others. This demonstrates that it is highly effective in distinguishing between complaints and non-complaints, making it a valuable tool for enhancing customer service in Indonesia's e-commerce sector.

Corresponding Author:

Achmad Rizki Aditama
Master of Information Technology, Universitas Indonesia
Faculty of Computer Science Building, UI Campus Depok 16424, Indonesia
Email: achmad.rizki02@ui.ac.id

1. INTRODUCTION

The e-commerce industry in Indonesia has experienced substantial growth in recent years, particularly due to the COVID-19 pandemic, which forced many businesses to move online [1]. The rapid growth of e-commerce has been a key factor in driving the digital transformation of the national economy [2]. The market is anticipated to grow by about 105.5% from 2025 to 2030, driven primarily by the rising number of internet and smartphone users, facilitating easier online shopping [3].

As the e-commerce sector expands, companies must improve their customer handling processes. Quick and effective responses to customer complaints are crucial for enhancing satisfied and loyalty [4]. An integrated customer relationship management system can significantly boost customer satisfaction [5]. One effective strategy is utilizing multiple channels, including social media, to engage with customers. Social media enables customers to quickly file complaints and receive responses, helping companies manage these interactions efficiently. Moreover, social media activities can significantly impact customer satisfaction and engagement, thereby enhancing brand loyalty and public perception [6]. However, social media interactions are not limited to complaints; they also include positive remarks, general queries, and spam, which must be filtered out [7]. Companies need to differentiate actual complaints from other interactions to ensure timely customer feedback management, thereby improving customer satisfaction and operational efficiency.

In addressing these challenges, machine learning can help by automatically sorting interactions into complaints and non-complaints. This makes the company's operations more efficient, as resources are only used for issues needing special attention. We argue that this improves customer interaction management and satisfaction, supporting the growth of e-commerce in Indonesia. Despite this potential, there are some challenges, including technical, operational, and human resource issues. Therefore, it is important to understand how they currently manage social media interactions and which ones need further handling.

Recent studies show that machine learning is effective for classifying customer complaints. Key techniques include obtaining numerical representations of text data using bidirectional encoder representations from transformers (BERT) [8]. For the Indonesian language, alternative models or localized versions of BERT may be necessary. Another approach is generating numerical text representations based on term frequency-inverse document frequency (TF-IDF), which can be combined with various algorithms [9]. Machine learning methods, including support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost), neural networks (NN), and BERT, have been effective in classifying different data classes [10]–[12]. Model evaluation is critical for ensuring reliable predictions. K-fold cross-validation involves dividing the dataset into k-folds, using each fold for testing while training on the others, and repeating this process k times, providing stable performance estimation [13]. A confusion matrix assesses model performance using metrics such as accuracy, precision, recall, and F1-score [14]. Additionally, the McNemar test can compare the performance of two models on the same dataset to determine if differences in prediction errors are statistically significant [15].

This research proposes developing a machine learning model to automatically classify social media interactions as complaints or non-complaints, enhancing resource use effectiveness and social media interaction management. Techniques such as BERT and TF-IDF have proven effective with text data in previous studies. Algorithms like SVM, RF, XGBoost, NN, and BERT have demonstrated efficacy in various classification tasks. Evaluating these models using k-fold cross-validation, confusion matrix, and McNemar test ensures reliable predictions. The key research question is: “To what extent can customer complaint interactions be automatically detected from social media?” This model aims to enhance operational efficiency and support the growth of Indonesia's e-commerce sector.

2. METHOD

The study research design applies the cross-industry standard process for data mining (CRISP-DM) framework, excluding the deployment phase. CRISP-DM is a general reference for data mining, explaining the different stages of the data mining project life cycle. This model is known for its reliability, simplicity, compatibility, flexibility in its application, iterative processes, and efficient time consumption [16]. This study involves several stages: business understanding, data understanding, data preparation, modeling, and evaluation.

Initially, it is crucial to understand the problems this business faces in identifying and addressing customer interactions. In the data understanding phase, social media data is manually annotated. During the data preparation phase, data cleaning involves removing special characters, converting text to lowercase, and eliminating stop words. The cleaned text is then used for feature extraction and converted into numeric vectors. These vectors are used in the modeling stage for classification. The data is split into 5-fold cross-validation and classified using algorithms such as SVM, RF, XGBoost, NN, and BERT. Model performance is evaluated using a confusion matrix and metrics-accuracy, precision, recall, and F1-measure-applying the McNemar test to determine which model best classifies customer complaints.

2.1. Business understanding

There are several limitations encountered in implementing social media as a channel for customer interactions. These limitations include technical aspects, operational challenges, and human resources that are not yet fully optimized to leverage the full potential of social media. Therefore, it is crucial to understand how the company currently identifies and manages customer interactions occurring on social media.

2.2. Data understanding

Through the data collection process, this study utilizes interaction data from one of Indonesia e-commerce accounts on social media X. Data collection was conducted using a crawling method. The data was collected from interactions excluding replies.

2.3. Data preparation

In this stage, irrelevant characters and numbers are filtered out, followed by converting all valid characters to lowercase and removing stop words. The goal is to produce clean text ready for further analysis

[17]. This is crucial as social media data often includes random texts with meaningless words, characters, or emojis. Character filtering removes meaningless characters or emojis and unrelated punctuation marks, using regex to eliminate all punctuation marks. Case folding changes all uppercase letters to lowercase to unify the text using the lower function. Stop words are removed using a rule-based method with a list of irrelevant common words. Feature extraction then transforms this cleaned data into representations suitable for machine learning algorithms by identifying and selecting the most relevant attributes from the raw data. This enhances model performance by converting text data into simpler numerical features, allowing machine learning models to process and analyze the data more effectively [18]. Two methods used for feature extraction are TF-IDF and BERT.

TF-IDF measures how important a word is in a document compared to other documents. Term frequency TF counts how often a word appears, while inverse document frequency (IDF) reduces the weight of common words appearing in many documents. The combination of TF and IDF gives higher scores to significant words in a specific document context [19]. While TF-IDF is popular in traditional machine learning, BERT is an advanced model based on neural networks for automatic feature extraction from textual documents [20]. Traditional methods like TF-IDF process text in a unidirectional manner, either left to right or right to left, potentially missing important contextual information. BERT understands the context of words bidirectionally, allowing for deeper and more accurate text comprehension. BERT uses a transformer architecture with an attention mechanism to contextualize words based on all other words in the sentence and can be used for various text mining applications, including text classification [21].

2.4. Classification modeling

The next stage, classification modeling involves categorizing text into predefined classes or labels using the extracted features. Text classification is a technique in text mining that aims to categorize text into predefined classes or labels [22]. It uses features extracted from the text to learn and identify patterns for classification. This process classifies documents into predefined classes or labels based on patterns formed from previous data [19]. In this study, we set up five different models to find the best one, such as SVM, RF, XGBoost, NN, and transformer-based classifier.

SVM is a popular technique for classifying documents using discriminative classifiers. SVM works by finding the hyperplane that best separates data into classes. Initially used for binary classification tasks, SVMs have now been extended to multi-class problems. They are versatile and can be applied in various data mining areas, including text, images, and videos [13]. RF is an ensemble learning method that combines multiple decision trees to improve classification accuracy. Each tree in the forest is trained using a randomly selected subset of the data. The final prediction is made by averaging the predictions of all the trees. This method reduces overfitting and improves the model's ability to generalize [23].

XGBoost is an advanced ensemble method that uses boosting to improve classification performance. It builds trees sequentially, with each tree fixing the errors of the previous one. It is known for its high efficiency and accuracy, making it a popular choice for many machine learning competitions [24]. NN is designed to learn from data through multiple interconnected layers of nodes or neurons. Each layer processes the input data and passes it to the next layer, gradually extracting higher-level features. Neural network can accommodate complex relationships in data, making them suitable for a wide range of classification tasks [13].

Transformer is one of neural network architecture designed to handle sequential data, such as text, more effectively. It implements a mechanism called self-attention to weigh the importance of unique words in a sentence, allowing the model to understand context better. BERT is an encoder part of the transformer model that processes text bidirectionally, providing an accurate vector representation for a textual document [20], [25]. Additionally, we use the pre-trained BERT Bahasa (indobert-base-p1) can be used not only for feature extraction but also as an end-to-end classification model by adding a feed-forward neural network (FFN) layer on top of BERT [26].

Each modeling algorithm is run five times using the k-fold method. This method is used to ensure that the model's performance is consistent and reliable by averaging the results over multiple training and testing splits, reducing the variance associated with a single train-test split [27]. The entire dataset is divided into five versions of training and testing data. In the first iteration, the first ⅕ of the data will be the test data and the rest will be the training data. This process repeats until the last ⅕ of the data becomes the test data and the rest is training data. The prediction results from the test data in each iteration for each model are combined and evaluated based on the original dataset.

2.5. Evaluation

The evaluation of classification models utilizes the confusion matrix method, a table used to assess the performance of classification models by displaying predictions in rows and actual decisions in columns, consisting of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Commonly used metrics include accuracy, precision, recall, and F1-score. These evaluation metrics complement each

other to provide a comprehensive picture of the model's performance. Accuracy gives an overall view (1), precision informs about the accuracy of positive predictions (2), recall measures the model's ability to identify true positives (3), and F1-score combines precision and recall for an overall evaluation (4).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

$$F1 = (\frac{(2*precision*recall)}{(precision+recall)}) \quad (4)$$

Additionally, McNemar test is a statistical method used to compare the performance of two classification models on the same dataset. A threshold of 0.05 is commonly used to determine statistical significance in hypothesis testing, including McNemar test, where a p-value of less than 0.05 indicates a statistically significant difference between model performances. This method is suitable when data is categorized as correct or incorrect predictions by both models. McNemar test calculates the difference between the number of errors made by both models and tests for statistical significance, helping to determine if the difference in predictive errors between the two models is significant.

3. RESULTS AND DISCUSSION

Several stages were conducted to examine the classification of customer complaints on social media. Initially, customer interaction data from social media was collected and manually labeled. The preprocessing stage involved cleaning the data by removing irrelevant characters, converting all text to lowercase, and eliminating stop words. Cleaned text was then converted into numerical vectors for feature extraction using TF-IDF and BERT. These vectors were used in the classification modeling stage, where the data was split using 5-fold cross-validation and classified using five different models: SVM, RF, XGBoost, NN, and BERT. Each model was run multiple times using the k-fold method for consistency and reliability. Finally, model performance was evaluated using a confusion matrix and metrics like accuracy, precision, recall, F1-measure, and the McNemar test to determine the best model for classifying customer complaints.

3.1. Data collection

This study uses interaction data from two social media accounts on X, both belonging to a single Indonesian e-commerce company. One account is dedicated to handling complaints (account A), while the other serves as the general company account (account B). Data collection was conducted using a crawling method, focusing on Indonesian-language content due to its relevance to the Indonesian e-commerce sector.

The results were manually annotated based on predefined criteria to ensure consistent labeling of complaints and non-complaints. The annotation process involved three trained annotators, who are typically responsible for handling and processing customer complaints, ensuring that their expertise was applied in accurately labeling the social media interactions. Inter-annotator agreement was calculated to ensure reliability and consistency in the labeling process. During the collection period, 10,792 interactions were gathered, of which 6,220 were labeled as complaints and 4,572 as non-complaints. After filtering using whitelisted keywords, the final dataset comprised 6,204 complaints and 4,395 non-complaints, as shown in Table 1.

Table 1. Data distribution

Interaction	Complaint	Non-complaint
Mentioning account A	2,649	789
Mentioning account B	1,692	3,693
Mentioning both account	1,870	89

3.2. Model performance

Based on the results in Tables 2 and 3, it can be concluded that BERT model with a FFN classifier is the best-performing model. This model achieves an impressive overall F1-score of 97.69% and an accuracy

of 97.75%, highlighting its effectiveness in both feature extraction and classification tasks. For the complaint class, BERT-FFN achieves an F1-score of 98.08%, significantly outperforming other models, demonstrating its superior ability to accurately identify and classify complaint instances. In the non-complaint class, BERT-FFN also achieves an F1-score of 97.29%, indicating a balanced performance across different classes and ensuring high precision and recall rates.

Comparatively, other models such as TF-IDF with SVM, RF, NN, and XGBoost show commendable performance but fall short of BERT-FFN. TF-IDF with SVM, for instance, achieves an overall F1-score of 93.02% and an accuracy of 93.28%, while TF-IDF with RF scores an overall F1-score of 91.86% and an accuracy of 92.20%. Similarly, TF-IDF with NN records an overall F1-score of 91.56% and an accuracy of 91.85%, and TF-IDF with XGBoost achieves an overall F1-score of 92.05% and an accuracy of 92.30%. The size of the data used for training and evaluation is consistent across models.

From McNemar test results, significant differences were found between several model pairs, indicating their performance differences. The threshold of 0.05 is commonly used to determine statistical significance in hypothesis testing, including McNemar test, where a p-value less than 0.05 indicates a statistically significant difference between model performances [28]. The McNemar test results can be seen in Table 4.

Table 2. Model's performance for each class

Feature extraction	Classifier	Complaint				Non-complaint			
		Precision	Recall	F1-score	Size data	Precision	Recall	F1-score	Size data
TF-IDF	SVM	92.59%	96.23%	94.37%	6,204	94.37%	89.13%	91.67%	4,396
TF-IDF	RF	90.93%	96.28%	93.53%	6,204	94.27%	86.45%	90.19%	4,396
TF-IDF	NN	91.93%	94.36%	93.13%	6,204	91.73%	88.31%	89.99%	4,396
TF-IDF	XGBoost	92.80%	94.15%	93.47%	6,204	91.57%	89.70%	90.62%	4,396
BERT	FFN	98.04%	98.13%	98.08%	6,204	97.36%	97.23%	97.29%	4,396

Table 3. Overall model's performance

Feature extraction	Classifier	All class				
		Precision	Recall	F1-score	Accuracy	Size data
TF-IDF	SVM	93.48%	92.68%	93.02%	93.28%	10,600
TF-IDF	RF	92.60%	91.36%	91.86%	92.20%	10,600
TF-IDF	NN	91.83%	91.33%	91.56%	91.85%	10,600
TF-IDF	XGBoost	92.19%	91.92%	92.05%	92.30%	10,600
BERT	FFN	97.77%	97.68%	97.69%	97.75%	10,600

Table 4. McNemar test result

Model pair	p-value
SVM and RF	< 0.05
SVM and NN	< 0.05
SVM and XGBoost	< 0.05
SVM and BERT	< 0.05
RF and NN	0.21
RF and XGBoost	0.62
RF and BERT	< 0.05
NN and XGBoost	0.11
NN and BERT	< 0.05
XGBoost and BERT	< 0.05

Significant differences (p-value < 0.05) were observed between several model pairs: SVM and random forest, SVM and NN, SVM and XGBoost, SVM and FFN, random forest and FFN, NN and FFN, and XGBoost and FFN. This means these model pairs differ significantly in performance. On the other hand, no significant differences (p-value > 0.05) were found between random forest and NN, random forest and XGBoost, and NN and XGBoost. This indicates that these model pairs perform similarly in classifying the data. In other words, the effectiveness of these models in classification tasks is comparable, and choosing between them may depend on factors like computational efficiency or ease of implementation.

3.3. Model implementation

Based on the best classification model, the implementation can use probability values, confidence scores, and thresholds to decide if the automated results are reliable or need manual review. This helps to reduce FP, FN, and improve overall accuracy and precision. As shown in Figure 1, the error distribution for

FP is visualized using kernel density estimation (KDE). Figure 2 illustrates the KDE visualization for FN. These visualizations make it easy to see where the model works well and where it needs improvement, highlighting areas with high error density and identifying patterns that might require further tuning.

The KDE diagram above helps identify where the probability of the model is likely to make mistakes, allowing for better decision-making when reviewing uncertain predictions. This approach can improve the overall performance of the classification system. Table 5 shows the statistical measures of the model for predicting the chances of FP and FN.

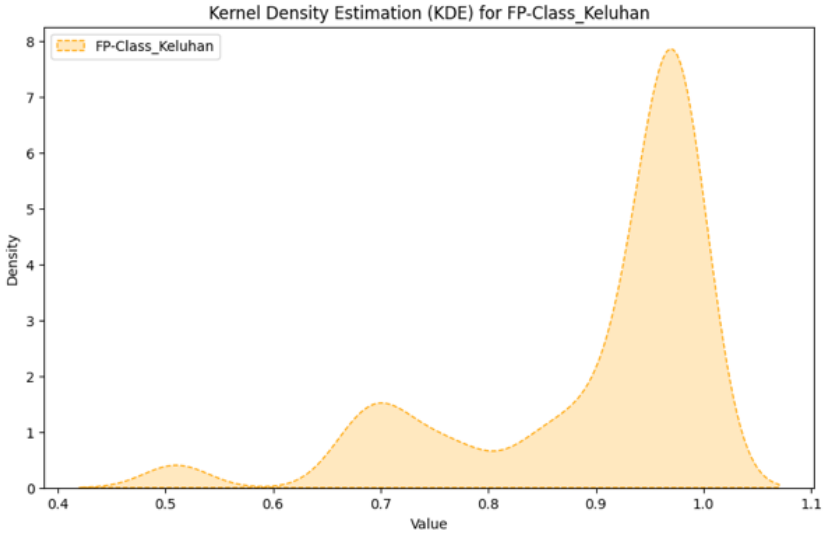

Figure 1. KDE for FP based on probability

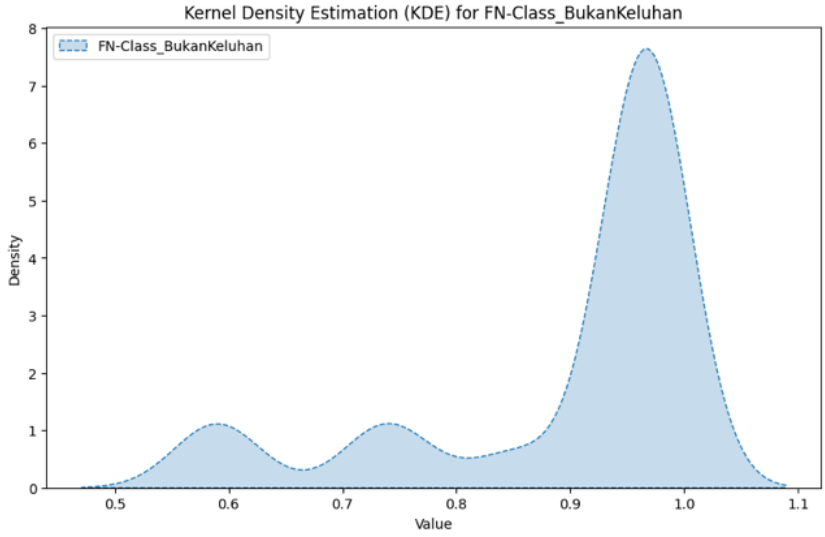

Figure 2. KDE for FN based on probability

Table 5. Statistical values of FP and FN

	FP	FN
mean	0.89	0.89
min	0.51	0.58
p25	0.86	0.89
p50	0.96	0.96
p75	0.98	0.97
max	1.00	1.00

To determine the optimal probability threshold, an analysis is needed to balance the trade-offs between different types of errors and the benefits of TP and TN outcomes. For complaints, based on the statistical values, a higher threshold, around the median FN value of 0.96 or higher, is recommended to minimize FN. For non-complaints, lowering the probability threshold helps in correctly identifying non-complaints, thereby reducing FP and increasing TN. A lower threshold, around the 25th percentile FP value of 0.86 or lower, is recommended to minimize FP.

These thresholds ensure a balanced trade-off between different types of errors and enhance the overall performance of the classification model. Additionally, fine-tuning these thresholds based on the specific context and the importance of different error types can further improve the model's accuracy and reliability. Regularly reassessing and adjusting the thresholds as new data becomes available will help maintain optimal performance over time. Furthermore, incorporating domain expertise in setting these thresholds can provide valuable insights, ensuring the model aligns with practical business needs and objectives.

In implementing the model in real-world settings, the integration with existing customer relationship management (CRM) systems is feasible through application programming interface (API). This integration facilitates real-time complaint identification and response, enabling companies to address customer issues promptly and effectively. The model is expected to process input text in the form of sentences obtained from various social media platforms, allowing for a thorough classification of interactions as either complaints or non-complaints.

The desired output from this model is a confidence score for each of the defined labels, which quantifies the model’s certainty regarding its classification. Based on these confidence scores, the CRM system can make informed decisions on whether to act on the predictions, adhering to the recommended threshold for classification accuracy. If the confidence score falls below this predetermined threshold, the interaction will be routed to the current annotators for further review, indicating that the model has not yet reached the necessary level of accuracy to classify the interaction reliably. This process not only enhances the efficiency of customer service operations but also ensures that complex or ambiguous cases are handled appropriately by human experts, maintaining a high standard of customer care.

4. CONCLUSION

From the above research and analysis, the following conclusions can be drawn. Using pre-trained BERT Bahasa (indoor-base-p1) as a feature extractor and classifier gives the best model, which achieved an F1-score of 98,3%. The best model can detect actual complaint and non-complaint interactions from the training dataset effectively. The result of the McNemar test indicates that SVM behaved significantly differently compared to random forest, NN, XGBoost, and BERT. Similarly, random forest, NN, and XGBoost showed an immense difference from BERT. Meanwhile, the performances for random forest and NN, random forest and XGBoost, and NN and XGBoost did not have a statistical difference in performance, showing a similarity in their process.

From further analysis, mostly the error comes from FP and FN classification. Some of complain data is being classified as non-complaint, and vice versa. This happens due to the lack of adequate context to know the intent of the interaction. Moreover, there are a few cases of unclear labeling in the dataset because common words in the opposite label exist. To tackle this in the implementation phase, probability values can be used to determine whether the automated results should be trusted or required a manual review. This ensures that the classification is done properly and minimizes the potential for errors. It is essential to find such a threshold for the balance of errors and the benefit to TP and TN. However, for complaints, a higher threshold of around 0.96 should minimize FN, while for non-complaints, a lower threshold of around 0.86 makes FP lower. These thresholds compromise different types of errors and boost overall performance.

The suggestion for further research includes addressing more features to support the classification of complaints and non-complaints, as users on social media can discuss a wide variety of topics. Expanding the classification to include other types of interactions, such as suggestions for improvements or compliments on features, would provide a more comprehensive understanding of user interactions. To improve generalizability, future research should use data from multiple companies within Indonesia's e-commerce. Implementing these suggestions will enhance the model's effectiveness and improve classification accuracy across various types of social media interactions, leading to deeper insights. The classification categories can be defined based on the primary problems or needs identified in the case study.

ACKNOWLEDGEMENTS

The author declares that there are no conflicts of interest. This research did not receive any specific grant or financial support from funding agencies in the public, commercial, or not-for-profit sectors. All data

and materials used in this study were obtained without any external financial influence or bias, ensuring the integrity and impartiality of the research findings.

ACKNOWLEDGMENTS

The author declares that there are no conflicts of interest. This research did not receive any specific grant or financial support from funding agencies in the public, commercial, or not-for-profit sectors. All data and materials used in this study were obtained without any external financial influence or bias, ensuring the integrity and impartiality of the research findings.

FUNDING INFORMATION

The author states that no funding was involved for this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Achmad Rizki	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Aditama Alfan Farizki Wicaksono	✓	✓		✓						✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The author declares that there are no conflicts of interest.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, ARA. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

[1] S. D. Negara and E. S. Soesilowati, "E-commerce in Indonesia: impressive growth but facing serious challenges," *ISEAS Perspective*, no. 102, pp. 1–14, 2021.
[2] A. Fashola and F. Kusuma, "E-Commerce Development for the Digital Economy in Indonesia," *Activa Yuris: Jurnal Hukum*, vol. 4, Aug. 2024, doi: 10.25273/ay.v4i2.20842.
[3] Mordor intelligence, "Indonesia e-commerce market size & share analysis-growth trends & forecasts (2025-2030)," *Mordor intelligence*, 2024. Accessed: Feb. 14, 2025. [Online]. Available: https://www.mordorintelligence.com/industry-reports/indonesia-ecommerce-market
[4] T. Bayır and S. Bozyiğit, "Consumer complaints management in the digital era," in *Proceedings of the 2023 International Conference on Consumer Behavior and Marketing*, Dec. 2023, pp. 57-90, doi: 10.4018/979-8-3693-0428-0.ch003.
[5] D. S. M. R. Kurniawan, "The effect of complaint handling on customer loyalty and its impact on customer satisfaction (study on consumers of PT Telkom Witel of West Kalimantan)," *East African Scholars Journal of Economics, Business and Management*, vol. 6, no. 1, pp. 11-16, Jan. 2023, doi: 10.36349/easjebm.2023.v06i01.002.
[6] K. Jamil, L. Dunnan, R. F. Gul, M. U. Shehzad, S. H. M. Gillani, and F. H. Awan, "Role of social media marketing activities in influencing customer intentions: A perspective of a new emerging era," *Frontiers in Psychology*, vol. 12, Jan. 2022, doi: 10.3389/fpsyg.2021.808525.
[7] Shopify, "Omnichannel ecommerce: what it is and how to implement it," *Shopify*, 2024. https://www.shopify.com/enterprise/blog/omnichannel-ecommerce (accessed Mar. 24, 2024).
[8] A. R. Manjrekar, S. S. Gokhale, and R. A. Wilson, "Detection of anti-human rights discourse from Colombian social media conversations using advanced transformer models," in *Proceedings - 22nd IEEE International Conference on Machine Learning and Applications, ICMLA 2023*, Dec. 2023, pp. 1501–1507, doi: 10.1109/ICMLA58977.2023.00226.
[9] M. Khalil and M. Azzeh, "Truth seeker of the largest social media content using machine learning algorithms," in *Proceedings* -

22nd IEEE International Conference on Machine Learning and Applications, ICMLA 2023, Dec. 2023, pp. 1606–1610, doi: 10.1109/ICMLA58977.2023.00243.
[10] J. Hussey, K. Stone, and T. Camp, "Positive and unlabeled learning for mobile application traffic classification," in *Proceedings - IEEE Military Communications Conference MILCOM*, Nov. 2022, vol. 2022-Novem, pp. 25–30, doi: 10.1109/MILCOM55135.2022.10017699.
[11] N. Thair Ali, K. Falih Hassan, M. Najim Abdullah, and Z. Salam Al-Hchimy, "The application of random forest to the classification of fake news," *BIO Web of Conferences*, vol. 97, p. 49, 2024, doi: 10.1051/bioconf/20249700049.
[12] K. Purwandari, R. B. Perdana, J. W. C. Sigalingging, R. Rahutomo, and B. Pardamean, "Automatic smart crawling on Twitter for weather information in Indonesia," *Procedia Computer Science*, vol. 227, pp. 795–804, 2023, doi: 10.1016/j.procs.2023.10.585.
[13] C. M. Bishop, *Pattern recognition and machine learning*. Springer New York, 2006.
[14] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*, 2nd ed. Pearson, 2020.
[15] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *arXiv:1811.12808*, Nov. 2018, doi: 10.48550/arXiv.1811.12808.
[16] P. Chapman *et al.*, "CRISP-DM 1.0: Step-by-step data mining guide," *SPSS inc*, vol. 78, pp. 1–78, 2000.
[17] S. Wankhede, R. Patil, S. Sonawane, and P. A. Save, "Data preprocessing for efficient sentimental analysis," in *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, 2018, pp. 723–726, doi: 10.1109/ICICCT.2018.8473277.
[18] T. B. Brown *et al.*, "Language models are few-shot learners," *arXiv:2005.14165*, May 2020.
[19] Md, "Blending weighted TF-IDF BERT for improving semantic search," in *ICARC 2022 - 2nd International Conference on Advanced Research in Computing: Towards a Digitally Empowered Society*, Feb. 2022, pp. 154–159, doi: 10.1109/ICARC54489.2022.9753875.
[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, Oct. 2018.
[21] H. Fan and Y. Qin, "Research on text classification based on improved TF-IDF algorithm," in *Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)*, 2018, pp. 501–506, doi: 10.2991/ncce-18.2018.79.
[22] A. Sinha, M. N. B. J. Naskar, M. Pandey, and S. S. Rautaray, "Text classification using machine learning techniques: comparative analysis," in *2022 OITS International Conference on Information Technology (OCIT)*, Dec. 2022, pp. 102–107, doi: 10.1109/OCIT56763.2022.00029.
[23] H. A. Salman, A. Kalakech, and A. Steiti, "Random forest algorithm overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.
[24] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[25] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[26] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.
[27] D. Berrar, "Cross-validation," in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 542–545.
[28] G. Di Leo and F. Sardanelli, "Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach," *European Radiology Experimental*, vol. 4, no. 1, Mar. 2020, doi: 10.1186/s41747-020-0145-y.

BIOGRAPHIES OF AUTHORS

Achmad Rizki Aditama is a master of information technology student in the Faculty of Computer Science at Universitas Indonesia. He obtained their bachelor's degree in computer science from the Faculty of Computer Science at Universitas Brawijaya. In addition to being an active student, he is working professionally in software engineering at one of the leading tech companies in Indonesia. His interest in modern backend development, microservices architecture, scalable system design and emerging technologies such as machine learning. He can be contacted at email: achmad.rizki02@ui.ac.id.

Alfan Farizki Wicaksono serves as a director at Universitas Indonesia Center for Legal Informatics, where his expertise significantly contributes to the center's research and development initiatives. With a Ph.D. from the University of Melbourne, an M.Sc. from the Korea Advanced Institute of Science and Technology, and a Bachelor of Engineering (S.T.) from Institut Teknologi Bandung. His research interests lie at the intersection of information retrieval and representation learning. He can be contacted at email: alfan@cs.ui.ac.id.