

Public Sentiment Analysis Toward Indonesian House of Representatives Member Allowances using Naïve Bayes

Khairul Siregar^{1*}, Relita Buaton², Muammar Khadapi³

^{1,2,3}*Sekolah Tinggi Manajemen Informatika dan Komputer Kaputama, Binjai, Indonesia*
khairulsiregar169@gmail.com^{1*}, bbcbuaton@gmail.com², khdafi5@gmail.com³

Abstract

This study analyzes public sentiment toward allowances for members of the Indonesian House of Representatives using a Multinomial Naïve Bayes classifier. The initial dataset contained 14,000 public posts and comments collected from X and YouTube between July 2022 and February 2026. After validation and preprocessing, 12,869 records were retained. Sentiment labels were generated through an Indonesian lexicon-based procedure with two-token negation handling, while textual features were represented using augmented TF-IDF. The dataset was divided into 10,295 training records and 2,574 testing records. The resulting labels comprised 10,675 negative, 1,912 positive, and 282 neutral records. Testing produced an accuracy of 84.50%, macro precision of 0.5410, macro recall of 0.3767, and macro F1-score of 0.3819. The negative class achieved the strongest performance, whereas the neutral class was not successfully recognized because of severe class imbalance. The classification results were deployed through a Streamlit application that presents sentiment distribution, class-specific word clouds, temporal and source distributions, evaluation metrics, and new-text prediction. The findings show that Naïve Bayes can provide a useful overall classification result, but balanced class representation remains essential for reliable performance across all sentiment categories.

Keywords: *Augmented TF-IDF; Naïve Bayes; Public policy; Sentiment analysis; Streamlit*

1. Introduction

Social media has become an important digital public sphere in which citizens express support, criticism, and rejection of public policies. These expressions can shape collective perception and provide an empirical indication of how institutions are viewed by society. Public sentiment toward government policy has therefore become a growing research topic, particularly when public discussion is distributed across large volumes of unstructured text [1]. The social media presence of the Indonesian House of Representatives also creates a direct channel through which citizens respond to legislative activities and institutional decisions [13].

One issue that repeatedly attracts public attention is the allowance received by members of the Indonesian House of Representatives. The issue is associated with state expenditure, social fairness, public trust, and perceptions of legislative performance. Although X and YouTube contain a large number of relevant comments, the data cannot be interpreted objectively without a systematic text-analysis procedure. Sentiment analysis addresses this limitation by classifying opinions into positive, negative, and neutral categories.

Previous studies have demonstrated the usefulness of Naïve Bayes for social-media sentiment classification. Putri et al. obtained 80% accuracy when analyzing 1,546 tweets concerning the performance of the House of Representatives [2]. Aprilia and Isnain applied Naïve Bayes to an anti-corruption campaign on Twitter [3], while Fadli et al. reported 65.60% accuracy for public opinion concerning fuel-price increases [4]. Dani compared Naïve Bayes, support vector machines, and random forests on 78,866 YouTube comments about legislative allowances; Naïve Bayes achieved 75%, whereas the best support vector machine configuration reached 93% [5]. These studies confirm that the algorithm is practical for text classification, but performance depends strongly on feature representation and class distribution.

Existing studies generally emphasize model comparison and accuracy, while fewer studies combine multiple social-media sources with an interactive analytical interface. This study addresses that gap by combining public text from X and YouTube, applying lexicon-based labeling with negation handling, representing text through augmented TF-IDF, and training a Multinomial Naïve Bayes model. The results are deployed through Streamlit so that sentiment distribution, evaluation results, source and time patterns, and predictions for new text can be examined within one application.

The study has three objectives: to implement Naïve Bayes for classifying public sentiment toward legislative allowances, to evaluate the model using accuracy, precision, recall, F1-score, and a confusion matrix, and to present the classification results through an understandable web-based visualization. The article is organized into related work, research method, results and discussion, and conclusion.

2. Related work

2.1. Sentiment analysis and Naïve Bayes

Sentiment analysis is a text-mining task that identifies the polarity expressed in textual data. Social-media text commonly contains informal vocabulary, abbreviations, inconsistent spelling, symbols, and repeated expressions; consequently, preprocessing is required before classification. Typical operations include case folding, cleaning, tokenization, normalization, stopwords removal, and stemming [8]. After preprocessing, the text must be transformed into numerical features that can be processed by a machine-learning algorithm.

Naïve Bayes is a probabilistic classifier based on Bayes' theorem and a conditional-independence assumption among features. Despite the simplicity of this assumption, the method remains effective for high-dimensional sparse text data and has a low computational cost [12]. TF-IDF weighting is commonly used to represent the relative importance of terms before classification [9]. In this study, the standard representation is extended through augmented term frequency and L2 normalization, and the resulting non-negative feature matrix is used to train Multinomial Naïve Bayes.

2.2. Previous studies

Table 1 summarizes the most relevant studies and clarifies the distinction between the present work and earlier research.

Table 1: Summary of related sentiment-analysis studies

Study	Data and method	Main result
Putri et al. [2]	1,546 Twitter posts; Naïve Bayes	Accuracy of 80%; negative sentiment was dominant.
Fadli et al. [4]	Twitter data on fuel prices; Naïve Bayes and decision tree	Naïve Bayes accuracy of 65.60%.
Dani [5]	78,866 YouTube comments; Naïve Bayes, SVM, and random forest	Naïve Bayes achieved 75%; SVM reached 93% after SMOTE.
Ismardani and Fatah [6]	YouTube comments on government policy; Naïve Bayes and TF-IDF	The model remained useful despite imbalanced data.
Pratama et al. [7]	Brand24 social-media monitoring of legislative allowances	Public discussion was dominated by negative sentiment.

The present study differs by combining X and YouTube data, using augmented TF-IDF, reporting both aggregate and per-class evaluation, and integrating the model into a Streamlit application.

3. Research method

3.1. Research framework and data

The research followed the six CRISP-DM stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The stages were implemented as an integrated pipeline, as illustrated in Fig. 1, so that the preparation, modeling, evaluation, and deployment procedures remained consistent throughout the study.

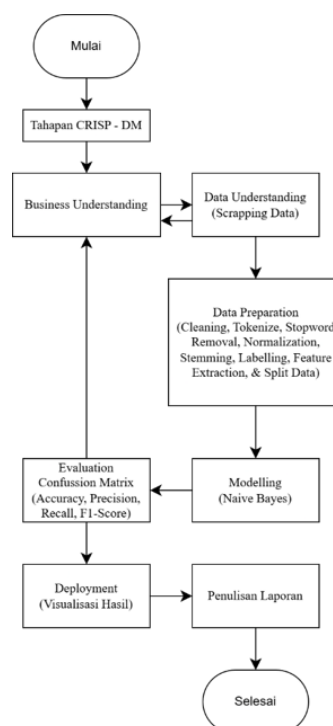


Fig. 1: Research workflow based on CRISP-DM

The initial data consisted of 14,000 Indonesian-language public posts and comments related to legislative allowances, salaries, and performance. The records were collected from X and YouTube within the period from 28 July 2022 to 22 February 2026. Validation removed unusable, empty, duplicate, or structurally invalid records, leaving 12,869 valid observations for analysis.

Table 2: Dataset and model configuration

Component	Configuration
Initial records	14,000 public posts and comments
Valid records	12,869 records
Sources	X and YouTube
Observation period	28 July 2022–22 February 2026
Sentiment classes	Negative, neutral, and positive
Training–testing split	10,295 records (80%) and 2,574 records (20%)
Feature representation	Augmented TF-IDF with L2 normalization
Vocabulary size	16,859 features
Classifier	Multinomial Naïve Bayes
Deployment	Streamlit web application

3.2. Text preparation and sentiment labeling

Text preparation was applied consistently to the full dataset. The process included case folding, removal of URLs and unnecessary symbols, tokenization, normalization of informal Indonesian words, stopword removal, and stemming. Source names and date values were also standardized to support source-based and temporal visualization.

Sentiment labels were generated using an Indonesian sentiment lexicon. Each recognized token contributed its lexicon weight to the document score. A two-token context window was used to detect the negation terms “tidak,” “bukan,” “jangan,” “belum,” and “tanpa”; when a negation occurred before a sentiment-bearing word, the word weight was reversed. The implementation used the same decision rule as the deployed system: scores greater than 10 were labeled positive, scores lower than 10 were labeled negative, and a score equal to 10 was labeled neutral. This automatic labeling stage produced the target classes used for supervised training [10].

3.3. Feature representation and model training

The labeled dataset was divided into training and testing subsets using an 80:20 stratified split. The training subset contained 10,295 records and the testing subset contained 2,574 records. Stratification preserved the class proportions in both subsets and ensured that the reported evaluation used only records that were not used for model fitting.

A count-based vocabulary was fitted on the training data. Term frequency was transformed through augmented scaling, which reduces the dominance of highly repeated words within a single document. The augmented values were multiplied by inverse document frequency derived from the training subset and then normalized using the L2 norm. The same fitted vocabulary and IDF values were applied to the training set, testing set, and new text entered through the prediction page. The final feature space contained 16,859 features.

The non-negative augmented TF-IDF matrix was used to train a Multinomial Naïve Bayes classifier. Model fitting, prediction, evaluation, and artifact storage were executed within the application pipeline. No test records were used to fit the vocabulary, calculate IDF, or estimate model parameters.

3.4. Evaluation and deployment

Evaluation used accuracy, macro precision, macro recall, macro F1-score, a three-class confusion matrix, and precision, recall, F1-score, and support for each class. Macro averages were emphasized because they assign equal importance to the three sentiment classes and therefore reveal performance limitations that can be hidden by overall accuracy in an imbalanced dataset. The trained model and analytical outputs were then deployed through Streamlit with four menus: Home, Data Process, Dashboard, and Text Prediction.

4. Results and discussion

4.1. Sentiment distribution

The automatic labeling process classified 10,675 records as negative, 1,912 as positive, and 282 as neutral. These values correspond to 82.95%, 14.86%, and 2.19% of the valid dataset, respectively. Fig. 2 shows that the dataset was severely imbalanced, with the negative class accounting for more than four fifths of all records.

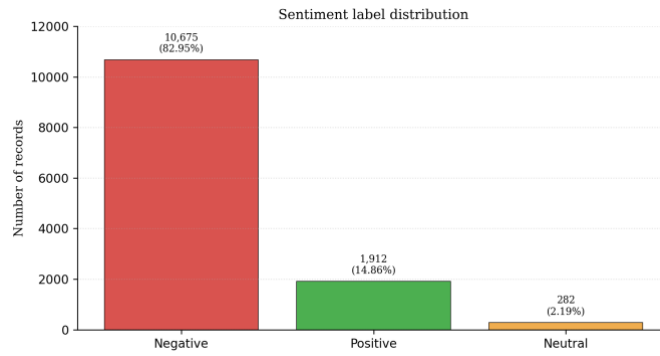


Fig. 2: Distribution of sentiment labels in 12,869 valid records

The dominance of negative sentiment is consistent with earlier studies concerning legislative performance and allowance-related public discussion [2][7]. Nevertheless, the distribution must be interpreted as the output of the configured lexicon-labeling procedure rather than as a population estimate of all Indonesian citizens. The limited neutral class also creates a difficult learning condition because the classifier receives comparatively few neutral examples.

4.2. Overall model performance

The overall evaluation results are presented in Table 3. The model achieved 84.50% accuracy on 2,574 test records. However, macro precision, recall, and F1-score were substantially lower because the macro calculation gives equal weight to the minority classes.

Table 3: Overall Multinomial Naïve Bayes evaluation

Metric	Value
Accuracy	0.8450 (84.50%)
Macro precision	0.5410
Macro recall	0.3767
Macro F1-score	0.3819
Training records	10,295 (80%)
Testing records	2,574 (20%)
Vocabulary	16,859 features

The difference between accuracy and macro F1-score indicates that the high overall correctness was strongly influenced by the majority negative class. Accuracy alone would therefore provide an incomplete assessment of the classifier. This finding supports the use of class-level measures when sentiment labels are imbalanced [6][11].

4.3. Confusion matrix and class-level performance

The confusion matrix in Fig. 3 and Table 4 shows how the 2,574 testing records were classified. The model correctly identified 2,123 negative and 52 positive records, but it did not predict any record as neutral.

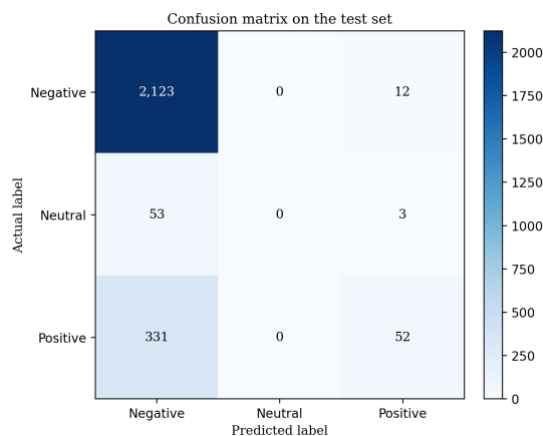


Fig. 3: Confusion matrix for the 2,574 testing records

Table 4: Confusion matrix values

Actual / Predicted	Negative	Neutral	Positive	Total
Negative	2,123	0	12	2,135
Neutral	53	0	3	56
Positive	331	0	52	383
Total	2,507	0	67	2,574

Table 5: Precision, recall, and F1-score by sentiment class

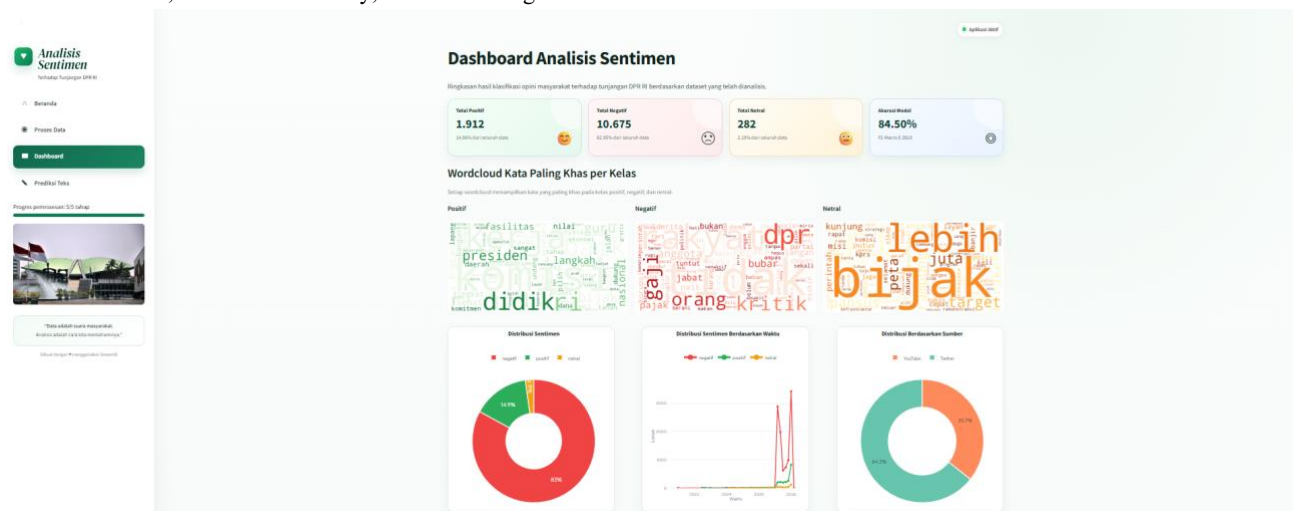
Class	Precision	Recall	F1-score	Support
Negative	0.8468	0.9944	0.9147	2,135
Neutral	0.0000	0.0000	0.0000	56
Positive	0.7761	0.1358	0.2311	383

The negative class achieved precision of 0.8468, recall of 0.9944, and F1-score of 0.9147. Its very high recall means that 2,123 of 2,135 negative test records were recognized. The positive class obtained relatively strong precision of 0.7761, but recall was only 0.1358 because just 52 of 383 positive records were identified. The neutral class had zero precision, recall, and F1-score because none of the 56 neutral records was predicted as neutral.

These results demonstrate that the classifier learned the dominant negative pattern effectively but did not learn balanced boundaries for the two minority classes. The large number of positive records predicted as negative and the absence of neutral predictions explain the low macro recall and macro F1-score. Future improvements should therefore prioritize a more balanced dataset, broader lexicon coverage, review of automatically generated labels, and explicit handling of informal language, sarcasm, and more varied negation patterns.

4.4. Streamlit implementation

The classification pipeline was integrated into a Streamlit application. The Home page introduces the study, the Data Process page executes labeling, feature construction, training, and evaluation, the Dashboard presents the analytical results, and the Text Prediction page classifies new opinions. The dashboard displays the number and percentage of each sentiment class, class-specific word clouds, temporal distribution, source distribution, and model accuracy, as shown in Fig. 4.

**Fig. 4:** Streamlit dashboard for sentiment-analysis results

The prediction page applies exactly the same preprocessing, vocabulary, IDF values, and trained model used during evaluation. It presents the predicted class, confidence level, probability for each class, processed tokens, and detected features. This consistency prevents a mismatch between model training and operational prediction and makes the classification process easier to interpret.

5. Conclusion

This study successfully implemented Multinomial Naïve Bayes to classify public sentiment toward allowances for members of the Indonesian House of Representatives. From 14,000 initial records, 12,869 valid X and YouTube records were processed through text normalization, lexicon-based labeling with negation handling, augmented TF-IDF representation, an 80:20 train-test split, and model evaluation. The label distribution was dominated by 10,675 negative records, followed by 1,912 positive and 282 neutral records.

The model achieved 84.50% accuracy, macro precision of 0.5410, macro recall of 0.3767, and macro F1-score of 0.3819. Although the negative class achieved an F1-score of 0.9147, the positive class reached only 0.2311 and the neutral class was not recognized. The results confirm that overall accuracy must be interpreted together with macro and class-level metrics when the data are imbalanced. The Streamlit implementation successfully presented the distribution, word clouds, temporal and source patterns, confusion matrix, evaluation metrics,

and new-text predictions. Future work should improve class balance and label quality to increase recognition of positive and neutral opinions.

References

- [1] I. Permana and K. D. Maani, "Publication trend of public sentiment towards Indonesia government policies," *Sinkron*, vol. 8, no. 3, pp. 2061–2069, 2024, doi: 10.33395/sinkron.v8i3.13843.
- [2] D. D. Putri, G. F. Nama, and W. E. Sulistiono, "Analisis sentimen kinerja Dewan Perwakilan Rakyat (DPR) pada Twitter menggunakan metode Naive Bayes Classifier," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 10, no. 1, 2022, doi: 10.23960/jitet.v10i1.2262.
- [3] N. W. A. S. Aprilia and A. R. Isnain, "Analisis sentimen terhadap media sosial Twitter dengan kasus kampanye anti-korupsi di Indonesia menggunakan Naive Bayes," *Jurnal Media Informatika Budidarma*, vol. 8, no. 2, p. 695, 2024, doi: 10.30865/mib.v8i2.7582.
- [4] M. Fadli, A. Triayudi, and E. T. E. Handayani, "Sentiment analysis on Twitter social media application on fuel oil price hike using Naïve Bayes and decision tree algorithms," *SANA: Journal of Blockchain, NFTs and Metaverse Technology*, vol. 2, no. 2, pp. 114–122, 2024, doi: 10.58905/sana.v2i2.271.
- [5] J. Rama Dani, "Analisis sentimen komentar YouTube terhadap kenaikan tunjangan DPR RI menggunakan Naïve Bayes, SVM, dan Random Forest," *Building of Informatics, Technology and Science (BITS)*, vol. 7, no. 3, 2025, doi: 10.47065/bits.v7i3.8513.
- [6] T. Ismardani and Z. Fatah, "Sentiment analysis of YouTube user comments on government policies using the Naïve Bayes method," *Journal of Data Insights*, vol. 3, no. 2, pp. 113–123, 2025.
- [7] Z. R. Pratama et al., "Analisis sentimen publik terhadap isu tunjangan DPR menggunakan Brand24: Studi kasus strategi komunikasi publik dalam menghadapi konflik di media sosial," *Jurnal Multidisiplin Ilmu Akademik*, vol. 2, no. 2, 2025, doi: 10.61722/jmia.v2i6.7382.
- [8] I. Hulu and H. Budiati, "Implementasi text mining pada pengukuran sentimen opini masyarakat terhadap Universitas Kristen Immanuel," *Jurnal Sains dan Komputer*, vol. 7, no. 2, pp. 1–7, 2023, doi: 10.61179/jurnalinfact.v7i02.445.
- [9] W. Trisnawati and A. Wibowo, "Sentiment analysis of ICT service user using Naive Bayes Classifier and SVM methods with TF-IDF text weighting," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 3, pp. 709–719, 2024, doi: 10.52436/1.jutif.2024.5.3.1784.
- [10] S. A. Nugraha, "Penerapan lexicon based untuk analisis sentimen masyarakat Indonesia terhadap Danantara," *Jurnal Mahasiswa Teknik Informatika*, vol. 9, no. 3, 2025.
- [11] Y. I. Kurniawan et al., "Naive Bayes Classifier with SMOTE for sentiment analysis of Blibli app reviews on the Google Play Store," *Jurnal Penelitian Inovatif*, vol. 5, no. 3, pp. 2675–2688, 2025, doi: 10.54082/jupin.1842.
- [12] S. Raschka, Y. H. Liu, and V. Mirjalili, *Machine Learning with PyTorch and Scikit-Learn: Develop Machine Learning and Deep Learning Models with Python*. Birmingham, U.K.: Packt Publishing, 2022.
- [13] Y. William and R. Loisa, "Peran media sosial dalam membangun opini publik: Studi kasus akun media sosial DPR RI," *Kiwari*, vol. 4, no. 3, pp. 426–434, 2025.