Indonesian Journal of Data and Science

Volume 6 Issue 1
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v6i2.235

Research Article

A Comparative Study of Public Opinion on Indonesian Police: Examining Cases in the Aftermath of the Kanjuruhan Football Disaster

Purnawansyah [1], Roesman Ridwan Raja [2,*], Herdianti Darwis [3]

[1] Universitas Muslim Indonesia, Makassar, Indonesia, purnawansyah@umi.ac.id
[2] Kyushu Institute of Technology, Iizuka City, Jepang, raja.roesman-ridwan757@mail.kyutech.jp
[3] Universitas Muslim Indonesia, Makassar, Indonesia, herdianti.darwis@umi.ac.id

Correspondence should be addressed to Roesman Ridwan Raja; raja.roesman-ridwan757@mail.kyutech.jp

Received 10 January 2025; Accepted 20 March 2025; Published 31 July 2025

© Authors 2025. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:

This research explores public sentiment towards the Indonesian police using sentiment analysis and machine learning techniques. The study addresses the challenge of understanding public opinion based on social media comments related to significant police cases. The aim is to compare reported satisfaction levels with actual public sentiment. Utilizing the Indonesian RoBERTa base IndoLEM sentiment classifier, comments were analyzed and preprocessed. The classification was conducted using Random Forest (RF) and Complement Naive Bayes (CNB) models, incorporating unigram and bi-gram features. Oversampling techniques were applied to handle data imbalance. The best-performing model, Random Forest with bi-gram features, achieved high evaluation scores, including a precision of 0.91 and accuracy of 0.91. The findings reveal significant insights into public opinion, contributing to improved law enforcement strategies and public trust.

Keywords: Indonesian Police, Kanjuruhan, Machine Learning, Public Opinion, Sentiment Analysis.

1. Introduction

The role of law enforcement agencies in maintaining public order and safety is critical to the stability and development of any nation. In Indonesia, the National Police (Polri) plays a pivotal role in this regard. Public opinion of the police force can significantly influence the efficacy of their operations and the overall perception of safety within the community. Understanding this dynamic becomes even more crucial in the aftermath of significant incidents involving law enforcement, such as the Kanjuruhan football disaster of 2022 [1].

The Kanjuruhan tragedy, which resulted in numerous casualties and widespread public outcry, placed the Indonesian police under intense scrutiny. Public sentiment towards the police in the aftermath of such incidents often reveals deeper insights into societal trust and the perceived effectiveness of law enforcement. This study aims to explore the public opinion of the Indonesian police in the context of major cases following the Kanjuruhan disaster.

A recent survey conducted by Litbang Kompas from October 23-31, 2023, indicates a high level of public satisfaction with the police force, with 87.8% of respondents expressing overall satisfaction with Polri's performance [2], and 92.1% satisfied with public service delivery by Polri [3]. These figures suggest a generally positive perception of the police among the public. However, this study seeks to investigate whether these survey results align with the public sentiment observed in social media and other online platforms regarding nine major cases involving the police post-Kanjuruhan.

To achieve this, the research will employ the Indonesian RoBERTa base IndoLEM sentiment classifier model to analyze and label sentiments expressed in comments related to these nine cases [4]. The sentiment data will then be preprocessed and fed into Random Forest [5] and Complement Naive Bayes models for further analysis and evaluation.

By juxtaposing the quantitative survey results with qualitative sentiment analysis [6] from social media, this study aims to provide a comprehensive understanding of public opinion towards the Indonesian police. The findings are expected to offer valuable insights into the consistency between reported satisfaction levels and actual public sentiment, thereby contributing to the broader discourse on law enforcement efficacy and public trust in Indonesia. This research holds the potential to inform policy-making and strategic communication efforts within the Polri, enhancing their engagement with the community and improving their public image in the long run.

2. Method:

Research Flow

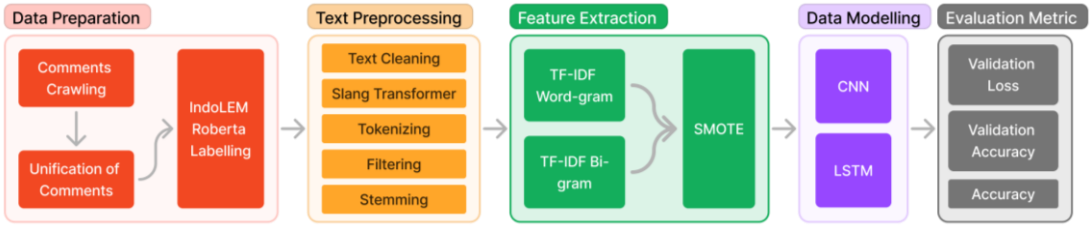


Figure 1. Research Flow

The research flow, as depicted in **Figure 1**, encompasses several key stages: data preparation, text preprocessing, feature extraction, data modeling, and evaluation metrics. Initially, comments are crawled and unified before being labeled using the indoLEM Roberta model. In the text preprocessing phase, techniques such as text cleaning, slang transformation, tokenizing, filtering, and stemming are employed to refine the data. The refined text then undergoes feature extraction using TF-IDF word-gram and BP-gram methods, along with SMOTE for balancing. Subsequently, the processed features are modeled using Random Forest and Complement Naive Bayes algorithms. Finally, the models are evaluated based on metrics including validation loss, validation accuracy, and overall accuracy.

Police Cases

This study focuses on several high-profile police-related cases that have had a significant impact on public opinion towards the Indonesian police following the Kanjuruhan disaster of 2022. These cases were selected due to their perceived influence on societal attitudes and the lingering trauma associated with the Kanjuruhan incident. The cases span various years, platforms, and media publishers, providing a comprehensive overview of public sentiment. **Table 1** below lists the cases, detailing the year, post link, media, and platform for each incident.

Table 1. Police Cases following the Kanjuruhan disaster of 2022

No.	Case	Year	Post Link	Media	Platform
1	Afif Maulana Persecution Case	2024	www.instagram.com/p/C8_nNypyt1Z	Narasi Newsroom	Instagram
2	Arrest of Pegi Setiawan, Suspect in Vina Cirebon Murder	2024	www.instagram.com/p/C7grxkzhfOE	Narasi Newsroom	Instagram
3	Release of Pegi Setiawan, Suspect in Vina Cirebon Murder	2024	www.instagram.com/p/C9JZ1bbSSBT	Narasi Newsroom	Instagram
4	Arrest of village head admits police robbed students	2024	www.youtube.com/watch?v=PIdKwMxeYXc	Tribun MedanTV	YouTube
5	Police Investigating Jessica's Case: Promoted, Gets Stars (Ice Cold Documentary)	2023	www.youtube.com/watch?v=rmvk_XjZh_k	Narasi Newsroom	YouTube

No.	Case	Year	Post Link	Media	Platform
6	Collision of University of Indonesia Students by Retired Police Officer	2023	www.instagram.com/p/Cn9XpIZhsZY	Narasi Newsroom	Instagram
7	Polemic about the dismissal of the Head of the North Kalimantan Police Propam Division when Handling Illegal Fuel Cases	2023	www.youtube.com/watch?v=AMAlZWrN6aY	Kompas	YouTube
8	Kanjuruhan Incident: Civil Society Coalition Finds 12 Initial Findings	2022	www.instagram.com/p/CjpOTlkrBDU	Narasi Newsroom	Instagram
9	Police Prostrate to Apologize to Kanjuruhan Victims	2022	www.youtube.com/watch?v=XxjBVunR4tA	CNN Indonesia	YouTube

Data Scrapping

In this study, data scrapping was conducted using a combination of BeautifulSoup and Selenium tools, integrated with the Chrome browser driver [7]. This approach was employed to extract comments related to each case from various social media platforms. BeautifulSoup facilitated the parsing and extraction of HTML content, while Selenium enabled dynamic interaction with web pages, ensuring comprehensive data retrieval even from pages requiring user interactions, such as scrolling or clicking. This robust methodology ensured the collection of a rich and diverse dataset of public comments for each of the highlighted cases.

Data Pre-processing Methods

In order to prepare the dataset for feature extraction, several data pre-processing methods were employed. These steps ensured that the data was clean, normalized, and suitable for analysis. The pre-processing pipeline included text cleaning, slang transformation, tokenizing, filtering, and stemming, all implemented using the Sastrawi library.

Text Cleaning

Text cleaning involves removing unwanted characters, symbols, and formatting from the text data. This step includes converting all text to lowercase, removing punctuation, numbers, special characters, and any HTML tags that might be present. The goal is to standardize the text and remove any extraneous elements that do not contribute to the sentiment analysis.

Slang Transformer

Slang transformation is the process of converting slang terms, abbreviations, and informal language into their standard forms [8]. This is particularly important in social media comments, where informal language is common. Using a predefined dictionary of slang terms, each occurrence of slang is replaced with its formal equivalent, ensuring consistency in the dataset.

Tokenization

Tokenization involves breaking down the text into individual words or tokens [9]. This step is crucial for transforming the text data into a format that can be easily analyzed and processed by machine learning algorithms. Each comment is split into its constituent words, which are then used for further processing and feature extraction.

Filtering

Filtering [10] is the process of removing stop words, which are common words that do not carry significant meaning for sentiment analysis, such as "and", "the", "is", etc. By eliminating these words, the focus is shifted to the more meaningful words that are more likely to influence the sentiment. This step enhances the efficiency and effectiveness of the subsequent analysis.

Stemming

Stemming [11] reduces words to their root forms using the Sastrawi library [12], [13], an Indonesian-language-specific stemmer. This step ensures that different forms of a word (e.g., "*berlari*", "*lari*") are treated as the same word ("*lari*"), thereby reducing redundancy and improving the accuracy of the sentiment analysis. Stemming is crucial for handling the morphological variations of words in the Indonesian language.

TF-IDF Feature Extraction

In this study, Term Frequency-Inverse Document Frequency (TF-IDF) [14] is utilized for feature extraction, employing both unigram and bi-gram approaches. This method transforms the text data into numerical representations that reflect the importance of each term within the document set.

The IDF is calculated as shown in Equation 1:

$$IDF(t) = log\left(\frac{N}{DF}\right) \quad (1)$$

Shown in Equation 1 where N is the number of documents and DF is the number of documents containing term (t). TF-IDF is a great way to transform text Representation of information in the Vector Space Model (VSM). Suppose we have a 200-word document and from these 200 words the Police appears 10 times as often as the time window, 10/250=0.04. Suppose you have 50,000 documents, only 500 of which contain a "Police". From IDF (Police) = 50000/500 = 100 and TF-IDF (Police) is 0.04 * 100 = 4. It should be understood that the more frequently a word occurs in a document, the higher the term frequency of the document, and the less frequently a word occurs in a document, the higher the Keyword Importance (IDF) to be searched. For a given document .TF-IDF is nothing but the multiplication of term frequency (TF) and inverse document frequency (IDF). To compute TF-IDF:

$$TF - IDF = tf * idf \quad (2)$$

Two separate datasets are created for further processing: one using unigram (single words) and the other using bi-gram (two consecutive words) features. This dual approach allows for capturing both individual word importance and contextual word pairs, enhancing the depth of the sentiment analysis.

Indonesian RoBERTa IndoLEM Sentiment Classifier

The Indonesian RoBERTa Base IndoLEM Sentiment Classifier [4] is a text classification model based on the RoBERTa architecture. Initially, it was the pre-trained Indonesian RoBERTa Base model, which was later fine-tuned using the IndoLEM Sentiment Analysis dataset [15], containing Indonesian tweets and hotel reviews. A 5-fold cross-validation was conducted, with splits provided by the dataset authors. This particular model was trained on fold 0. Models trained on folds 1 through 4 are also available via their respective links.

Imbalance Data Handling with SMOTE

Before the data is modeled, based on **Figure 2** below, the scrapped dataset is imbalanced, which strongly affects subsequent evaluation results. Therefore, treatment of imbalanced data is performed before proceeding with modeling. One way to handle imbalanced datasets is to oversample the minority classes [16]. The easiest way is to duplicate the examples in the minority class, but these examples add no new information to the model. Alternatively, you can synthesize new examples from existing examples. In this paper, the dataset over-sampled with Synthetic Minority Oversampling Technique (Synthetic Minority Oversampling Technique), or SMOTE for short [17], [18]. This is a type of minority tier data multiplication.

Data Modelling

In this section, we explore the use of Random Forest and Complement Naive Bayes as classifiers in our research. These models are selected for their robustness and effectiveness in handling textual data for sentiment analysis.

Random Forest

Random Forest [19] is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes for classification tasks. It is highly effective for text data due to its ability to handle large datasets and complex interactions between features. In this study, Random Forest is utilized to process the preprocessed textual data, learning to identify patterns and dependencies that signify sentiment within the comments. Its ensemble nature helps in reducing overfitting and improving the model's generalization capability.

Complement Naive Bayes

Complement Naive Bayes (CNB) is a variant of the standard Naive Bayes algorithm, particularly suited for imbalanced data [20], [21]. CNB is designed to handle class imbalance more effectively by focusing on the complements of each class rather than the classes themselves. This makes it well-suited for text classification tasks where some sentiments may be less frequent. In this research, CNB is employed to analyze the textual data, leveraging its ability to handle imbalance and improve classification performance by estimating probabilities that better reflect the underlying data distribution.

By utilizing both Random Forest and Complement Naive Bayes models, this study aims to comprehensively evaluate and compare their effectiveness in sentiment classification, providing insights into the most suitable approach for analyzing public opinion on the Indonesian police.

Evaluation Metrics

In this research, we utilize the confusion matrix [22] and classification report [23] to assess the performance of our models. The confusion matrix provides a detailed breakdown of the model's predictions, showing the counts of true positives, true negatives, false positives, and false negatives. This allows for a comprehensive understanding of the model's accuracy, precision, recall, and F1-score.

The classification report complements the confusion matrix by summarizing these metrics, providing a clear overview of the model's performance across different classes. Precision measures the accuracy of positive predictions, recall measures the ability to identify positive instances, and the F1-score provides a harmonic mean of precision and recall.

By employing the confusion matrix and classification report, we ensure a rigorous assessment of our models' capabilities in accurately classifying sentiments related to public opinion on the Indonesian police.

3. Results and Discussion

Dataset Preparation

In this study, we collected sample comments from various social media posts related to significant police cases. The number of sample comments and the sentiment of the posts towards the police are detailed in **Table 2** below. These cases were chosen due to their substantial impact on public opinion regarding the police, following the Kanjuruhan disaster of 2022.

Table 2. Dataset Summary

No.	Case	Number of Sample Comments taken	Post's Sentiment towards the police
1	Afif Maulana Persecution Case	233	Negative
2	Arrest of Pegi Setiawan, Suspect in Vina Cirebon Murder	238	Positive
3	Release of Pegi Setiawan, Suspect in Vina Cirebon Murder	235	Neutral
4	Arrest of village head admits police robbed students	500	Positive
5	Police Investigating Jessica's Case: Promoted, Gets Stars (Ice Cold Documentary)	500	Negative
6	Collision of University of Indonesia Students by Retired Police Officer	235	Negative
7	Polemic about the dismissal of the Head of the North Kalimantan Police Propam Division when Handling Illegal Fuel Cases	500	Positive
8	Kanjuruhan Incident: Civil Society Coalition Finds 12 Initial Findings	94	Negative
9	Police Prostrate to Apologize to Kanjuruhan Victims	500	Positive

The dataset includes a diverse range of sentiments, with comments reflecting negative, and positive opinions. This variety ensures a comprehensive analysis of public sentiment towards the police in the aftermath of significant events.

Exploratory Data Analysis

Sentiment Polarity

Sentiment analysis was conducted to examine the public's perception of the police's handling of the nine cases. The analysis was performed using a sentiment analysis tool that utilizes a lexicon-based approach to classify text as positive, negative, or neutral. The tool considers a variety of factors, including word choice, punctuation, and sentence structure, to determine the sentiment of a given text.

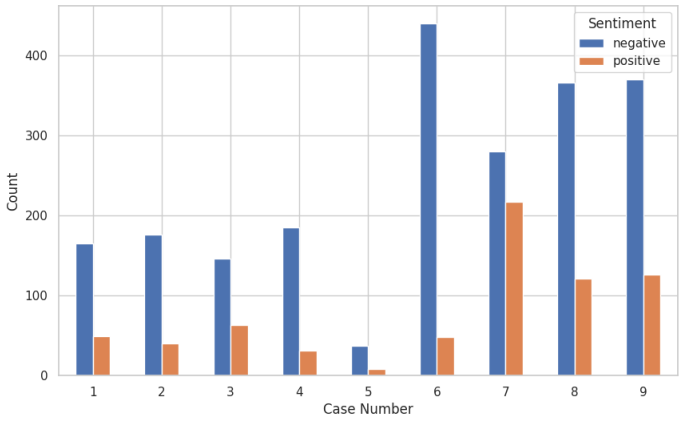


Figure 2. Sentiment Polarity by Case Number

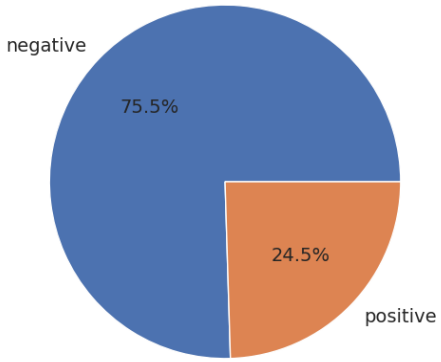


Figure 3. Sentiment Polarity

Figure 2 illustrates the sentiment distribution across nine cases. The x-axis represents the case number, while the y-axis denotes the count of posts. The bar chart reveals a predominance of negative sentiment across most cases, with all cases except case number 7 exhibiting a significantly higher count of negative posts compared to positive ones. Conversely, cases number 7 demonstrate a more balanced sentiment distribution, with a higher proportion of positive sentiment against other cases.

Figure 3 provides a comprehensive overview of the overall sentiment polarity. The pie chart indicates that approximately 75.5% of the total posts exhibit negative sentiment, while the remaining 24.5% express positive sentiment. This finding underscores the prevailing negative sentiment within the dataset.

Dataset Vocabulary

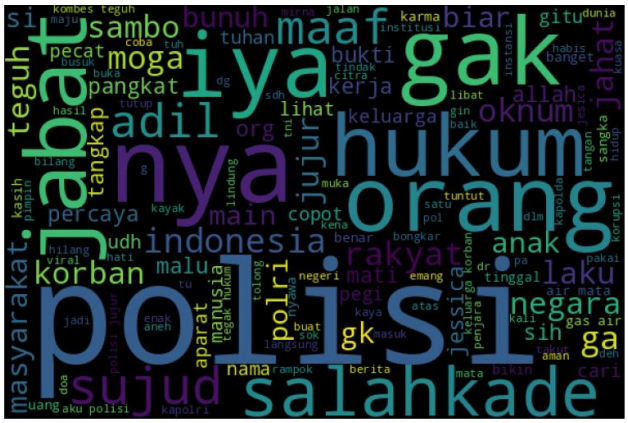


Figure 4. Word cloud of comments data

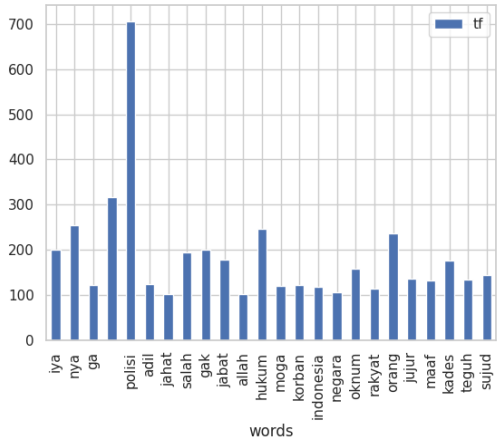


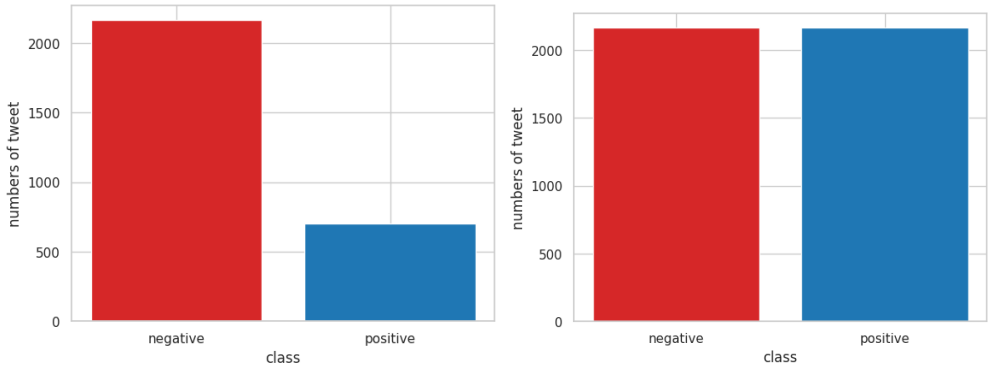
Figure 5. Frequency of the words

Figure 4 presents a word cloud visualization of the dataset vocabulary. The word cloud highlights the most frequently used words, with larger font sizes indicating higher frequency. The prominent words include "*negara*" (country), "*rakyat*" (people), "*hukum*" (law), "*korban*" (victim), "*indonesia*" (Indonesia), "*oknum*" (individual), "*allah*" (God), "*moga*" (hopefully), "*maaf*" (sorry), and "*kades*" (village head). These words reflect the central themes and issues surrounding the nine cases analyzed in the study.

Figure 5 depicts the frequency distribution of the words in the dataset. The x-axis represents the word frequency, while the y-axis denotes the number of words. The graph reveals a power law distribution, with a small number of words appearing frequently and a large number of words appearing infrequently. This distribution is characteristic of natural language data.

Imbalance Data Handling

Imbalanced datasets pose a significant challenge in sentiment analysis, as they can lead to biased model predictions.



(a) (b)

Figure 6. (a). Sentiment Polarity Before Oversampling (b). Sentiment Polarity After Oversampling

Figure 6 illustrates the sentiment polarity distribution before and after oversampling. In Figure 6 (a), it is evident that the dataset is heavily skewed towards negative sentiments, with a much smaller proportion of positive comments. To address this imbalance, oversampling techniques were applied, resulting in a more balanced dataset as shown in Figure 6 (b). This adjustment ensures that the model receives an equal representation of both positive and negative sentiments, thereby enhancing its ability to accurately classify sentiments in the data.

Classification Evaluation

The performance of the classification models used in this research is comprehensively analyzed through evaluation metrics. As shown in Figure 7, the confusion matrix results indicate that the Random Forest (RF) model with bi-grams achieves the highest number of true positives and true negatives, showcasing its accuracy in correctly classifying sentiments. On the other hand, the Complement Naive Bayes (CNB) model with unigrams records a higher count of false positives and false negatives, highlighting its limitations in sentiment detection.

In Figure 8, a detailed classification report further emphasizes the differences in model performance. The Random Forest model with bi-grams stands out with the highest precision, recall, F1-score, and accuracy, confirming its effectiveness in sentiment analysis. Conversely, the Complement Naive Bayes model with unigrams shows lower scores, particularly in recall and F1-score, indicating its relative ineffectiveness.

Overall, the Random Forest model with bi-grams demonstrates the best performance, while the Complement Naive Bayes model with unigrams proves to be the least effective. These findings highlight the critical role of model selection and feature engineering in achieving reliable sentiment classification.

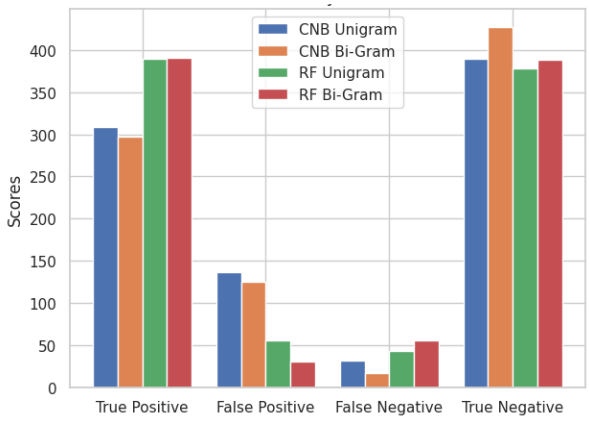


Figure 7. Confusion Matrix by Model and Metric

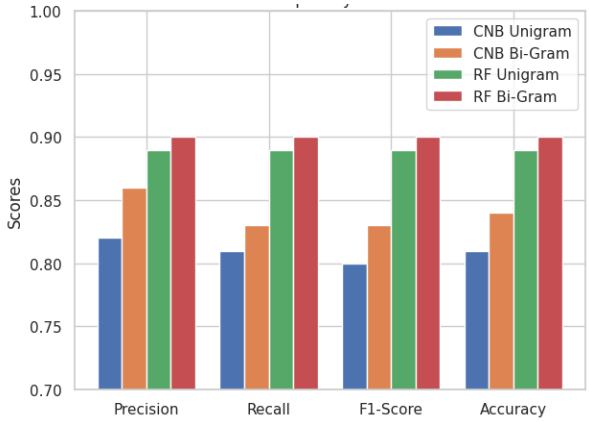


Figure 8. Classification Report by Model and Metric

4. Conclusion

This research provides a comprehensive analysis of public sentiment towards the Indonesian police through a combination of qualitative and quantitative methods. By employing the Indonesian RoBERTa base IndoLEM sentiment classifier for sentiment labelling, and utilizing Random Forest and Complement Naive Bayes classifiers, we have effectively captured the complexities of public opinion as expressed in social media comments.

Our findings indicate that the Random Forest model with bi-grams is the most effective classifier, demonstrating superior performance across precision, recall, F1-score, and accuracy metrics. Conversely, the Complement Naive Bayes model with unigrams was found to be less effective. This underscores the importance of model selection and feature engineering in sentiment analysis.

Through the juxtaposition of survey results and sentiment analysis, this study reveals valuable insights into the consistency between reported satisfaction levels and actual public sentiment. The results contribute to the broader discourse on law enforcement efficacy and public trust in Indonesia, providing a foundation for informed policy-making and strategic communication efforts within the Polri.

Ultimately, this research highlights the potential for advanced sentiment analysis techniques to enhance understanding of public opinion, guiding improvements in community engagement and public image for the Indonesian police.

Acknowledgments

This research was funded by a grant from Universitas Muslim Indonesia.

References:

[1] F. Sokoy, “Kanjuruhan Indonesia Football Tragedy (Culture, Management, Governance, and Justice),” International Journal of Human Movement and Sports Sciences, vol. 11, no. 4, pp. 753–761, 2023, doi: 10.13189/saj.2023.110408.

[2] S. Aprilia, A. Wijaya, and S. Suryadi, “Efektivitas Website Sebagai Media E-Government dalam Meningkatkan Pelayanan Elektronik Pemerintah Daerah (Studi Pada Website Pemerintah Daerah Kabupaten Jombang),” Wacana, Jurnal Sosial dan Humaniora, vol. 17, no. 3, pp. 126–135, Jul. 2014, doi: 10.21776/ub.wacana.2014.017.03.3.

[3] A. Saputra, “Tingkat Kepercayaan Masyarakat terhadap Kinerja Polri Tahun 2021,” Jurnal Litbang Polri, vol. 25, no. 1, pp. 23–34, Apr. 2022, doi: 10.46976/.v25i1.174.

[4] L. Sandra, “Are University Students Independent: Twitter Sentiment Analysis of Independent Learning in Independent Campus Using RoBERTa Base IndoLEM Sentiment Classifier Model,” 2021 International

Seminar on Machine Learning, Optimization, and Data Science, ISMODE 2021, pp. 249–253, 2022, doi: 10.1109/ISMODE53584.2022.9743110.

[5] Nurul Amelina Nasharuddin and Nurul Shuhada Zamri, "Non-Parametric Machine Learning for Pollinator Image Classification: A Comparative Study," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 34, no. 1, pp. 106–115, Nov. 2023, doi: 10.37934/araset.34.1.106115.

[6] Harika vanam and Jeberson Retna Raj, "Novel Method for Sentiment Analysis in Social Media Data Using Hybrid Deep Learning Model," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 32, no. 1, pp. 272–289, Aug. 2023, doi: 10.37934/araset.32.1.272289.

[7] S. Singh, G. Srivastava, V. Dubey, and G. R. Mishra, "Triggering an Email Alert Based on Price Comparison by Web Scraping Using Python," in Lecture Notes in Electrical Engineering, vol. 1065, Singapore: Springer, 2024, pp. 145–155. doi: 10.1007/978-981-99-4795-9_14.

[8] S. C. Permatasari and C. H. Karjo, "The Influence of Fandom Language in the Word Formation of Indonesian Internet Slangs," E3S Web of Conferences, vol. 388, p. 04040, May 2023, doi: 10.1051/e3sconf/202338804040.

[9] Peter E. Shawky, Saleh Mesbah ElKaffas, and Shawkat K Guirguis, "Effect of typos on text classification accuracy in word and character tokenization," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 40, no. 2, pp. 152–162, Feb. 2024, doi: 10.37934/araset.40.2.152162.

[10] P. Garg, "A systematic review on spam filtering techniques based on natural language processing framework," Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering, pp. 30–35, 2021, doi: 10.1109/Confluence51648.2021.9377042.

[11] Rianto, "Improving stemming techniques for non-formal indonesian sentences using incorbiz," ICIC Express Letters, vol. 15, no. 1, pp. 67–74, 2021, doi: 10.24507/icicel.15.01.67.

[12] X. Y. Tan and C. W. Tan, "From Complexity to Clarity: Tatabahasa-Centric Lemmatization in Malay Texts," in 2024 3rd International Conference on Digital Transformation and Applications (ICDXA), IEEE, Jan. 2024, pp. 1–5. doi: 10.1109/ICDXA61007.2024.10470673.

[13] A. Amalia, M. S. Lidya, A. Andrian, E. M. Zamzami, and S. M. Hardi, "OLCBot: Dissemination Of Interactive Information Related To Indonesia's Omnibus Law With The Implementation of Fuzzy String Matching Algorithm and Sastrawi Stemmer," in 2022 6th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM), IEEE, Nov. 2022, pp. 178–181. doi: 10.1109/ELTICOM57747.2022.10037966.

[14] B. Bakiyev, "Method for Determining the Similarity of Text Documents for the Kazakh language, Taking Into Account Synonyms: Extension to TF-IDF," SIST 2022 - 2022 International Conference on Smart Information Systems and Technologies, Proceedings, 2022, doi: 10.1109/SIST54437.2022.9945747.

[15] 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, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[16] A. Sharma, A. Purohit, and H. Mishra, "A Survey on Imbalanced Data Handling Techniques for Classification," International Journal of Emerging Trends in Engineering Research, vol. 9, no. 10, pp. 1341–1347, Oct. 2021, doi: 10.30534/ijeter/2021/089102021.

[17] Nur Liana Ab Majid and Syahid Anuar, "Machine Learning Modelling for Imbalanced Dataset: Case Study of Adolescent Obesity in Malaysia," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 36, no. 1, pp. 189–202, Dec. 2023, doi: 10.37934/araset.36.1.189202.

[18] Al-Ogaidi Ali Hameed Khalaf, Raihani Mohamed, and Abdul Rafiez Abdul Raziff, "Detection Model for Ambiguous Intrusion using SMOTE and LSTM for Network Security," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 39, no. 2, pp. 191–203, Feb. 2024, doi: 10.37934/araset.39.2.191203.

[19] Dhasaradhan Kaveripakam and Jaichandran Ravichandran, "Comparative Analysis of Machine Learning Algorithms for Diabetic Disease Identification," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 45, no. 1, pp. 40–50, May 2024, doi: 10.37934/araset.45.1.4050.

[20] S. Ismail, M. Nouman, D. W. Dawoud, and H. Reza, "Towards a lightweight security framework using blockchain and machine learning," Blockchain: Research and Applications, vol. 5, no. 1, p. 100174, Mar. 2024, doi: 10.1016/j.bcra.2023.100174.

[21] P. Rakshit, S. Gupta, and T. Das, "Sentiment Analysis to Find Sentence Polarity on Tweet Data," in Lecture Notes in Networks and Systems, vol. 498, Singapore: Springer, 2023, pp. 197–202. doi: 10.1007/978-981-19-5090-2_19.

[22] Nurul Ehsan Ramli, Zainor Ridzuan Yahya, and Nor Azinee Said, "Confusion Matrix as Performance Measure for Corner Detectors," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 29, no. 1, pp. 256–265, Dec. 2022, doi: 10.37934/araset.29.1.256265.

[23] Alawiyah Abd Wahab, Huda H. Ibrahim, Shehu M. Sarkintudu, Maslinda Mohd. Nadzir, and Zhamri Che Ani, "Comparative Performance of Machine Learning Algorithms for Predicting Future Committer in Blockchain Projects," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 34, no. 2, pp. 72–87, Dec. 2023, doi: 10.37934/araset.34.2.7287.