



Indonesian Chatbot Message Classification Using TF-IDF and Multinomial Naive Bayes

Farizi Sharfan^{1*}

¹Department of Mathematics, Universitas Sumatera Utara, Medan, Indonesia
farizisharfan@students.usu.ac.id

ABSTRACT

Customer-service chatbots receive messages with different communicative purposes, including abusive content that may require further review. This study examines the classification of Indonesian chatbot messages using term frequency–inverse document frequency (TF-IDF) and multinomial naive Bayes. The reported preprocessing comprises case folding, text normalization, removal of punctuation and nonalphabetic characters, and stopword removal. Messages were assigned to four categories: harassment, neutral, racist, and violence. An 80:20 stratified training–test split was reported. The available confusion matrix contains 2,141 test messages, of which 1,674 were correctly classified, giving an accuracy of 78.19%. Harassment had the highest precision (0.893), while violence had the highest recall (0.844). Neutral messages had the lowest F1-score (0.738), and 98 neutral messages were classified as violence. These results establish the performance of the reported classifier on the available test set, but they do not establish customer satisfaction, because the labels describe message content rather than satisfaction ratings. Incomplete documentation of corpus construction, annotation, model settings, and timing procedures limits reproducibility and deployment claims. The model therefore provides a preliminary baseline for content categorization, with further validation needed before its outputs inform customer-service decisions.

Keywords: Indonesian text; TF-IDF; multinomial naive Bayes; abusive language; chatbots

Received: 2 January 2026

Revised: 4 February 2026

Published: 30 March 2026

INTRODUCTION

Chatbots provide an interface through which customers seek information, request assistance, and express concerns. Their usefulness depends on the quality of the interaction as well as the information they provide. Studies of chatbot experiences and continued use examine pragmatic assistance, interaction quality, and user satisfaction as related but distinct aspects of service delivery (Ashfaq et al., 2020; Følstad & Brandtzaeg, 2020; Hsu & Lin, 2023). Classifying messages can support the organization of these interactions, although classification performance alone does not demonstrate an improvement in service quality.

Research on chatbot communication has examined links between communication style and customer satisfaction (Xu et al., 2022), business- and customer-oriented activities (Akdemir & Bulut, 2024), and service characteristics in online shopping (Kappi & Marlina, 2023). Such work motivates attention to customer experiences, but a message classifier requires a clearly defined target. A satisfaction measure ordinarily needs an explicit operational definition, such as a user rating or a validated annotation scheme. Content categories cannot automatically serve as substitutes for that measure.

The distinction is particularly relevant when labels refer to abusive language. Offensive language and hate speech are overlapping but nonidentical concepts, and their definitions affect both annotation and model evaluation (Davidson et al., 2017; Waseem et al., 2017). Indonesian-language research likewise distinguishes abusive-language detection from broader hate-speech classification and recognizes the importance of dataset-specific labeling decisions (Ibrohim & Budi, 2018; Ibrohim & Budi, 2019). A message labeled neutral is not necessarily evidence of a satisfied customer, while a message containing violent or racist language does not by itself identify the cause or degree of dissatisfaction.

This study evaluates a TF-IDF and multinomial naive Bayes pipeline for four-category classification of Indonesian chatbot messages. The objective is to describe its held-out classification performance and identify the main error patterns. The analysis is confined to harassment, neutral, racist, and violence labels. Customer satisfaction and service improvement remain possible subjects for subsequent studies requiring independent outcome measures.

METHODS

The study uses Indonesian chatbot conversation messages as individual classification units. Each message receives one of four labels: harassment, neutral, racist, or violence. The available study record does not specify the complete corpus size, collection period, platform, sampling procedure, or annotation protocol. Accordingly, these categories are treated as the dataset labels rather than as independently validated measures of sentiment or customer satisfaction. The test-set counts are available and can be checked against the reported confusion matrix.

The reported preprocessing includes case folding, normalization, removal of punctuation and nonalphabetic characters, and stopword removal using Indonesian and English stopword lists. These steps standardize the text before feature extraction. Their precise implementation, including the normalization dictionary and stopword-list versions, is not documented. Removing punctuation or other tokens may also discard information useful for interpreting abuse, so the benefits of each step would require a separate preprocessing comparison.

TF-IDF represents messages through weighted terms, assigning importance according to their occurrence within a message and their prevalence across the corpus. Term weighting is an established approach in text retrieval (Salton & Buckley, 1988). In this study, the resulting features are supplied to a multinomial naive Bayes classifier, which combines class priors with term-based evidence under a conditional-independence assumption. Naive Bayes variants have also been examined for Indonesian sentiment datasets (Umar & M. Adnan Nur, 2022); those findings provide methodological context rather than a directly comparable benchmark for the present labels.

An 80:20 stratified training–test split is reported. Stratification preserves class proportions approximately; it does not make the classes equally frequent or eliminate sampling bias. The available confusion matrix identifies 2,141 evaluation messages. Training-set counts, the random seed, vocabulary settings, smoothing parameter, and software versions are not specified. The record also does not establish whether feature fitting was restricted to the training partition or whether messages from the same conversation or user were separated across partitions. These omissions prevent a complete assessment of leakage and reproducibility.

Evaluation uses accuracy and class-specific precision, recall, and F1-score. Accuracy is the number of correct predictions divided by the total number of test messages. For each class, precision is true positives divided by all predictions assigned to that class, recall is true positives divided by all actual instances of that class, and F1 is the harmonic mean of precision and recall. Reporting several measures is necessary because they emphasize different aspects of classification performance (Sokolova & Lapalme, 2009). The values below are recalculated from the supplied confusion matrix; no new model training or independent replication is reported.

Reproducible interpretation also requires documentation of data provenance, language characteristics, annotation decisions, and intended uses. Data statements and dataset documentation frameworks provide guidance for recording these details (Bender & Friedman, 2018; Gebru et al., 2021). The present evaluation does not establish the collection permissions, de-identification procedure, or suitability of the messages for public release. These matters require clarification before data sharing or operational use.

RESULT AND DISCUSSION

The evaluation set contains 497 harassment messages, 546 neutral messages, 501 racist messages, and 597 violence messages. Figure 1 reproduces the original distribution graphic. Its bar values match the test-set row totals in Table 1, so they are interpreted as the held-out distribution rather than the distribution of the complete corpus. Violence is the largest category, followed by neutral, racist, and harassment. The largest category accounts for 27.88% of the test messages.

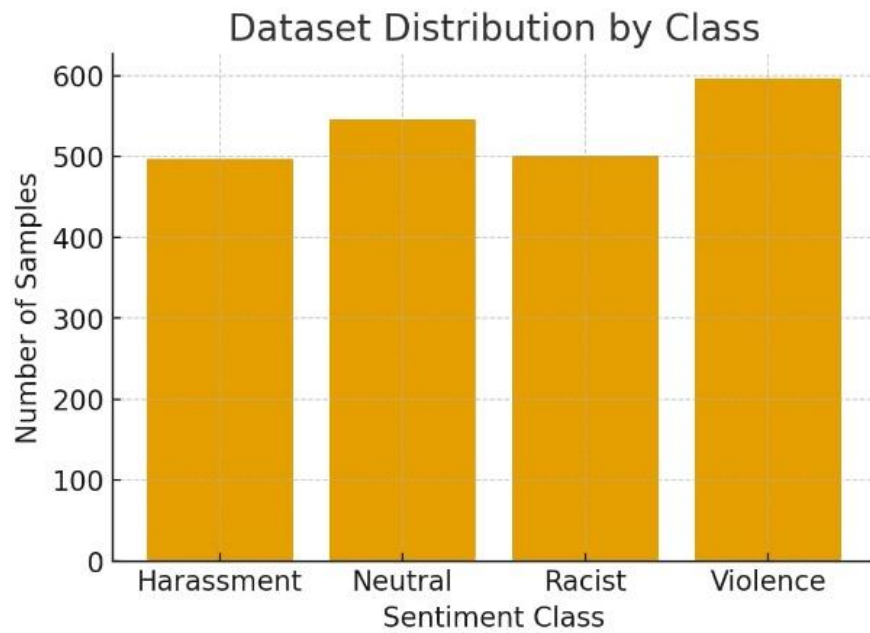


Figure 1. Class counts in the held-out evaluation set.

Note. The original graphic uses “sentiment” terminology; the labels represent content categories. Counts match the test set.

Table 1 presents the confusion matrix with actual classes in rows and predicted classes in columns. The diagonal contains 1,674 correct predictions out of 2,141 messages. The resulting accuracy is 78.19%, leaving 467 classification errors. This aggregate result describes the available evaluation partition; it does not quantify uncertainty across alternative splits or performance on messages collected in a different service setting.

Table 1. Confusion matrix for the held-out evaluation set.

Actual / predicted	Harassment	Neutral	Racist	Violence	Total
Harassment	385	25	9	78	497
Neutral	21	399	28	98	546
Racist	6	62	386	47	501
Violence	19	50	24	504	597
Total	431	536	447	727	2141

Note. Rows denote actual labels; columns denote predicted labels. Diagonal total = 1,674.

Table 2 reports the class-specific measures calculated from Table 1. Harassment has the highest precision (0.893): 385 of the 431 messages predicted as harassment carry that label in the test set. Violence has the highest recall (0.844), with 504 of 597 instances correctly identified. However, violence also has the lowest precision (0.693), because 223 messages from other classes are assigned to it. The classifier therefore detects many violence-labeled messages while generating substantial cross-class errors.

Table 2. Classification measures recalculated from the confusion matrix.

Class	Precision	Recall	F1-score	Support
Harassment	0.893	0.775	0.830	497
Neutral	0.744	0.731	0.738	546
Racist	0.864	0.770	0.814	501
Violence	0.693	0.844	0.761	597

Note. Metrics are rounded to three decimal places. Overall accuracy = 78.19% (1,674/2,141).

Figure 2 retains the original comparison of precision, recall, and F1-score, which displays values rounded to two decimal places. The three-decimal values in Table 2 provide a more precise account of the same confusion matrix. Harassment has the highest F1-score (0.830), followed by racist (0.814), violence (0.761), and neutral (0.738). Consequently, the statement that neutral messages have the lowest performance is accurate for recall and F1, but not for precision.

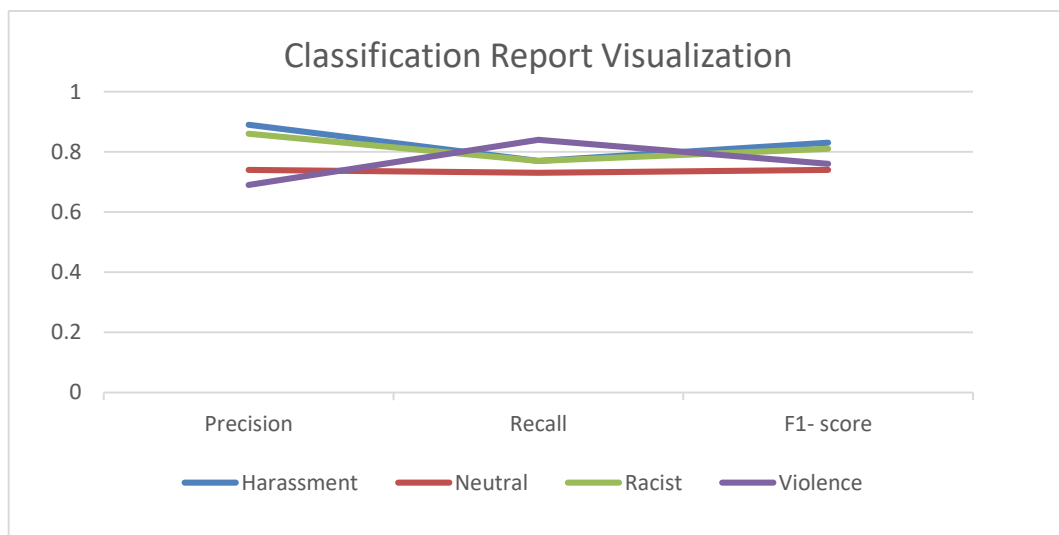


Figure 2. Original class-specific precision, recall, and F1-score.

Note. The original chart rounds metrics to two decimals; see Table 2 for recalculated values.

The most frequent individual off-diagonal entry is the assignment of 98 neutral messages to violence. This represents 17.95% of all neutral messages and 66.67% of the 147 errors within that class. In addition, 78 harassment messages and 47 racist messages are predicted as violence. Overall, the model predicts violence 727 times, compared with 597 actual violence labels. These counts suggest that the violence decision boundary deserves closer examination, but the matrix alone cannot identify the linguistic causes of these errors.

Possible explanations include overlapping vocabulary, inconsistent label definitions, or information lost when messages are treated independently of their conversations. These remain hypotheses because misclassified examples and annotation records are unavailable. Work on abusive-language datasets shows why category definitions and annotation procedures matter (Founta et al., 2018). Analysis of racial bias in other hate-speech datasets further motivates auditing subgroup effects, although those findings cannot be transferred directly to this Indonesian corpus (Davidson et al., 2019). No subgroup performance or fairness claim is supported by the current results.

The findings should be interpreted as content-classification results rather than evidence of customer satisfaction. The study includes neither satisfaction ratings nor a validated mapping between its four labels and satisfaction levels. It therefore cannot determine whether dissatisfied customers are recognized more accurately than satisfied customers, or whether using the classifier improves service outcomes. Linking content predictions to such outcomes would require a separate study with an explicit satisfaction measure and a suitable evaluation design.

The reported pipeline offers a baseline for further comparison. Indonesian language-understanding benchmarks and social-media language models provide relevant examples of alternative evaluation resources and representations (Koto et al., 2021; Willie et al., 2020). They do not establish that a more complex model would outperform this classifier on the present data. A meaningful comparison would use the same documented corpus, preprocessing conditions, and evaluation partition, and would report both predictive performance and computational requirements.

No reproducible timing protocol accompanies the original claim of prediction in less than 10 ms. Hardware, batch size, preprocessing overhead, repeated measurements, and latency variability are not documented. The available evidence therefore does not establish a latency threshold or production readiness. Before operational use, evaluation should include end-to-end timing and review of the consequences of false positive and false negative predictions. Model reporting should also state intended uses, limitations, and evaluation conditions, consistent with model-card guidance (Mitchell et al., 2019).

The principal limitations are incomplete corpus and annotation documentation, a single reported split, unknown conversation-level overlap, and the absence of external validation. The observed accuracy is useful as a description of this test set, but its generalizability remains unresolved. Future work should first make the existing experiment reproducible, examine representative errors, and evaluate additional data before drawing conclusions about customer-service benefits.

CONCLUSION

TF-IDF and multinomial naive Bayes classified 1,674 of 2,141 Indonesian chatbot messages correctly, corresponding to 78.19% accuracy. Performance varied across the four content categories: harassment had the highest precision and F1-score, violence had the highest recall, and neutral had the lowest recall and F1-score. Frequent assignments of other categories to violence indicate a specific area for further error analysis.

These results support the use of the reported experiment as a preliminary message-classification baseline. They do not validate customer-satisfaction measurement, service improvement, or a sub-10-ms prediction claim. Complete documentation of the corpus, annotation process, model settings, and evaluation procedure, followed by independent validation, is needed before broader conclusions or operational deployment.

BIBLIOGRAPHY

Akdemir, D. M., & Bulut, Z. A. (2024). Business and customer-based chatbot activities: The role of customer satisfaction in online purchase intention and intention to reuse chatbots. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(4), 2961–2979. <https://doi.org/10.3390/jtaer19040142>

- Ashfaq, M., Yun, J., Yu, S., & Loureiro, S. M. C. (2020). I, chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics*, 54, Article 101473. <https://doi.org/10.1016/j.tele.2020.101473>
- Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6, 587–604. https://doi.org/10.1162/tacl_a_00041
- Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 25–35). <https://doi.org/10.18653/v1/w19-3504>
- Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 512–515. <https://doi.org/10.1609/icwsm.v11i1.14955>
- Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1). <https://doi.org/10.1609/icwsm.v12i1.14991>
- Følstad, A., & Brandtzaeg, P. B. (2020). Users' experiences with chatbots: Findings from a questionnaire study. *Quality and User Experience*, 5(1), Article 3. <https://doi.org/10.1007/s41233-020-00033-2>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
- Hsu, C.-L., & Lin, J. C.-C. (2023). Understanding the user satisfaction and loyalty of customer service chatbots. *Journal of Retailing and Consumer Services*, 71, Article 103211. <https://doi.org/10.1016/j.jretconser.2022.103211>
- Ibrohim, M. O., & Budi, I. (2018). A dataset and preliminaries study for abusive language detection in Indonesian social media. *Procedia Computer Science*, 135, 222–229. <https://doi.org/10.1016/j.procs.2018.08.169>
- Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 46–57). <https://doi.org/10.18653/v1/w19-3506>
- Kappi, C. M. K., & Marlina, L. (2023). The effect of chatbot services on online shop customer satisfaction. *Brilliance: Research of Artificial Intelligence*, 3(2), 252–261. <https://doi.org/10.47709/brilliance.v3i2.3133>
- Koto, F., Lau, J. H., & Baldwin, T. (2021). Indobertweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10660–10668). <https://doi.org/10.18653/v1/2021.emnlp-main.833>

- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). <https://doi.org/10.1145/3287560.3287596>
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Umar, N., & M. Adnan Nur (2022). Application of Naïve Bayes algorithm variations on Indonesian general analysis dataset for sentiment analysis. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 6(4), 585–590. <https://doi.org/10.29207/resti.v6i4.4179>
- Waseem, Z., Davidson, T., Warmusley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 78–84). <https://doi.org/10.18653/v1/w17-3012>
- Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). Indonlu: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 843–857). <https://doi.org/10.18653/v1/2020.aacl-main.85>
- Xu, Y., Zhang, J., & Deng, G. (2022). Enhancing customer satisfaction with chatbots: The influence of communication styles and consumer attachment anxiety. *Frontiers in Psychology*, 13, Article 902782. <https://doi.org/10.3389/fpsyg.2022.902782>