

Bot Comment Detection in Indonesian Online News Using IndoBERT

M. Wanda ^{1*}, Rizal ^{2*}, Yesy Afrillia ^{3*}

* Informatics Engineering, Malikussaleh University

wanda.210170041@mhs.unimal.ac.id ¹, rizal@unimal.ac.id ², yesy.afrillia@unimal.ac.id ³

Article Info

Article history:

Received 2026-09-07

Revised 2026-09-25

Accepted 2026-10-06

Keywords:

BERT,
Bot Comment Detection,
IndoBERT,
Natural Language Processing,
Text Classification.

ABSTRACT

Bot-generated comments in online news discussions can reduce discussion quality and complicate content moderation, particularly when automated comments resemble natural human language. This study aims to evaluate IndoBERT for classifying Indonesian online-news comments into bot and human classes and to compare its performance with Random Forest as a conventional machine-learning baseline. The dataset consists of 1,000 comments collected through web scraping from six Indonesian news and discussion platforms, namely Detik.com, Kompas.com, Kompasiana.com, Kaskus.co.id, Quora.com, and CNN Indonesia. The dataset contains 500 bot comments and 500 human comments. Comments were labeled using rule-guided manual annotation based on repetitive patterns, external promotional links, textual similarity, templated structures, contextual irrelevance, and posting-frequency patterns. The data were preprocessed and divided into 70% training data and 30% evaluation data. IndoBERT was fine-tuned for binary classification using three training epochs. The evaluation produced 95.67% accuracy, 91.56% precision, 100% recall, 95.59% F1-score, and 0.9164 MCC, while Random Forest achieved an F1-score of 91.84%. The results indicate that IndoBERT achieved higher F1-score than Random Forest on the dataset and evaluation procedure used in this study. However, the findings should be interpreted within the study setting because external, temporal, cross-platform validation, repeated runs, and broader baseline comparisons were not performed. The contribution of the study lies in the empirical evaluation of IndoBERT for Indonesian online-news bot-comment classification, while the web application serves as an implementation artifact.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

A. Background, Operational Definition, and Problems

The comment column on the news portal provides a space for readers to express their opinions and interact on an issue. On the other hand, open interaction spaces can also be used to generate comments that are repetitive, promotional, irrelevant, or indicated to come from automated processes. Systematic reviews show that bot detection is evolving from a feature-based approach towards machine learning and transformers, although most research still focuses on social media [1], [2]. In this study, bot comments are operationally defined as comments that show a combination of textual characteristics and appearance patterns that indicate the possibility of automatic creation or distribution. These

characteristics include text repetition, repetitive links or promotions, templatic sentence structure, irrelevance to the context of the article, and unnatural patterns of frequency of occurrences.

Rule-based approaches and conventional machine learning can be used to detect specific patterns in text, but the contextual representation of transformer models provides an alternative to understanding the relationships between words in more depth [3]. BERT is a transformer model designed to produce bidirectional representations of language [3], while IndoBERT is a pre-trained model developed for the Indonesian context [4]. Previous research has applied BERT and its derivatives to a wide range of bot detection and text processing tasks [5],[6],[7],[8], but most bot detection research still focuses on social media. This

condition shows the need for evaluation in the context of Indonesian news commentary.

This study used comment data from Detik.com, Kompas.com, Kompasiana.com, Kaskus.co.id, Quora.com, and CNN Indonesia. The total dataset amounted to 1,000 comments consisting of 500 bot comments and 500 human comments. The focus of the study was classification based on the characteristics of the comments, not forensic identification of the account owner, as account metadata was not used as the main feature of the model.

B. Research Gap and Objectives

Previous research has shown that BERT has been used for bot detection and a variety of text classification tasks, including on Twitter data and comments [5],[6],[7],[8]. Indonesian-language research has also applied BERT or IndoBERT to comment analysis, sentiment, coarse language detection, and text classification [9],[10],[11],[12],[13],[14],[15], including the use of BERT in news classification [20]. However, research that specifically evaluated IndoBERT to distinguish between bot and human comments in the context of Indonesian-language online news is still limited. Studies using Random Forest, SVM, and XGBoost for bot detection have been reported [17], while SVM has also been used for the classification of opinions on news [18]. A comparison of IndoBERT with Random Forest on the classification of Indonesian texts has also been carried out [19]. This shows the importance of comparative evaluation between transformer models and baseline machine learning on the same dataset.

Based on these gaps, the objectives of the study are:

- 1) Measure IndoBERT's performance in classifying Indonesian news comments as bots or humans using Accuracy, Precision, Recall, F1-Score, and MCC.
- 2) Analyze the pattern of misclassification through the confusion matrix and the prediction example on the evaluation set.
- 3) Compare the performance of IndoBERT with Random Forest as a baseline on the same dataset.

The research contribution is limited to empirical evidence of IndoBERT's performance in the research dataset. This study does not claim that the model is a universal bot detector because external validation, temporal, cross-platform, repeated runs, or comparison with all relevant machine learning baselines have not been carried out.

II. METHOD

A. Data Sources

The research data was collected through web scraping of public comments on political news articles from six platforms, namely Detik.com, Kompas.com, Kompasiana.com, Kaskus.co.id, Quora.com, and CNN Indonesia. The final dataset consisted of 1,000 comments in Indonesian with two classes, namely bots and humans. The composition of the class was 500 bot comments and 500 human comments.

Data collection is carried out using a combination of Selenium WebDriver and BeautifulSoup to dynamically handle pages that contain content. The stored data includes the content of the comments as well as technical metadata that is available and required for the processing process. User identity is not used as the primary classification feature.

B. Labeling Procedure

Labeling was carried out using rule-guided manual annotation, not semi-supervised learning, because the experiment used did not carry out the process of forming pseudo-labels iteratively through the learning model. Comments are labeled as bots when they show a combination of the following indicators:

- 1) repetitive or identical textual patterns in different articles;
- 2) external links or repeated promotions that are not accompanied by relevant opinions;
- 3) textual similarity of more than 80% in repeated comments across articles;
- 4) very generic templatistic or promotional sentence structure;
- 5) irrelevance to the context of the article;
- 6) Unusual frequency patterns of occurrence based on available time information.

Comments are labeled human if they show natural language variation, relevance to the context of the article, and contextual opinions, experiences, or responses.

Labeling is performed by researchers based on these criteria. The research dataset does not document independent annotations by two or more annotators, so inter-annotator agreements such as Cohen's kappa are not counted and are not reported. These limitations are recognized as a potential source of subjectivity in the labeling process.

C. Duplication Control and Data Leakage Risk

The dataset consists of 1,000 comments from six platforms. The dataset has a class distribution of 50% bots and 50% humans. However, the recap of the number of comments per platform is not validated in the final dataset used for the experiment. Therefore, this study does not establish a distribution number per platform that cannot be verified.

Differences in platform characteristics, discussion patterns, language styles, and user composition have the potential to cause platform bias. Thus, the results of the study are not interpreted as a representation of all Indonesian news platforms. Further validation with an explicitly documented amount of data per platform is needed to measure the influence of platforms on model performance.

D. Dataset Sharing and Fine-Tuning

The data were divided sequentially into 700 samples for training and 300 samples for evaluation. The sequential division was retained as it was a procedure in the original experiment. However, the 300 evaluation samples were used for performance monitoring at each epoch. Therefore, in this revision the section is called an evaluation set, not an

independent held-out test set, and the results are not used to express universal generalizations.

The model used is IndoBERT for binary classification. Fine-tuning uses AdamW with a learning rate of 3×10^{-5} , epsilon 1×10^{-8} , weight decay 0.01, linear warm-up 10%, and three epochs. The performance evaluation was monitored after each epoch.

E. Evaluation and Baseline

Performance is measured using accuracy, precision, recall, F1-score, and Matthews correlation coefficient (MCC). MCC is used as a complement because it provides a measure of the correlation between actual and predicted labels on binary classifications [20]. Random Forest is used as a baseline to compare transformer models with conventional machine learning approaches. This study has not conducted SVM or XGBoost training so there are no experimental results for either model and no numbers claimed as additional baseline results.

The study also used a single training series with a 70:30 data split. Repeated cross-validation and multiple independent runs have not been performed. Therefore, the metric results are not interpreted as estimates of cross-run statistical variability. Follow-up evaluation with multiple runs, random seeds, or repeated cross-validation is required to test performance stability.

III. RESULTS AND DISCUSSION

A. Dataset Characteristics

The final dataset consists of 1,000 comments with a balanced class distribution, namely 500 bot comments and 500 human comments. Class balance helps reduce the dominance of one class on aggregate metrics, but it does not eliminate the possibility of bias due to platform sources, political topics, article patterns, or near-duplicates.

TABLE I.
DISTRIBUTION OF DATASET CLASSES

Classes	Quantity	Propose yourself
Bot	500	50%
Humans	500	50%

To reduce the risk of data leakage, follow-up research needs to conduct explicit checks on identical or near-duplicate comments, as well as ensure comments from the same account, article, or thread are not scattered across the train and evaluation. These checks have not been thoroughly documented in current experiments, so the results are not interpreted as evidence of generalizations across accounts, articles, threads, or time periods.

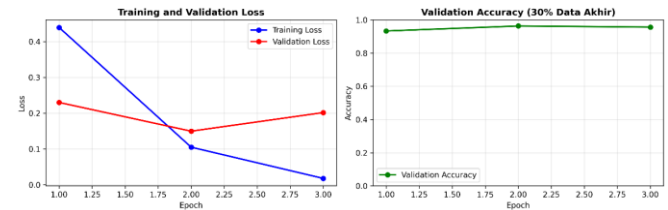
B. Training Dynamics

Table II shows a decrease in training loss from 0.4395 in the first epoch to 0.0180 in the third epoch. The evaluation loss decreased from 0.2301 in the first epoch to 0.1499 in the second epoch, then increased to 0.2018 in the third epoch. The evaluation accuracy reached 96.33% in the second

epoch and decreased to 95.67% in the third epoch. The pattern is consistent with the indication of mild overfitting in the third epoch, so the reported final results remain with the experimental configuration resulting in 95.67% in the evaluation set.

TABLE II.
RESULTS OF INDOBERT TRAINING PER EPOCH

Epoch	Training Loss	Evaluation Loss	Evaluation Accuracy
1	0,4395	0,2301	93,33%
2	0,1052	0,1499	96,33%
3	0,0180	0,2018	95,67%



Gambar 1. Kurva training loss, evaluation loss, dan evaluation accuracy.

C. IndoBERT's Performance in the Evaluation Set

In the evaluation set of 300 comments, IndoBERT produced an Accuracy of 95.67%, Precision of 91.56%, Recall of 100%, F1-Score of 95.59%, and MCC of 0.9164. This value is the result of evaluation in the experimental configuration used in this study.

Confusion matrix is used to identify true and false prediction patterns in more detail [20]. Based on the results of the reported evaluation, 141 True Positive, 146 True Negative, 13 False Positive, and 0 False Negative were obtained. A recall of 100% is a direct consequence of the absence of False Negatives in the evaluation set. With the formula $\text{Recall} = \text{TP}/(\text{TP}+\text{FN})$, $141/(141+0) = 1.00$ or 100% is obtained. Therefore, the recall value is maintained as an empirical result of the evaluation set, not as a claim that the model will result in a 100% recall on the new data.

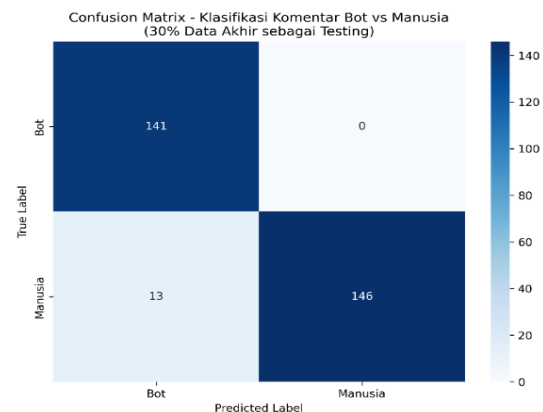


Figure 2. IndoBERT confusion matrix in 300 evaluation samples.

TABLE III.
RESULTS OF INDOBERT'S EVALUATION

Metric	Value
Accuracy	95,67%
Precision	91,56%
Recall	100%
F1-score	95,59%
MCC	0,9164

D. Error Analysis

The error analysis showed that the main error pattern on the evaluation set had to do with comments that had linguistic characteristics that could resemble bot comments or promotional patterns, but were still labeled human. These conditions showed that the model faced not only simple class differences, but also comments that had boundary characteristics between organic and automated patterns.

Examples of predictions that are inconsistent with the final confusion matrix are not used as quantitative evidence of the final result. The numerical interpretation of the study refers to the confusion matrix and the final evaluation output. Thus, the value of FN=0 is only maintained if it matches the output of the final prediction used to calculate the metric.

E. Comparison with Random Forest

Random Forest obtained an accuracy of 92%, precision of 100%, recall of 84.91%, and an F1-score of 91.84%. IndoBERT had an F1-score of 95.59%, or 3.75 percentage points higher on the same dataset. Random Forest had higher precision, but lower recall. The pattern shows that in this experiment IndoBERT provided a better precision-recall balance. This comparison should not be interpreted as evidence of universal superiority because only one baseline and one data division were used.

TABLE IV.

COMPARISON OF INDOBERT AND RANDOM FOREST EVALUATION METRICS

Models	Accuracy	Precision	Recall	F1 Score
IndoBERT	95,67%	91,56%	100%	95,59%
Random Forest	92%	100%	84,91%	91,84%

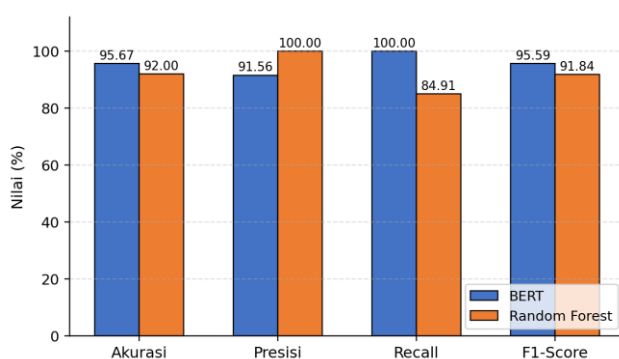


Figure 3. Comparison of evaluation metrics between IndoBERT and Random Forest.

These results show that IndoBERT assigns a higher F1-Score than Random Forest to the dataset and evaluation protocols of this study. The findings are relevant to studies that used Random Forest, SVM, and XGBoost for bot

detection [17], as well as a study comparing IndoBERT with Random Forest on the classification of Indonesian texts [19]. However, the results of this study cannot be used to conclude the superiority of IndoBERT over all machine learning algorithms because the baseline tested directly is only Random Forest.

F. Error Analysis and Overclaim Risk

The confusion matrix shows 13 false positives and 0 false negatives. This means that the aggregate errors reported are entirely from human comments that are predicted to be bots. This condition is consistent with 91.56% accuracy and 100% recall. The analysis also explains why the recall looks so high: the recall denominator has no false negatives on the evaluation set.

In the source manuscript, there is an example of a comment that was previously given a false negative status, namely a promotional comment in the form of a question that was predicted to be human. This example is mathematically contradictory to the confusion matrix TP=141 and FN=0. To prevent publication errors, the example is not used as quantitative proof and needs to be re-verified against the final prediction output before publication.

Reviewers also requested repeated testing or cross-validation. The current experiment does not provide repeated cross-validation or multiple independent runs, so this revision explicitly states these limitations and does not claim statistical stability beyond a single data sharing.

G. System Implementation as a Supporting Artifact

The IndoBERT model is then integrated into a Flask-based web application prototype to demonstrate the use of the model. The prototype provides data acquisition, preprocessing, detection, and presentation of results. Black-box testing of the application function shows the tested scenario is running as expected output. However, the web application is positioned as a supporting implementation artifact, rather than as a major scientific contribution. The main contribution of the research remains on the operational definition of labels, classification experiments, evaluations, and model comparisons.

H. Ethics and Reproducibility of Data Collection

The data used comes from publicly accessible comments. Collection is done only on data that is necessary for the research; private data or access that requires authentication is not used as a research feature. Account identity is not used as a model feature. Because platform policies are subject to change, scraping procedures need to be carried out with due regard to each platform's access conditions and technical limitations. For reproducibility, advanced research should store collection scripts, retrieval dates, source URLs, article hash/IDs, and deduplication logs without exposing unnecessary personal data.

I. Research Contributions and Limitations

The main contribution of the research lies in the empirical evaluation of IndoBERT for the classification of bot and

human comments on online news in Indonesian. The web application built serves as an implementation artifact that allows the model to be used in acquisition and detection flows, but is not the main scientific contribution of the research.

Several limitations need to be noted. First, the dataset only consists of 1,000 comments and focuses on the context of political news. Second, the distribution of the number of comments per platform has not been quantitatively documented in the final dataset so that the influence of platform bias has not been statistically tested. Third, labeling is carried out manually with rule guidelines without measuring inter-annotator agreement. Fourth, leakage checks do not include comprehensive grouping of accounts, articles, threads, and near-duplicates. Fifth, the evaluation uses a 70:30 data division and a series of trainings so that repeated cross-validation or multiple independent runs are not yet available. Sixth, the experimental baseline only uses Random Forest. Due to these limitations, the results of Accuracy 95.67%, Precision 91.56%, Recall 100%, F1-Score 95.59%, and MCC 0.9164 should be understood as performance in the evaluation set of this study. These results are not enough to state universal generalizations on other platforms or time periods.

IV. CONCLUSION

This study evaluated IndoBERT for the classification of bot and human comments on 1,000 online news comments in Indonesian. In the evaluation set, IndoBERT obtained an Accuracy of 95.67%, Precision of 91.56%, Recall of 100%, F1-Score of 95.59%, and MCC of 0.9164. Random Forest obtained an F1-Score of 91.84%, so that IndoBERT gave a higher F1-Score of 3.75 percentage points on the same evaluation dataset and protocol.

The confusion matrix shows 141 True Positives, 146 True Negatives, 13 False Positives, and 0 False Negatives in the reported evaluation results. The Recall value is 100% because no False Negatives were found in the evaluation set. The findings show high performance on the research dataset, but are not intended as universal generalization evidence.

Thus, the contribution of the research is to provide empirical evidence of the use of IndoBERT for the classification of bot comments on online news in Indonesian as well as comparison with Random Forest. Subsequent developments need to expand the dataset, document the distribution per platform, strengthen annotation procedures, control data leakage more strictly, add baselines, and perform repeated runs and external validation to test the stability and generalization of the model.

BIBLIOGRAPHY

- [1] K. Hayawi, S. Saha, M. M. Masud, S. S. Mathew, and M. Kaosar, "Social media bot detection with deep learning methods: a systematic review," *Neural Comput. Appl.*, vol. 35, no. 12, pp. 8903–8918, 2023, doi: 10.1007/s00521-023-08352-z.
- [10] A. M. Priyatno, M. M. Muttaqi, F. Syuhada, and A. Z. Arifin, "Detection of twitter spammer bots based on time interval entropy and global vectors for word representations tweet's hashtag," *Regist. J. Ilm. Teknol. Last. Inf.*, vol. 5, no. 1, pp. 37–46, 2019, doi: 10.26594/register.v5i1.1382.
- [11] M. Firdaus and P. Nur Miftahur Rizki, "BIJAKAWEB: A Web-Based Platform for Hate Speech Detection in Indonesian News Comments," *J. Technology. Inf. and Computing Science.*, vol. 11, no. 4, pp. 939–948, 2024, doi: 10.25126/jtiik.1148719.
- [12] R. Fatmasari, R. K. Septiani, T. H. Pinem, D. Fabiyanto, and W. Gata, "Implementation of the BERT Algorithm on Comments of Academic and Non-Academic Services of Open Universities on Social Media," *Science, Apps. Computing and Technology. Inf.*, vol. 5, no. 2, p. 96, 2024, doi: 10.30872/jsakti.v5i2.13915.
- [2] Z. Ellaky, F. Benabbou, and S. Ouahabi, "Systematic Literature Review of Social Media Bots Detection Systems," vol. 35, 2023, doi: 10.1016/j.jksuci.2023.04.004.
- [3] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NaacI-Hlt 2019*, no. Mlm, pp. 4171–4186, 2018, [Online]. Available: <https://aclanthology.org/N19-1423.pdf>
- [4] F. Koto and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," pp. 757–770, 2020.
- [5] D. Duki, D. Ke, and D. Stipi, "Are You Human ? Detecting Bots on Twitter Using BERT," 2020, doi: 10.1109/DSAA49011.2020.00089.
- [6] M. Heidari and J. H. J. Jones, "Using BERT to Extract Topic-Independent Sentiment Features for Social Media Bot Detection," 2020.
- [7] Q. Guo, H. Xie, Y. Li, W. Ma, and C. Zhang, "Social bots detection via fusing bert and graph convolutional networks," *Symmetry (Basel)*, vol. 14, no. 1, pp. 1–14, 2022, doi: 10.3390/sym14010030.
- [8] S. Kumar, S. Garg, Y. Vats, and A. S. Parihar, "Content based bot detection using bot language model and BERT embeddings," in *Proceedings of ICCSP*, 2021, pp. 285–289. doi: 10.1109/ICCSP52374.2021.9465506.
- [9] A. Sallah et al., "Fine-Tuned Understanding: Enhancing Social Bot Detection with Transformer-Based Classification," *IEEE Access*, vol. 12, no. August, pp. 118250–118269, 2024, doi: 10.1109/ACCESS.2024.3440657.
- [13] S. Imron, E. I. Setiawan, and J. Santoso, "Detection of E-Commerce Review Aspects Using IndoBERT Embedding and CNN," *J. Intell. Syst. Comput.*, vol. 5, no. 1, pp. 10–16, 2023, doi: 10.52985/insyst.v5i1.267.
- [14] F. A. Pratama and A. Romadhony, "Identification of toxic comments with BERT," *e-Proceeding Eng.*, vol. 7, no. 2, pp. 7941–7949, 2020, doi: 10.34818/eoe.v7i2.13073.
- [15] A. D. Rendragraha, M. A. Bijaksana, and A. Romadhony, "Transformers method approach for the detection of foul language in Indonesian online news commentary," *e-Proceeding Eng.*, vol. 8, no. 2, pp. 3385–3395, 2021.
- [16] A. Fardhina et al., "A Natural Language Processing (NLP) Algorithm-Based Hoax Detection System Using Bert," vol. 4, no. 1, pp. 450–461, 2025, doi: 10.70247/jumistik.v4i1.156.
- [17] M. Averina, W. B. T. Handaya, and A. J. Santoso, "Prediction of bot accounts on social media X using Random Forest, XGBoost, and SVM," *J. Inform. Atma Jogja*, vol. 6, no. 2, pp. 57–68, 2025, doi: 10.24002/jiaj.v6i2.12948.
- [18] R. Tjut Adek, M. Fikry, and U. Khalil, "News Opinion Classification Application With Support Vector Machine Algorithm Using Framework Codeigniter," *J. Informatics Telecommun. Eng.*, vol. 5, no. 1, pp. 160–166, 2021, doi: 10.31289/jite.v5i1.5189.
- [19] S. Rohman and I. W. P. Agung, "Comparison of IndoBERT, SVM, and Random Forest Algorithms in Sentiment Analysis of the Free Nutritious Meal Program," *JATI*, vol. 9, no. 6, pp. 9464–9471, 2025.
- [20] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, vol. 17, no. 1, pp. 168–192, 2021, doi: 10.1016/j.aci.2018.08.003.