BERT-Based Multi-Task Classification Model for Free Health Check-Up Program Analysis

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

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

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

Andria Arisal
Research Center for Data and Information Science
National Research and Innovation Agency (BRIN)
Bandung, Indonesia
andr015@brin.go.id

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

2025 International Conference on Computer, Control, Informatics and its Applications (IC3INA) | 979-8-3315-9188-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/IC3INA68387.2025.11325397

***Abstract*—Assesing public perception is vital for ensuring the effectiveness government program. This study analyzes public responses to the government's Free Health Check-Up Program using data extracted from tweets on Platform X posted between October 20, 2024 and March 31, 2025. The study employs a multi-task classification approach in which the first task classifies tweets as either news or public service announcements, while the second task performs sentiment analysis. For comparison, single-task and multi-class classification model were also developed as baselines, allowing a comprehensive evaluation across different strategies. The dataset was preprocessed using a combination of the NLTK method, custom-built dictionary, and the InSet Lexicon-Based method for automatic sentiment labeling. The classification model used was IndoBERT, which is a BERT-based model and was optimized using the Design of Experiment technique. The multi-task classification approach demonstrated optimal performance. Achieving Content type classification reached an accuracy of 87.05% and F1-score of 84.04%, while sentiment classification achieved an even higher accuracy of 89.76% and F1-score of 87.07%. The model was trained with dropout rate of 0.2, learning rate of 5e-5, batch size of 32, weight decay of 0.1, and 5 training epochs. These result demonstrate the potential of multi-task classification and the IndoBERT model for analyzing public sentiment on social media platforms.**

***Index Terms*—free health check-up, sentiment analysis, multi-task classification, mutli-class classification, IndoBERT**

I. Introduction

Public perception plays a crucial role in evaluating government policies and the success of public programs in the digital era. Social media has transformed into a primary space where citizens express their opinions, shape perceptions, and mobilize support or criticism toward government initiatives [1], including the Free Health Check-Up Program. Therefore, regularly monitoring conversations on social media is essential for promptly addressing new problems and maintaining constructive engagement with the community.

However, data generated from social media platforms often exhibits highly complex characteristics both structured, unstructured and frequently contains informal language, abbreviations, symbols, and emoticons [2]. This complexity presents challenges in analysis and renders conventional approaches less effective, particularly in accurately identifying public sentiment and opinion. Therefore, effective data preprocessing techniques are essential to prepare such data for textual analysis [3]. In this context, the InSet Lexicon-Based approach by Fajri and Gemala has also been implemented to facilitate sentiment analysis through an Indonesian sentiment dictionary [4].

The application of BERT-based pre-trained models architecture and methods for Natural Language Processing (NLP) is becoming more relevant to tackle these problems [5]. BERT utilizes methods like Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) to comprehend the context of words in both directions within a sentence [6]. These capabilities enable BERT to perform a wide range of NLP tasks with greater accuracy [7] [6]. Built on the BERT architecture, IndoBERT is a language model created particularly for Indonesian [8]. The model was trained on various data sources, including Indonesian Wikipedia, news articles, and local web corpora [7]. Consequently, IndoBERT offers advantages in understanding the linguistic diversity and local context of Indonesian, particularly in the analysis of social media data [8].

Nevertheless, studies that jointly analyze content and sentiment using a multi-task classification approach in Indonesian-language social media particularly in the context of public health policy evaluation remain limited [9]. Most existing research focuses on a single classification task, such as topic classification [10] or sentiment analysis [11], and primarily adopts single-task or multi-label approaches [12]. In fact, social media data related to public policies like the Free Health Check-Up Program often contains dual-layered information both content and sentiment which can enrich interpretation

if analyzed simultaneously [9] [13].

In this study, We have employed a multi-task classification framework with IndoBERT to simultaneously analyze content and sentiment. Alongside this, we also evaluate single-task classification and a multi-class combined label approach. Both serve as baselines for comparison against the proposed multi-task framework. To optimize model performance We apply Design of Experiment (DoE) through hyperparameter tuning [14], while addressing class imbalance by assigning class weights [15]. This paper is organized as follows: Section II provides an overview of related research, Section III details the framework of our proposed method, and Section IV presents and discusses the experimental results. Section V concludes the paper.

II. Related Work

Public sentiment regarding government policies, services, and public figures is often examined through sentiment analysis, particularly on social media platforms like X, YouTube, and TikTok [16]. Study [17], through a systematic literature review, identified that sentiment analysis in Indonesian-language studies commonly focuses on topics such as entertainment, economy, politics, and public services. However, most of these studies rely on single-task approaches that separate content classification and sentiment analysis into distinct processes [18]. This separation introduces inefficiencies, as two different models or analysis pipelines must be developed and executed independently. It often leads to the loss of contextual information that could otherwise enhance the analysis [19]. Multi-task classification [20] has been introduced to address this limitation by enabling content and sentiment to be analyzed simultaneously in a more integrated and efficient manner.

A study similar to this research is conducted by [10], where a multi-task learning approach was applied to classify news categories and sentiment in order to support systems for suggesting content and finding information, with the purpose of delivering content that is more customized and relevant to each user. Another comparable study is the All-in-One framework [13], which employed a multi-task approach to simultaneously address four types of problems related to emotion and sentiment analysis, demonstrating superior performance compared to handling each task separately.

IndoBERT has been evaluated across various Indonesian NLP benchmarks and has demonstrated superior performance on most tasks [5] [21]. Comparative evaluation studies show that IndoBERT consistently outperforms baseline models such as multilingual BERT (mBERT), delivering significantly higher results [8]. For semantic level tasks, IndoBERT also outperforms other models in both sentiment analysis and extractive summarization [22]. Despite IndoBERT's proven effectiveness for a range of single-task NLP problems, research on incorporating IndoBERT into multi-task classification frameworks remains limited. As a result, the potential correlations between tasks within Indonesian-language texts have not been systematically explored.

III. Methodology

The experiment was carried out utilizing the framework displayed in Figure 1. The accumulated textual data underwent preprocessing and labelling based on its contextual relevance. To avoid overfitting, imbalanced data handling methods were used on the evaluation results. Multi-task classification approach was used for a IndoBERT model to analyze how the public views the Free Health Check-Up program.

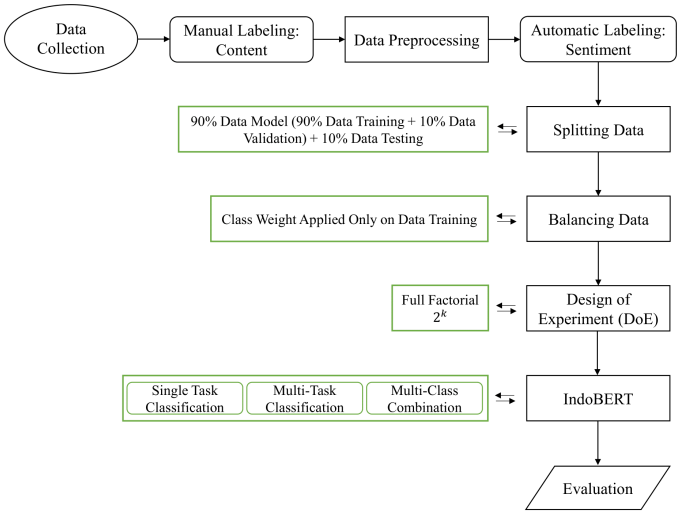

Fig. 1. Experiment Framework

A. Dataset

Text data related to the Free Health Check-Up Program was collected from the X platform using the keyword "cek kesehatan gratis". The dataset, with 8,393 entries gathered between October 20, 2024 and March 31, 2025, was preprocessed through cleaning, case folding, normalization, stopwords removal, tokenizing, and stemming. These methods were enhanced by a specialized dictionary [23], combined with the Natural Language Toolkit (NLTK). After preprocessing, the final dataset consisted of 6,635 entries. It was then visualized in a wordcloud Fig. 2, excluding "*cek*," "*sehat*," and "*gratis*" to highlight other relevant terms. The dataset was partitioned with 90% allocated for model development and 10% reserved for testing, where the development set was further split into training (90%) and validation (10%). The detailed distribution of content and sentiment labels is presented in Table I.

B. Labelling

Labeling was carried out for two categories: manual labeling for content type and automatic labeling for sentiment. The annotation process involved two annotators, with inter-annotator reliability assessed using Cohen's Kappa to ensure consistency [24]. We categorized the types of content representing the Free Health Check-Up Program into two classes. Inter-annotator agreement for this task achieved a Cohen's Kappa score of 0.89, indicating an almost perfect agreement. Disagreements

TABLE I. SUMMARY OF DATA DISTRIBUTION

Task		Data Distribution	Train Distribution	Val Distribution	Test Distribution	Class Weight	Average Sentence
Content	Conversation	4,997	4,046	452	499	0.6639	8.3178
	News	1,638	1,327	146	165	2.0253	11.6832
Sentiment	Negative	1,127	912	102	113	1.9624	10.4605
	Neutral	2,086	1,689	188	209	1.0602	7.7780
	Positive	3,422	2,772	308	342	0.6463	9.2250

Fig. 2. Wordcloud Free Health Check-Up Program

Fig. 4. Neutral Distribution Sentiment Maping

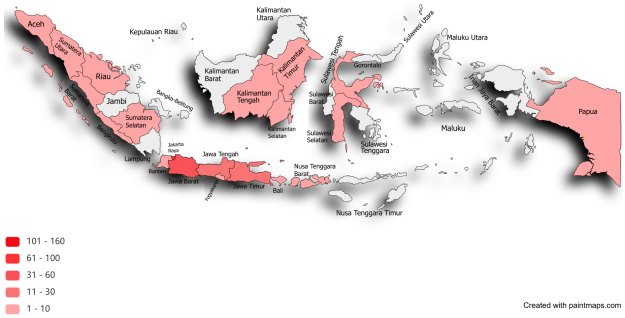

Fig. 3. Negative Distribution Sentiment Maping

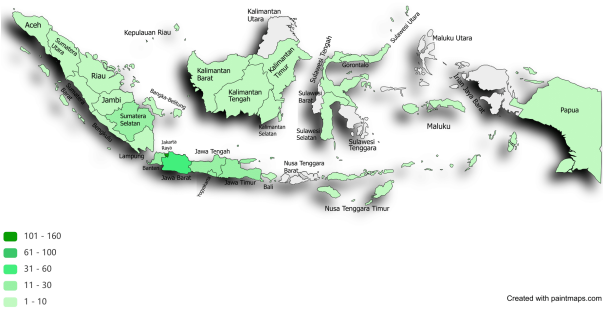

Fig. 5. Positive Distribution Sentiment Maping

were resolved through consensus, and the final agreed labels were then used as the ground truth:

- 0 (conversation) = Posts that encourage interaction help users build community and connections.
- 1 (news) = Posts sharing public service information and recognizable news content frequently marked by event hashtags are covered.

Sentiment labeling was performed automatically using the InSet lexicon-based method. To assess labeling reliability, a manual annotation was conducted on a randomly selected 10% subset of the data. The comparison yielded 628 matches and 36 mismatches, with a Cohen's Kappa score of 0.91, indicating almost perfect agreement between annotators. This demonstrates strong reliability of the labeling process, and the subsequent accuracy of 94.5% further highlights the alignment between the manual labels and the InSet lexicon-based method. The final classes are defined as follows:

- 0 (negative) = This category includes public opinions that express doubt, hate speech, or dissatisfaction with the program.
- 1 (neutral) = Public opinions that do not express an explicitly positive or negative attitude.
- 2 (positive) = This category includes public perspectives that are optimistic and enthusiastic about the program.

The sentiment distribution was visualized using a geographical map of Indonesia. As shown in Fig. 3, Fig. 4, and Fig. 5. Only tweets with available user location metadata were included in this visualization. Since location information is often incomplete or inconsistent on Platform X, these maps are not intended as a normalized representation of sentiment by population, but rather as a real distribution of geotagged tweets.

C. Design of Experiment (DoE)

Design of Experiments (DoE) uses a methodical, mathematical process to plan, perform, and analyze experiments in order to learn how input and output variables relate to one another. [14]. DoE is a methodology used for optimizing processes, making data-driven decisions, and modeling complex systems in a manner that saves time and resources. The DoE process includes the selection of important factors, the execution of

controlled experiments, the analysis of data, and the validation of results. These steps are taken to ensure that experimental objectives are met in an objective and measurable way. Various DoE methods, such as the full factorial design [25], have been implemented in materials science and engineering. The full factorial design is an experimental method where all possible combinations of factor levels are tested. For k factors, a full factorial design requires 2^k experiments.

D. Classification

Within the field of NLP, classification refers to the task of assigning labels to text based on predefined categories [10]. Several strategies can be applied to this problem, including single-task, multi-class, and multi-task classification. Single-task treats each problem independently, which limits the ability to exploit shared information [12]. Multi-class combines multiple labels into a single label space, but this can increase complexity and reduce flexibility [9]. In contrast, multi-task classification enables related tasks such as content and sentiment analysis to be learned jointly, sharing representations that improve generalization and reduce overfitting while being computationally efficient [11]. For these reasons, multi-task learning is the central focus of this study, while single-task and multi-class approaches are included as baselines for comparison.

E. IndoBERT

Evaluated across multiple Indonesian NLP benchmarks, IndoBERT which is based on the BERT architecture has demonstrated superior performance in the majority of tasks, particularly impressive given its relatively smaller model size [6]. IndoBERT's flexible architecture and strong contextual representation capabilities make it a suitable choice for this study. These characteristics are crucial when a single model needs to perform multiple classification tasks concurrently, as the model can extract general information applicable to different tasks [7] [8]. The pre-trained `indobenchmark/indorbert-base-p1` model was adopted as a shared encoder, with two task-epsific heads implemented as fully connected layers with softmax. The model applies a hard parameter sharing strategy, where both task jointly update the encoder. Training used cross-entropy losses for content and sentiment weighted 0.5 each, optimized with `AdamW` and backpropragated through the entire architecture, as illustrated in Fig. 6.

IV. Results and Discussion

Based on the prepared dataset and model architecture described in the methodology, the experimental stage was carried out to evaluate the proposed multi-task classification approach. The evaluation began with hyperparameter tuning to identify the most effective configuration. Since hyperparameters cannot be determined automatically, they were optimized through experimentation with different values. To obtain the best hyperparameter combinations (HC), we implemented a DoE approach, conducting 32 experiments. The top three configurations, shown in Table II, were selected based on validation performance and then applied during the final training and evaluation. The hyperparameters considered include:

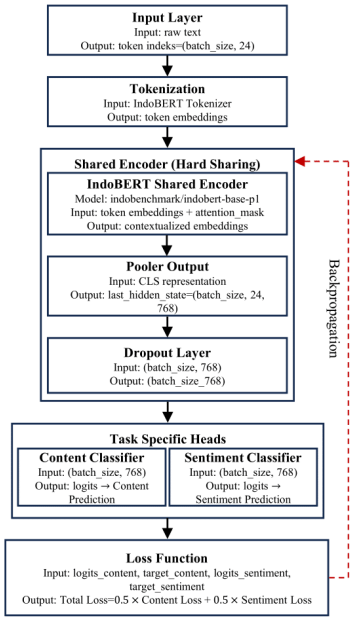

Fig. 6. IndoBERT Multi-Task Architecture

- Dropout is utilized as a stochastic regularization technique to reduce overfitting.
- Learning rate determines the extent to which the weight are updated.
- Batch size representing the number of samples processed before the weight are updated.
- Weight decay adds a penalty to large weights in the loss function.
- Epoch provide the model opportunities to learn better, especially when combined with an early stopping technique.

TABLE II. Design of Experiment Hyperparameter

Experiment	Hyperparameter	Value
HC 1	Dropout	0.3
	Learning Rate	5e-5
	Batch Size	32
	Weight Decay	0.1
	Epoch	10
HC 2	Dropout	0.2
	Learning Rate	5e-5
	Batch Size	32
	Weight Decay	0.1
	Epoch	5
HC 3	Dropout	0.2
	Learning Rate	5e-5
	Batch Size	16
	Weight Decay	0.01
	Epoch	5

TABLE III. PERFORMANCE COMPARISON ACROSS CLASSIFICATION APPROACHES

Classification	Experiment	Task	Accuracy	F1-Score
Single-Task Classification	HC 1	Content	0.8584	0.8163
		Sentiment	0.8916	0.8771
	HC 2	Content	0.8599	0.8098
		Sentiment	0.8705	0.8495
	HC 3	Content	0.8313	0.7942
		Sentiment	0.8720	0.8525
Multi-Class Combination Labels	HC 1	Content_Sentiment	0.7154	0.6702
	HC 2	Content_Sentiment	0.7455	0.6936
	HC 3	Content_Sentiment	0.7590	0.7166
Multi-Task Classification	HC 1	Content	0.8690	0.8270
		Sentiment	0.9051	0.8955
	HC 2	Content	**0.8810**	**0.8404**
		Sentiment	**0.8886**	**0.8707**
	HC 3	Content	0.8720	0.8297
		Sentiment	0.8961	0.8861

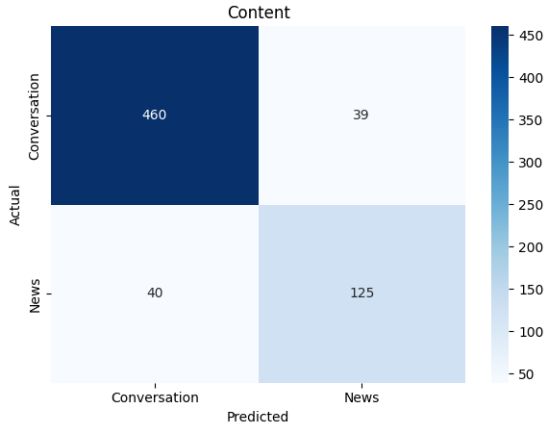

Fig. 7. Confusion Matrix Content

Fig. 8. Confusion Matrix Sentiment

We assess the effectiveness of each classifier using both accuracy and the F1-score to evaluate model performance. Accuracy indicates the general ability of the model to classify correctly across all classes, but it may not fully reflect performance when dealing with datasets exhibiting class imbalance. In these situations, focusing exclusively on accuracy can be deceptive because the model might excel in predicting common classes but overlook less frequent ones. Therefore, the F1-score becomes a crucial complementary metric, calculating the F1-score for each class independently and then averaging them, thus giving equal weight to each class regardless of its frequency in the dataset. This ensures the model's performance is fairly evaluated across all categories, including those that are underrepresented. By combining these two metrics, we obtain a more holistic and nuanced understanding of the model's strengths and limitations.

This study used the IndoBERT model optimized with a DoE approach as shown in Table III. The best overall performance was achieved by the multi-task framework under HC 2. This configuration produced an accuracy of 0.8810 and F1-score of 0.8404 for content classification. It also achieved an accuracy of 0.8886 and F1-score of 0.8707 for sentiment classification. The single-task approach reached competitive results with HC 1, where sentiment accuracy was slightly higher at 0.8916 but content performance was lower than the multi-task model. The multi-class combined label approach performed worst with its best result at HC 3, where accuracy was 0.7590 and F1-score was 0.7166. These findings highlight that the multi-task approach provides the most balanced and robust performance, offering consistent gains across both tasks while avoiding the drawbacks of single-task fragmentation and multi-class sparsity.

Effectiveness of the multi-task framework under HC 2 is demonstrated through the confusion matrix shown in Fig. 7 and Fig. 8. Content classification shows strong results and misclassifications are balanced across categories. This suggests that the model distinguishes well between the two types but overlapping linguistic patterns still cause confusion. For sentiment classification the model performs robustly across all classes correctly identified. Most errors occur in neutral detection tweets are misclassified, which indicates that tweets with moderate or less explicit emotional cues are harder to categorize. This shows the model's relative strength in capturing polarity. Overall the confusion matrices derived from

the evaluation set provide a clear view of precision across categories and highlight patterns of misclassification while emphasizing that the multi-task framework achieves balanced performance across classes.

V. Conclusion

This study presents a content and sentiment analysis of data from a Free Health Check-Up Program. It employed a multi-task classification approach based on the IndoBERT model and compared it with single-task and multi-class combined label approaches. The results showed that the multi-task framework performed optimally, with content classification reaching 88.10% accuracy and 84.04% F1-score, and sentiment classification achieving 88.86% accuracy and 87.07% F1-score under HC 2. While the single-task model obtained slightly higher sentiment accuracy at 89.16%, it performed worse on content classification and lacked the efficiency of a unified framework. The multi-class combined label approach produced the lowest results, with a maximum accuracy of only 75.90% and F1-score of 71.66%. These findings highlight the superiority of the multi-task framework in jointly handling dual classification tasks while avoiding the drawbacks of fragmented single-task models and imbalanced multi-class formulations.

VI. Acknowledgement

I would like to express my deepest gratitude to my supervisors, Purnomo Husnul Khotimah and Ekasari Nugraheni, 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] Y.-P. Yuan, Y. K. Dwivedi, G. W.-H. Tan, T.-H. Cham, K.-B. Ooi, E. C.-X. Aw, and W. Currie, "Government digital transformation: Understanding the role of government social media," *Government Information Quarterly*, vol. 40, no. 1, p. 101775, 2023.

[2] M. Saveski, B. Roy, and D. Roy, "The structure of toxic conversations on twitter," in *Proceedings of the Web Conference 2021 (WWW '21)*. Ljubljana, Slovenia: ACM, 2021, pp. 3294–3305.

[3] M. Hasan, T. Ahmed, M. R. Islam, and M. P. Uddin, "Leveraging textual information for social media news categorization and sentiment analysis," *PLOS ONE*, vol. 19, no. 7, p. e0307027, 2024.

[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)*. Singapore: IEEE, Dec 2017, pp. 391–394.

[5] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of indonesian tweets using large scale transfer learning via indobert," *Social Network Analysis and Mining*, vol. 15, no. 22, 2025, accepted: 24 December 2024.

[6] F. Bahruddin and M. F. Naufal, "Fine-tuning indobert for indonesian exam question classification based on bloom's taxonomy," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 225–233, 2023.

[7] F. Indriani, R. A. Nugroho, M. R. Faisal, and D. Kartini, "Comparative evaluation of indobert, indobertweet, and mbert for multilabel student feedback classification," *JRESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 6, pp. 748–757, 2024.

[8] A. F. A. Farizi and Y. Sibaroni, "Implementation of bilstm and indobert for sentiment analysis of tiktok reviews," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 1, pp. 96–106, 2025, available online 1 March 2025.

[9] W. O. Vihikan and I. N. P. Trisna, "Indonesian health question multi-class classification based on deep learning," *Journal of Information Systems and Informatics*, vol. 6, no. 3, September 2024.

[10] P. Shah, H. Patel, and P. Swaminarayan, "Multitask sentiment analysis and topic classification using bert," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 12, no. 1, July 2024.

[11] J. Zhang, K. Yan, and Y. Mo, "Multi-task learning for sentiment analysis with hard-sharing and task recognition mechanisms," *Information*, vol. 12, no. 5, p. 207, 2021.

[12] Riccosan and K. E. Saputra, "Multilabel multiclass sentiment and emotion dataset from indonesian mobile application review," *Data in Brief*, vol. 50, p. 109576, 2023.

[13] M. S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, and S. Kurohashi, "All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 285–299, 2022.

[14] 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, pp. 334–341, 2023.

[15] M. N. Razali, N. Arbaiy, P.-C. Lin, and S. Ismail, "Optimizing multiclass classification using convolutional neural networks with class weights and early stopping for imbalanced datasets," *Electronics*, vol. 14, no. 4, p. 705, 2025.

[16] G. P. R. Ariono and W. Alrasyid, "Ai-based sentiment analysis of social media to detect public opinion on government policies," *Journal Basic Science and Technology*, vol. 14, no. 2, pp. 61–68, June 2025.

[17] Y. Fauziah, B. Yuwono, and A. S. Aribowo, "Lexicon based sentiment analysis in indonesia languages: A systematic literature review," *RSF Conference Series: Engineering and Technology*, vol. 1, no. 1, pp. 397–405, November 2022.

[18] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, p. 102048, 2024.

[19] U.-J. Baek, B. Kim, J.-T. Park, J.-W. Choi, and M.-S. Kim, "A multi-task classification method for application traffic classification using task relationships," *Electronics*, vol. 12, no. 23, p. 3897, 2023.

[20] S. Huang, W. Peng, J. Li, and D. Lee, "Sentiment and topic analysis on social media: A multi-task multi-label classification approach," in *Proceedings of the 5th Annual ACM Web Science Conference (WebSci'13)*. Paris, France: Association for Computing Machinery, 2013, pp. 172–181.

[21] P. Khotimah, A. Arisal, A. 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 and Computer Engineering (IJECE)*, vol. 13, pp. 957–971, Feb. 2023.

[22] 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 (COLING)*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770.

[23] E. Nugraheni, F. I. Haekal, A. Arisal, and R. S. Perdana, "Optimizing indonesian tweet preprocessing on halal domain," in *2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2024, pp. 434–439.

[24] M. Li, Q. Gao, and T. Yu, "Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters," *BMC Cancer*, vol. 23, no. 799, 2023.

[25] G. Al-Kharusi, N. J. Dunne, S. Little, and T. J. Levingstone, "The role of machine learning and design of experiments in the advancement of biomaterial and tissue engineering research," *Bioengineering*, vol. 9, no. 10, p. 561, 2022.