2025 International Conference on Computer, Control, Informatics and its Applications (IC3INA) | 979-8-3315-9188-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/IC3INA68387.2025.11325554

Comparative Lexicon-Based Sentiment Analysis Towards Indonesian *'Cek Kesehatan Gratis'* Program in X

Fatur Rozak
Faculty of Mathematics and Natural Sciences
University of Lampung
Lampung, Indonesia
faturrozak759@gmail.com

Oja Widiyatama
Faculty of Mathematics and Natural Sciences
University of Lampung
Lampung, Indonesia
ojawidiyatama795@gmail.com

Andri Fachrur Rozie
Research Center for Data and Information Science
National Research and Innovation Agency (BRIN)
Bandung, Indonesia
andr035@brin.go.id

Ekasari Nugraheni
Research Center for Data and Information Science
National Research and Innovation Agency (BRIN)
Bandung, Indonesia
ekasari.nugraheni@brin.go.id

Dian Kurniasari
Faculty of Mathematics and Natural Sciences
University of Lampung
Lampung, Indonesia
dian.kurniasari@fmipa.unila.ac.id

***Abstract*—Sentiment analysis aims to understand a person's perspective or attitude toward a particular topic. One of its main challenges is the labeling approach, especially in the context of the Indonesian language. This study evaluates six labeling approaches: manual labeling, InSet Lexicon, modified InSet Lexicon, VADER, VADER-Translate, and SentIL. Each approach classifies data into three categories: positive, neutral, and negative. The dataset consists of 8,393 posts from the X application related to the Free Health Check Program. Three models were used for the evaluation process: Support Vector Machine (SVM), Convolutional Neural Network (CNN), and IndoBERT. The results show that the combination of the modified InSet Lexicon and IndoBERT produced the best performance, achieving an accuracy of 94.04%, precision of 93.29%, recall of 93.47%, and an F1-score of 93.31%. These findings highlight the importance of selecting lexicon-based methods that align with the study domain and the linguistic characteristics of the data to improve the accuracy and reliability of sentiment analysis in public policy.**

***Index Terms*—sentiment analysis, free health check, lexicon, comparative.**

I. Introduction

Sentiment analysis aims to identify and classify public opinions about an object into three sentiment categories: positive, neutral, or negative [1]. In the context of public policy, this method plays a crucial role in rapidly and data-drivenly measuring public perceptions, especially through social media, which serves as a space for digital public expression. The results of sentiment analysis can provide strategic feedback for policymakers in evaluating programs, developing public communication narratives, and enhancing the responsiveness of public services.

However, one of the main challenges in sentiment analysis is the data labeling task, which involves assigning sentiment categories to text content [2]. One commonly used labeling approach is the lexical dictionary method, where words in the text are analyzed based on a list of sentiment-laden words. The issue that arises is that the availability and quality of Indonesian lexical dictionaries are still limited, and most of them are general in nature, not yet tailored to specific contexts or domains.

One of the most widely used lexicons is the Valence Aware Dictionary and sEntiment Reasoner (VADER), a sentiment dictionary well known for analyzing English-language social media texts [3]. In addition, there is the InSet Lexicon, an Indonesian sentiment dictionary that has not yet been adapted to any specific domain context [4]. Meanwhile, the Sentiment Indonesian Lexicon (SentIL) is a newly curated lexicon that has not yet been extensively tested in academic research [5]. These three approaches have certain limitations that open opportunities for further development. For example, VADER needs to be translated to be relevant for the Indonesian language. The InSet Lexicon can be modified to better align with the context of the data. SentIL, as a general lexical dictionary, needs to be empirically evaluated for its accuracy using data from specific domains. Therefore, to obtain valid and contextual results, manual labeling is still required as ground truth to measure the performance of these automated lexicons.

Therefore, this study aims to compare the six labeling approaches. These approaches are applied to data collected through scraping from the social media platform X using the keyword *"cek, kesehatan gratis"*. X was chosen because it is a dominant platform for expressing public opinion in real time in Indonesia and has been widely used in sentiment analysis research on public policy [6].

The policy context analyzed in this study is the *"Cek Kesehatan Gratis* (CKG)" or Free Health Check program,

which was launched by the 8th President of the Republic of Indonesia, Prabowo Subianto, on February 10, 2025. This program has received widespread public attention on social media, making it a relevant object for sentiment analysis.

To evaluate each labeling approach, three classification models were used: Support Vector Machine (SVM), Convolutional Neural Network (CNN), and IndoBERT. These three models were selected to represent traditional classification, deep learning, and transformer-based approaches, respectively. This design aims to ensure that the evaluation of sentiment lexicons can be thoroughly tested across various types of text classification models. Through this approach, the study is expected to identify the best combination of lexicon and model for Indonesian sentiment analysis, as well as provide empirical contributions to improving the quality of sentiment labeling and supporting data-driven policymaking.

II. Related Work

Sentiment analysis has been widely used to understand public opinion on government policies. For example, national policies have been evaluated by analyzing public reactions on the social media platform X [6].Public sentiment analysis is used to measure public attitudes toward the PPKM (Community Activity Restrictions) policy in Indonesia during the COVID-19 pandemic [7]. These studies demonstrate that sentiment analysis is an important tool for data-driven policy evaluation. Several studies have also focused on comparing sentiment data labeling approaches. One study compared automatic labeling using VADER and TextBlob [2].

When using the lexical approach, the InSet Lexicon was applied to an SVM model and achieved an accuracy of 89.9% [8]. The VADER lexicon, applied in another study, achieved an accuracy of 96% [3]. Meanwhile, SentIL, a newly developed lexicon created through a semi-automated curation process, reached an accuracy of up to 95.7% [5]. On the other hand, manual labeling is still considered the benchmark method because it ensures high-quality labels, as demonstrated in previous studies [9].

Various classification models have also been used in sentiment analysis. SVM has been shown to outperform Naive Bayes with an accuracy of 94.38% due to its ability to handle high-dimensional data [10]. CNN has also demonstrated good performance with an accuracy of 87.58% [1]. IndoBERT, as a transformer-based model specifically trained in the Indonesian language, achieved the highest accuracy of 95.16% [11].

However, to date, no study has comprehensively compared six labeling methods (manual, InSet, modified InSet, VADER, translated VADER, and SentIL) with three classification models (SVM, CNN, and IndoBERT) within a single experiment specifically for sentiment analysis of public policy in the Indonesian language. Therefore, this study was conducted to fill this gap and provide a meaningful contribution to identifying the most effective labeling methods and classification models in the local context.

III. Methodology

Fig. 1 presents the experimental framework of this study. The details of each stage will be explained in the following sections.

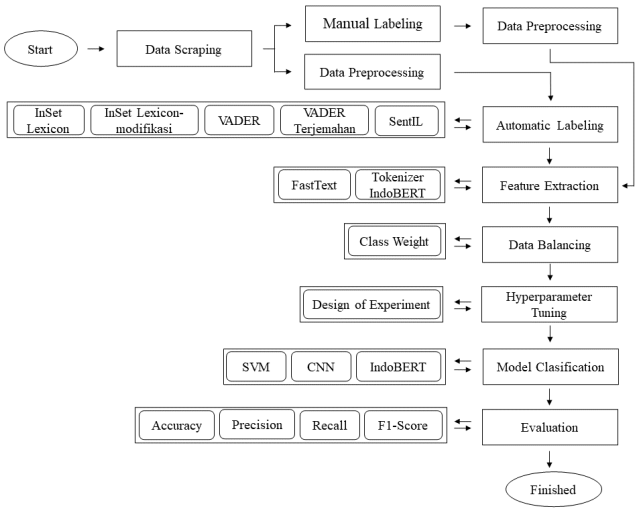

Fig. 1. Framework

A. Dataset

Data obtained from web scraping results on social media platform X related to public opinion on the CKG program. Data was taken from October 20, 2024 to March 31, 2025 with the keyword *"cek kesehatan gratis"*. The results obtained were 8,393 data.

B. Data Preprocessing

Preprocessing is useful for removing unimportant words that do not match the keywords [12]. The preprocessing phase involves several steps, including text cleaning, case folding, normalization, removing stopwords, tokenization, and stemming. In order for the preprocessing results to be as expected by the researcher, we added a special dictionary for the normalization, stopword removal and stemming sections [8].

C. Labelling

Data labeling in this study was carried out using two approaches: manual labeling and lexicon-based automatic labeling [13]. In the manual labeling stage, the process was conducted by two annotators with an academic background in Mathematics specializing in Statistics, who possess an understanding of data analysis and the research context. The use of two annotators was chosen considering the research's resource limitations, yet it was deemed sufficient to measure reliability.

Before the annotation process, the annotators were provided with a labeling guideline that had been prepared beforehand to explain the definition of each sentiment category (positive, negative, neutral) along with examples, ensuring consistent

perception between annotators. The guideline used the Innacoved dataset from the study by [14] as a reference.

If differences in labeling occurred due to disagreements or data ambiguity, discussions were held until a consensus was reached. The reliability of the manual labeling results, which serve as the ground truth, was tested using Cohen's Kappa, a statistical measure used to evaluate inter-annotator agreement [15]. The calculation of the Kappa value is based on Table I and Table II.

TABLE I. CONTINGENCY TABLE OF ANNOTATORS

Annotator 2	**Annotator 1**			
	Positive	**Neutral**	**Negative**	**Total**
Positive	a	b	c	m_1
Neutral	d	e	f	m_2
Negative	g	h	i	m_3
Total	n_1	n_2	n_3	N

The following is the formula for Cohen's Kappa:

- Observed agreement formula:

$$P_0 = \frac{\text{Total Same Class}}{\text{Total Data}} \tag{1}$$

- Expected agreement formula:

$$P_e = \left(\frac{n_1}{N} \cdot \frac{m_1}{N}\right) + \left(\frac{n_2}{N} \cdot \frac{m_2}{N}\right) + \left(\frac{n_3}{N} \cdot \frac{m_3}{N}\right) \tag{2}$$

- Cohen's Kappa formula:

$$K = \frac{P_0 - P_e}{1 - P_e} \tag{3}$$

TABLE II. INTERPRETATION OF KAPPA COEFFICIENT VALUES

Score Kappa	**Level of Agreement Interpretation**
< 0	No agreement / Worse than chance
0.01 – 0.20	Slight agreement / Minimal consistency
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

In addition, automatic labeling was also carried out using a corpus, namely the InSet Lexicon developed by [4]. This lexicon was originally created in Indonesian but has not yet been fully adapted to the characteristics of the data used in this study. Therefore, the researchers modified the InSet Lexicon by adding new vocabulary entries to the lexicon. The addition of vocabulary was done by identifying synonyms and root words of the existing entries in the InSet Lexicon to enrich its lexical coverage. An example of the modified vocabulary is presented in Table III.

The researchers were also interested in the VADER lexicon, which is originally in English. In its labeling process, the data will first be translated into English before undergoing labeling. Another experiment was conducted using a translated version

TABLE III. SAMPLE OF MODIFIED WORDS

Words in Inset Lexicon	**Weight**	**Modified Words**	**Weight**
keluhan	-3	*keluh*	-3
giliran	3	*gilir*	3
cepat	-3	*kebut*	-3
perubahan	4	*revolusi*	4

of the VADER lexicon in Indonesian, allowing the data to be labeled without being translated into English. Another lexicon of interest is SentIL [5].

D. Feature Extraction, Data Balancing and Hyperparameter Tuning

Classification models cannot directly understand raw word or text inputs. Therefore, a feature extraction process was carried out using FastText. FastText is a word embedding-based representation method developed by Facebook AI Research [16]. This method represents words as a combination of subword segments (character n-grams), enabling it to recognize new words, informal words, or misspellings. For example, the word "diabetes" with n-gram = 3 will be segmented into (dia, iab, abe, bet, ete, tes). This approach is highly suitable for informal Indonesian texts, which are rich in morphology [17]. The vector representations generated by FastText were used as input for the SVM and CNN models, while the IndoBERT model used its internal transformer-based word representations (IndoBERT tokenizer).

In addition, a common challenge in sentiment analysis is the imbalanced class distribution, for example when the number of positive-labeled data is much higher than that of negative or neutral labels. This class imbalance can cause the model to focus more on the majority class while ignoring the minority class [18]. To address this issue, the class weight technique was applied, in which higher weights are assigned to minority classes during the model training process [19]. Class weighting was chosen because it does not alter the original data distribution, thereby avoiding the risk of information loss as in undersampling or the risk of overfitting due to data duplication as in oversampling.

To achieve optimal performance in sentiment analysis, it is essential to select hyperparameters that match the characteristics of the data and labels. Design of Experiments (DoE) is a statistical approach that systematically designs experiments by treating hyperparameters as factors and levels. Using 2^k factorial designs and space-filling strategies, DoE produces hyperparameter combinations that balance model accuracy and computational efficiency [20].

E. Model Classification

We use several classifier models in this study. These classifier models are:

- SVM. SVM is a powerful classification algorithm that separates data into distinct classes by identifying the optimal hyperplane that maximizes the margin between them [21]. SVM is often used in sentiment analysis because of

its ability to produce sharp classification boundaries, even on high-dimensional data. The DoE results in optimizing the model are C: 0.1, gamma: auto, and kernel: rbf for the manual labeling, VADER, VADER translate, and SentIL approaches. Meanwhile, for the Inset Lexicon and its modifications, the parameters are C: 10, gamma: scale, and kernel: linear.

- CNN. CNN is one of the deep learning architectures originally designed for image processing, but it has been proven effective in text classification tasks such as sentiment analysis [1]. CNN extracts local patterns (n-grams) from word sequences represented as vectors such as word embeddings to determine sentiment. The CNN model was optimized with a batch size of 64, embedding dimension of 128, and an optimizer, while other parameters were tuned using DoE. The tuning results showed the best configurations for manual labeling (Filters: 64, Kernel size: 3, Dropout: 0.5, LR: 0.001), InSet Lexicon, InSet-modified, and VADER Translate (Filters: 128, Kernel size: 5, Dropout: 0.3, LR: 0.001), as well as VADER and SentIL (Filters: 64, Kernel size: 5, Dropout: 0.3, LR: 0.001)
- IndoBERT. IndoBERT is a language model based on the BERT transformer architecture, specifically designed for Indonesian text, with strengths in capturing local context and grammatical structures through bidirectional attention [22]. This study employed the DoE results as follows: manual labeling, InSet Lexicon, VADER, and translated VADER (Dropout: 0.3, Batch size: 32, LR: 5e-05, Weight decay: 0.05); modified InSet Lexicon (Dropout: 0.1, Batch size: 16, LR: 5e-05, Weight decay: 0.05); and SentIL (Dropout: 0.3, Batch size: 16, LR: 5e-05, Weight decay: 0.01)

F. Evaluation

There are four evaluation metrics that we use to see the best performance of various combinations of lexico and models, namely:

- Accuracy. Accuracy indicates the proportion of model predictions that match the actual labels from the entire dataset. Accuracy = (TP + TN) / (TP + FP + FN + TN)
- Precision. Precision reflects how accurately the model classifies a specific class, showing how many of the predicted instances for that class are truly relevant. Precision = (TP) / (TP + FP)
- Recall. Recall measures the ability of the model to correctly identify all instances belonging to a particular class. Recall = (TP) / (TP + FN)
- F1-score. F1-score represents the balance between precision and recall, and is especially useful when dealing with imbalanced class distributions. F1-Score = 2 * (Recall * Precision) / (Recall + Precision)

F1-score is chosen as the main metric to assess model performance because this metric combines precision and recall in a balanced way. So that it can provide a more comprehensive evaluation picture, especially on imbalanced datasets [23].

IV. Results and Discussion

Data awal yang diperoleh melalui proses scraping berjumlah 8,393. Data tersebut dilakukan pelabelan manual yang dilakukan disajikan pada Tabel IV.

TABLE IV. Two Annotators Labeling Results

Annotator 2	Annotator 1			
	Positive	**Neutral**	**Negative**	**Total**
Positive	6,343	3	0	6,346
Neutral	202	1,449	0	1,651
Negative	2	0	394	396
Total	6,547	1,452	394	8,393

The initial annotation results showed differences in classification between the annotators, particularly in the positive and neutral categories, which required discussions to reach an agreement on the final labels. After the discussion process, the final dataset consisted of 6,433 positive data points, 1,566 neutral data points, and 394 negative data points. Based on the results in Table IV, the obtained values were $P_0 = 0.9753$, $P_e = 0.6257$, and $K = 0.9338$, indicating a very high level of agreement between the two annotators, which is considered sufficient for building the model.

In addition, the data were processed through a series of predetermined preprocessing stages defined in the lexicon, including cleaning, case folding, normalization, stopword removal, tokenization, and stemming. After undergoing all preprocessing steps, the number of valid and usable data was reduced to 6,627. The preprocessing results are presented in Fig. 2.

Fig. 2. After Preprocessing

Fig. 2 clearly shows that irrelevant words, unrelated terms, and misspelled words have been removed. The word cloud display was also limited to 500 words, and the words *"cek," "kesehatan," and "gratis"* were excluded because we aimed to highlight words that truly represent the CKG program. The subsequent data labeling process showed that the positive class dominated the data distribution across all the lexicons used, as presented in Table V. This indicates that the majority of the public expressed positive responses toward the CKG program. These results also demonstrate that the choice of lexicon-based approach can influence the data distribution.

TABLE V. SENTIMENT DISTRIBUTION BY LEXICON APPROACH

Lexicon Approach	Positive	Neutral	Negative
Manual Labeling	4996	1249	382
InSet Lexicon	3401	2047	1151
InSet Lexicon-modification	3557	1762	1307
VADER	6234	259	133
Translated VADER	6049	288	289
SentIL	6088	317	221

On the other hand, the data were then split into training and testing sets with an 80:20 ratio, which were converted into matrix representations. Furthermore, the use of class weights for each label is presented in Fig. 3.

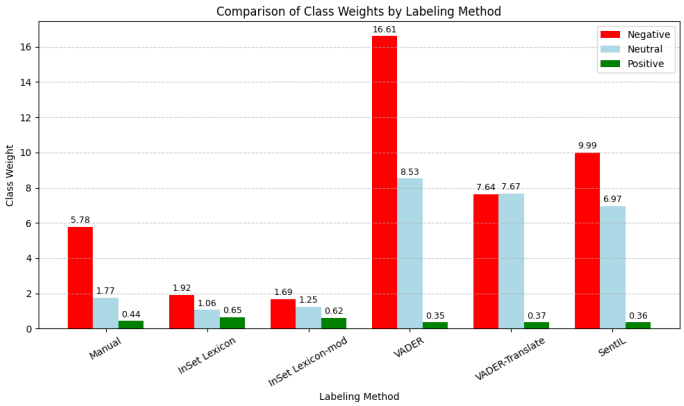

Fig. 3. Class Weight Values

TABLE VI. SVM PERFORMANCE

Lexicon Approach	Accuracy	Precision	Recall	F1-Score
Manual Labeling	75.41%	37.96%	59.27%	44.15%
InSet Lexicon	54.98%	53.69%	56.59%	**53.97%**
InSet Lexicon-modified	50.98%	50.49%	52.32%	49.34%
VADER	86.73%	37.25%	59.73%	38.87%
Translated VADER	84.77%	39.24%	54.49%	41.81%
SentIL	86.05%	41.98%	53.18%	44.12%

TABLE VII. CNN PERFORMANCE

Lexicon Approach	Accuracy	Precision	Recall	F1-Score
Manual Labeling	82.88%	67.91%	76.42%	71.30%
Inset Lexicon	89.14%	87.28%	88.31%	**87.70%**
Inset Lexicon-modified	87.56%	85.35%	85.46%	85.35%
VADER	97.21%	78.80%	78.97%	78.74%
Translated VADER	96.08%	80.51%	84.89%	82.50%
SentIL	96.68%	79.83%	84.37%	81.91%

TABLE VIII. INDOBERT PERFORMANCE

Lexicon Approach	Accuracy	Precision	Recall	F1-Score
Manual Labeling	85.14%	73.61%	85.90%	78.31%
Inset Lexicon	83.26%	80.22%	83.27%	81.57%
Inset Lexicon-modified	94.04%	93.05%	93.91%	**93.31%**
VADER	93.51%	64.03%	88.02%	71.99%
Translated VADER	90.12%	60.40%	83.73%	66.60%
SentIL	93.82%	63.52%	72.44%	65.02%

The final stage is evaluating all labels using the models presented in Tables VI, VII and VIII. Based on the comparison of F1-scores in Tables Tables VI, VII, and VIII, the Inset Lexicon-modified approach demonstrates competitive performance across all classification models. Although it did not achieve the highest scores in the SVM and CNN models 49.34% and 85.35% respectively—these results are still very close to the top scores achieved by Inset Lexicon (53.97% and 87.70%). Interestingly, when using the IndoBERT model, Inset Lexicon-modified outperformed all other approaches, achieving the highest F1-score of 93.31%. This finding indicates that the modifications made to the lexicon enhance the alignment between the data representation and transformer-based models like IndoBERT, resulting in more optimal classification performance. Overall, although it does not always achieve the highest scores in every model, Inset Lexicon-modified has proven to consistently remain competitive and demonstrates the best performance when combined with the IndoBERT model.

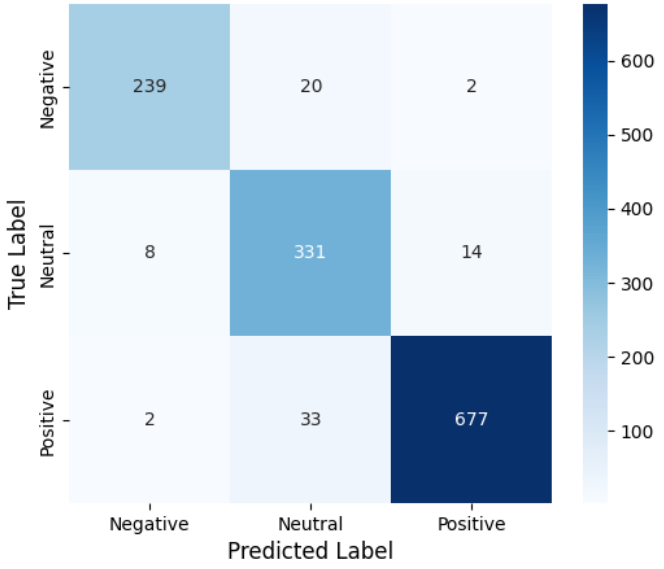

Fig. 4. Confusion Matrix InSet Lexicon-modified

The confusion matrix in Fig. 4 further reinforces this finding by showing that the model using the Modified Inset Lexicon is highly effective in classification, particularly for the positive class, which recorded the highest number of correct predictions (677 data points). Misclassifications were relatively minimal, such as positive data being incorrectly predicted as neutral (33 data points) or negative (2 data points), and neutral data being classified as positive (14 data points). This pattern suggests that the model is reliable in distinguishing polar sentiments (positive and negative), although some ambiguity remains within the neutral class. The distribution of predictions aligns with the high F1-score reported in Table VII, thereby strengthening the claim that the Modified Inset Lexicon is the most optimal strategy in this study.

The strength of this approach lies in the lexicon's design,

which is tailored to the context of the data, namely the CKG program. By focusing on health policy issues, the Modified Inset Lexicon is better equipped to capture the nuances of public sentiment compared to general-purpose lexicons such as VADER and its translated versions. This finding highlights the importance of employing context-specific lexical approaches, both linguistically and domain-wise, making it a valuable tool for analyzing public perceptions of policy programs in the future.

V. Conclusion

The results of the analysis indicate that the Inset Lexicon-modified approach demonstrates competitive performance across various classification models. Although it does not always achieve the highest scores in the SVM and CNN models, this approach performs closely to Inset Lexicon, which attained the best scores on those models. Furthermore, Inset Lexicon-modified achieved the highest performance when applied to the IndoBERT model, reaching an F1-score of 93.31%. These findings highlight that the modifications made to the lexicon improve the alignment between the training data and the characteristics of transformer-based models such as IndoBERT. Therefore, Inset Lexicon-modified can be considered the most promising lexicon-based approach for enhancing the accuracy and reliability of sentiment analysis on Indonesian language data.

VI. Acknowledgement

I would like to express my deepest gratitude to my supervisors, Purnomo Husnul Khotimah and Adria Arisal, for their invaluable guidance and continuous support throughout the course of this study. This study is also supported by Riset dan Inovasi untuk Indonesia Maju (RIIM) Batch 4 of the National Research and Innovation Agency (BRIN) and the Indonesia Endowment Fund for Education Agency (LPDP); Contract Number: B-3836/II.7.5/FR.06.00/11/2023.

References

[1] E. Nugraheni, P. H. Khotimah, A. Arisal, A. F. Rozie, D. Riswantini, and A. Purwarianti, "Classifying aggravation status of covid-19 event from short-text using cnn," in *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*. IEEE, 2020, pp. 240–245.

[2] W. N. S. W. Min, N. Z. Zulkarnain *et al.*, "Comparative evaluation of lexicons in performing sentiment analysis," *Journal of Advanced Computing Technology and Application (JACTA)*, vol. 2, no. 1, pp. 1–8, 2020.

[3] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.

[4] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for indonesian sentiment analysis in microblogs," in *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 2017, pp. 391–394.

[5] R. Wijayanti and A. Arisal, "Automatic indonesian sentiment lexicon curation with sentiment valence tuning for social media sentiment analysis," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 20, no. 1, pp. 1–16, 2021.

[6] R. Kurniawan and R. Rachmawati, "Indonesian twitter user sentiment towards pedulilindungi app in strengthening smart living during covid-19," *IKAT: The Indonesian Journal of Southeast Asian Studies*, vol. 6, no. 2, pp. 150–167.

[7] A. Girsang *et al.*, "Sentiment analysis of covid-19 public activity restriction (ppkm) impact using bert method," *arXiv preprint arXiv:2301.00096*, 2022.

[8] H. Firda, P. Putra, N. R. Oktadini, P. E. Sevtiyuni, and A. Meiriza, "Comparison of rating-based and inset lexicon-based labeling in sentiment analysis using svm (case study: Gobiz application reviews on google play store)," *Sistemasi: Jurnal Sistem Informasi*, vol. 14, no. 2, pp. 516–528, 2025.

[9] W. Van Atteveldt, M. A. Van der Velden, and M. Boukes, "The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms," *Communication Methods and Measures*, vol. 15, no. 2, pp. 121–140, 2021.

[10] M. Hadwan, M. Al-Sarem, F. Saeed, and M. A. Al-Hagery, "An improved sentiment classification approach for measuring user satisfaction toward governmental services' mobile apps using machine learning methods with feature engineering and smote technique," *Applied Sciences*, vol. 12, no. 11, p. 5547, 2022.

[11] P. H. Khotimah, A. Arisal, A. F. Rozie, E. Nugraheni, D. Riswantini, W. Suwarningsih, D. Munandar, and A. Purwarianti, "Monitoring indonesian online news for covid-19 event detection using deep learning." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 13, no. 1, 2023.

[12] L. N. Azizah, P. H. Khotimah, A. Arisal, A. F. Rozie, D. Munandar, D. Riswantini, E. Nugraheni, W. Suwarningsih, and D. Kurniasari, "The investigation into deep learning classifiers towards imbalanced text data," in *Proceedings of [Nama Konferensi, misalnya International Conference on AI and Data Science]*. Bandung, Indonesia: National Research and Innovation Agency (BRIN), 2025.

[13] S. Biswas, K. Young, and J. Griffith, "A comparison of automatic labelling approaches for sentiment analysis," *arXiv preprint arXiv:2211.02976*, 2022.

[14] G. A. D. Cahyo, P. H. Khotimah, A. F. Rozie, E. Nugraheni, A. Arisal, and A. Nuryaman, "Cnn-based hybrid performance evaluation towards online news sentiment classification task," in *2024 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*. IEEE, 2024, pp. 349–354.

[15] A. A. Chamid, R. Kusumaningrum *et al.*, "Labeling consistency test of multi-label data for aspect and sentiment classification using the cohen kappa method," *Ingenierie des Systemes d'Information*, vol. 29, no. 1, p. 161, 2024.

[16] A. Z. A. Muhabbab and R. A. F. Rizki, "Topic modelling berbasis embedding pada komentar youtube," in *Seminar Nasional Official Statistics*, vol. 2024, no. 1, 2024, pp. 873–884.

[17] R. Adipradana, B. P. Nayoga, R. Suryadi, and D. Suhartono, "Hoax analyzer for indonesian news using rnns with fasttext and glove embeddings," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2130–2136, 2021.

[18] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, "A survey of methods for addressing class imbalance in deep-learning based natural language processing," *arXiv preprint arXiv:2210.04675*, 2022.

[19] B. Bakirarar and A. H. Elhan, "Class weighting technique to deal with imbalanced class problem in machine learning: Methodological research," *Türkiye Klinikleri Biyoistatistik*, vol. 15, no. 1, pp. 19–29, 2023.

[20] C. Shi, A. K. Chiu, and H. Xu, "Evaluating designs for hyperparameter tuning in deep neural networks," *The New England Journal of Statistics in Data Science*, vol. 1, no. 3, pp. 334–341, 2023.

[21] GeeksforGeeks, "Support vector machine (svm) algorithm," 2025, accessed: September 17, 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/

[22] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of indonesian reviews using fine-tuning indobert and r-cnn," *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 348–354, 2022.

[23] S. Riyanto, S. S. Imas, T. Djatna, and T. D. Atikah, "Comparative analysis using various performance metrics in imbalanced data for multi-class text classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.