



TRANSFORMER-BASED EMOTION CLASSIFICATION FOR INDONESIAN POLITICAL SPEECH

Daryanto*¹, Triawan Adi Cahyanto²

^{1,2}Department of Informatics Engineering, Universitas Muhammadiyah Jember, East Java, Indonesia.

¹<https://orcid.org/0000-0001-5132-5655>, ²<https://orcid.org/0000-0002-2893-5313>

Email: *daryanto@unmuhjember.ac.id, triawanac@unmuhjember.ac.id

ARTICLE INFO

Article History

Received: January 31, 2026

Reviewed: July 29, 2026

Accepted: August 29, 2026

Published: September 30, 2026

Keywords:

Emotion Analysis,
Indonesian NLP,
Political Speeches,
Roberta,
Transformer.

ABSTRACT

This study aims to classify emotions in Indonesian political speeches using the RoBERTa Transformer model. The dataset comprises 2,952 paragraphs derived from transcripts of closed-meeting political speeches delivered between 2014 and 2024. The research methodology encompasses data collection, augmentation, automatic emotion labeling, text pre-processing, and the fine-tuning of the RoBERTa model. The model was trained with a batch size of 16, a learning rate of 1e-5, and over three epochs. Evaluation was performed using a multi-class confusion matrix based on accuracy, precision, recall, and F1-score. Experimental results demonstrate that RoBERTa performs effectively, achieving 90% accuracy with a macro-average precision of 91%, recall of 90%, and an F1-score of 90%. Notably, the "happy" label achieved perfect scores across all metrics, while other labels exhibited consistent results. These findings validate the feasibility of Transformer-based approaches for recognizing implicit emotions in formal Indonesian political discourse. This research contributes to expanding Indonesian NLP resources within the political domain and provides empirical evidence for computational political discourse analysis. Future work may focus on expanding emotion categories, enhancing expert validation, and testing alternative Transformer architectures.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Advances in information technology over the past two decades have led to significant improvements in text-based data processing research. Text analysis is no longer confined to computational linguistics; it has become integral to social, economic, and political studies. This development has been reinforced by advances in the field of Natural Language Processing (NLP), a discipline that focuses on efforts to enable machines to understand and produce human language semantically and contextually [1]. One rapidly developing application in NLP is emotion classification, which is the process of recognizing emotional categories contained in text [2]. Emotion identification has been proven to support decision making, public opinion mapping, and political discourse analysis [3]. In the context of modern politics, political speeches are a strategic communication medium that aims to convey policy agendas, build images, and influence public perceptions and emotional responses [4]. Thus, emotion analysis in political speeches is important for understanding the communication strategies of political elites and for assessing their impact on public opinion. However, classifying emotions in formal texts such as political speeches poses its own challenges because emotional meaning is often implicit, contextual, and rhetorical. Conventional lexicon-based approaches often fail in capturing these semantic nuances and are less adaptive to language variations [5]. The emergence of Transformer architecture has been a breakthrough in NLP due to its ability to process text representations more contextually through the *self-attention* mechanism [6]. Various Transformer-based models have been developed, one of which is RoBERTa (Robustly Optimized BERT Approach), which shows better performance than BERT in various language comprehension tasks, including text classification [7]. Although RoBERTa's performance has been widely proven, research on its application to Indonesian—especially for the task of emotion classification in the domain of formal political discourse—is still very limited. Most Indonesian NLP research still focuses on sentiment analysis based on social media, e-commerce, or online news [8], while studies that specifically target political speeches are still rare. This limitation indicates a research gap in Indonesian NLP studies, particularly for emotion recognition in formal political discourse.

In addition, there is no available dataset of presidential speeches for emotion classification tasks, so this domain does not yet have an adequate research benchmark. To address this gap, this study applies the RoBERTa model to classify emotions in a corpus of speeches by President Joko Widodo during his two terms in office. The contributions of this study include: (i) the development of a dataset of Indonesian political speeches obtained from video transcriptions; (ii) the application and evaluation of the RoBERTa model in the domain of formal political discourse; and (iii) the provision of a starting point for Indonesian NLP research in emotion analysis in political documents.

II. MATERIALS AND METHODS

This research was conducted in several main stages, namely data collection, automatic emotion labeling, data pre-processing, training and *fine-tuning* of the RoBERTa model, and model evaluation. Each stage is explained as follows.

II.1 DATA COLLECTION

The research dataset consists of transcripts of President Joko Widodo's speeches during the 2014 – 2024 period, taken from the YouTube channel of the Cabinet Secretariat. This study focuses on 273 closed meeting videos to ensure that there is only one main speaker in order to suit the purpose of classifying emotions in political speeches. The transcripts were divided into paragraphs by analyzing semantic coherence. Specifically, the Indonesian-SBERT-Large model [9] was used to generate sentence embeddings, and a similarity threshold was applied to identify significant semantic shifts, thereby determining paragraph boundaries. This process collected 1,645 paragraphs. The data showed an imbalance in label distribution, so minority class augmentation was performed using synthetic data generation techniques as described in [10]. After augmentation, the total dataset consisted of 2,952 paragraphs. The dataset was divided using a *train-validation-test* scheme of 80%–10%–10%.

II.2 AUTOMATIC EMOTION LABELING

Emotion labelling was performed using the Indo RoBERTa Emotion Classifier model through the text-classification pipeline from Hugging Face [11]. The model generates five emotion categories, namely *anger*, *fear*, *happy*, *love*, and *sadness*. The *automatic labeling* approach was chosen for its processing efficiency with thousands of samples, as has been applied in previous NLP studies [12]. To reduce potential bias arising from differences between the training domain and the formal political domain, the labelling results were manually rechecked (post-editing) by a linguist who performed manual validation on 60% of the dataset to ensure label accuracy.

II.3 PREPROCESSING

The preprocessing stage was carried out to ensure that the data could be processed optimally by the model. The procedures applied included:

1. *Cleaning text* to remove irrelevant characters, symbols, and excess *whitespace* using the spaCy library [13].
2. *Tokenization* to map sentences into token units, in accordance with common practice in NLP [14].

Emotion labels are then converted to binary representations through one-hot encoding to support the multi-class classification stage.

II.4 MODEL TRAINING AND FINE-TUNING

The training stage was conducted using the RoBERTa base model, which comprises 12 hidden layers, 12 multi-head attention layers, and a hidden size dimension of 768 [6]. The model was implemented utilizing the Transformers library from Hugging Face [15], with PyTorch serving as the underlying deep learning framework [16]. For optimization, the AdamW algorithm [17] was employed alongside a warmup-based learning rate scheduler. The fine-tuning process was specifically configured to run for 3 epochs, using a batch size of 16 and a learning rate of $1e-5$. This approach, proven effective in various downstream NLP tasks [18], follows the comprehensive pipeline detailed in Algorithm 1.

Algorithm 1: RoBERTa Fine-Tuning Pipeline for Emotion Classification.

Input:

- Dataset D (Transcripts)
- Learning Rate $\eta = 1e-5$,
- Batch Size $B = 16$
- Epochs $E = 3$
- Model $M = \text{RoBERTa-base}$

Output: Trained Model M_{trained}

```

1: // Preprocessing Phase
2: for each paragraph  $p$  in  $D$  do
3:   Clean text  $p$  using spaCy (remove symbols, whitespace)
4:   Tokenize  $p$  into input_ids and attention_mask
5:   Encode label  $l$  into one-hot vector  $y$ 
6: end for
7: // Dataset Splitting
8: Split  $D$  into Train_set (80%), Val_set (10%), Test_set (10%)
9: // Model Initialization
10: Initialize  $M$  with 12 hidden layers, hidden size 768

```

```

11: Initialize Optimizer AdamW with warmup scheduler
12: // Training Loop
13: for epoch = 1 to E do
14:   for batch in Train_set do
15:     outputs ← M(batch.input_ids, batch.attention_mask)
16:     loss ← CrossEntropyLoss(outputs.logits, batch.labels)
17:     loss.backward() // Backpropagation
18:     optimizer.step() // Update weights
19:     optimizer.zero_grad()
20:   end for
21:
22:   // Validation
23:   Calculate accuracy and loss on Val_set
24:   if Val_loss does not improve then
25:     Early stopping check (optional)
26:   end if
27: end for
28: Return M_trained

```

II.5 EVALUATION

Model evaluation was performed using a confusion matrix for multi-class classification and four standard metrics, namely accuracy, precision, recall, and F1-score. The confusion matrix was chosen because it is a commonly used evaluation method in multi-class NLP classification tasks [19], [20].

III. RESULTS AND DISCUSSIONS

III.1 EVALUATION

This section presents the results of each stage of the research, starting from data collection, data preprocessing, training and fine-tuning of the RoBERTa model, to model evaluation and the results of emotion classification analysis. The following is a complete explanation.

III.2 DATA COLLECTION RESULTS

The final dataset comprises 2,952 text paragraphs obtained after data balancing. The data is sourced from transcripts of President Joko Widodo's political speeches in closed meetings during the 2014-2024 period. The dataset was then divided into three parts: 80% for *training*, 10% for *validation*, and 10% for *testing*. The distribution of data per emotion label is presented in Table 1 below.

Table 1: Data Distribution per emotion label.

Label	Training	Validation	Testing
Anger	477	60	60
Fear	400	50	50
Happy	402	50	50
Love	496	62	62
Sadness	586	73	74

Source : Authors, (2026).

The final dataset consists of five emotion categories, with the following distribution: *sadness* is the most common label with 733 paragraphs (24.8%), while *fear* is the least common with 500 paragraphs (16.9%). Other labels include: *happy* (17%), *anger* (20.25%), and *love* (21%). This distribution is visualized in Figure 1, which shows the relative balance between emotion labels.

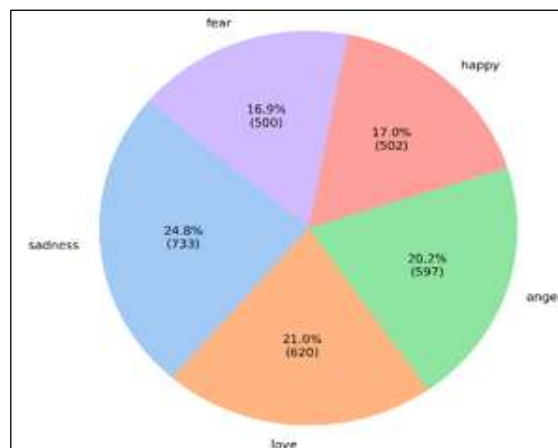


Figure 1: Data Distribution per Emotion Label.

Source: Authors, (2026).

III.3 DATA PREPROCESSING RESULTS

Before initiating the training phase, the text data underwent preprocessing to ensure optimal model input. The procedures included cleaning the text to remove irrelevant characters and symbols using the spaCy library [13], followed by tokenization to map sentences into token units. An example of the preprocessing outcome is presented in Table 2.

Table 2: Preprocessing and emotion labels.

No	Original paragraph	Pre-processing text
1	Assalamualaikum warahmatullahi wabarakatuh Rapat terbatas pada siang hari ini akan kita bahas lagi mengenai LRT di Palembang di Bandung dan di Jabodetabek Ini adalah sebuah rapat yang ke6 Artinya perlu harus ratas dan harus saya putuskan	assalamualaikum warahmatullahi wabarakatuh rapat terbatas pada siang hari ini akan kita bahas lagi mengenai lrt di palembang di bandung dan di jabodetabek ini adalah sebuah rapat yang ke6 artinya perlu harus ratas dan harus saya putuskan

Source : Authors, (2026).

After preprocessing is complete, the emotion labels are converted into numerical representations using the one-hot-encoding technique, as shown in Table 3. One-hot-encoding is a numerical representation technique for categorical data, where each category is represented as a binary vector.

Table 3: Emotion labeling and one hot encoding.

Emotion label	One hot encoding				
	Anger	Fear	Happy	Love	Sadness
<i>anger</i>	1	0	0	0	0
<i>fear</i>	0	1	0	0	0
<i>happy</i>	0	0	1	0	0

Source : Authors, (2026).

III.4 HYPERPARAMETER AND FINE-TUNING RESULTS

The training process using the RoBERTa model with the hyperparameter configuration is shown in Table 4.

Table 4: Hyperparameter roberta architecture.

Model	Batch Size	Dropout	Att Dropout	Epoch	Learning Rate
RoBERTa	16	0.2	0.2	3	1e-5

Source : Authors, (2026).

The selection of a batch size of 16 is based on training stability on a CPU device with limited capacity, while also being efficient for datasets with less than 10,000 data points. To regulate how quickly the model learns from each data point, the learning rate value in fine-tuning is optimally set in the range of 1e-5 to 5e-5 (Devlin et al., 2018). A value that is too high can cause the model to learn too quickly, while a value that is too low can make the learning process too slow and less effective. In this case, because the dataset used is relatively small, a learning rate of 1e-5 is used to prevent any indication of overfitting. Furthermore, Epoch 3 was used because with a dataset of 2952 and a batch size of 16, each epoch has 184 iterations, meaning that if 3 epochs are performed, the total is 552 iterations, which is sufficient to achieve convergence without losing model generalization. The advanced model tuning architecture is shown in Table 5.

Table 5: Fine-tuning roberta specification.

Model	Layer
RoBERTa	Input_ids = 256
	Attention_mask = 256
	Hidden Size = 768
	Output_layer = 5

Source : Authors, (2026).

III.5 PERFORMANCE RESULTS OF TRAINING HYPERPARAMETERS AND FINE-TUNING ROBERTA

During the training process, the model showed stable and convergent performance. Figure 2 shows a consistent decrease in the *loss* value in the training and validation data at each *epoch*. In the first *epoch*, there was a sharp decline, indicating an improvement in the model's ability to recognize patterns. Entering the second epoch, the rate of increase began to slow down but remained consistent. The model continued to improve, although not as quickly as in the first epoch. This pattern can be seen in the two lines that continue to decline with a balanced distance between them, which means that the process is stable with no indication of overfitting. In the third epoch, the validation line continued to decline, indicating that the model truly understood the basic patterns of the given data.

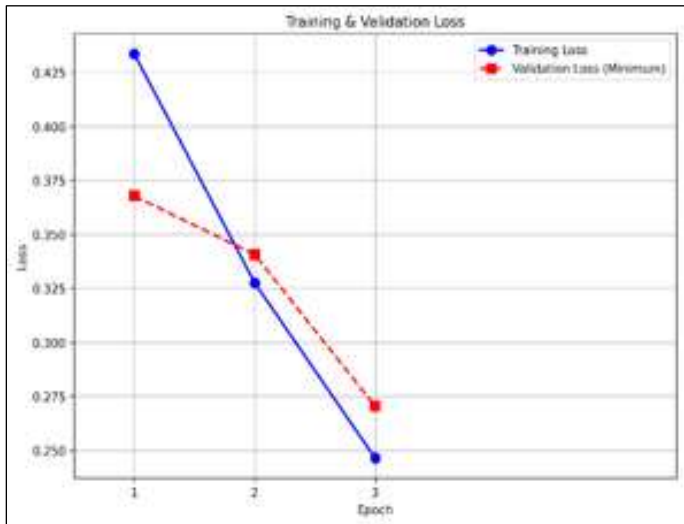


Figure 2: Training and validation loss.
Source : Authors, (2026).

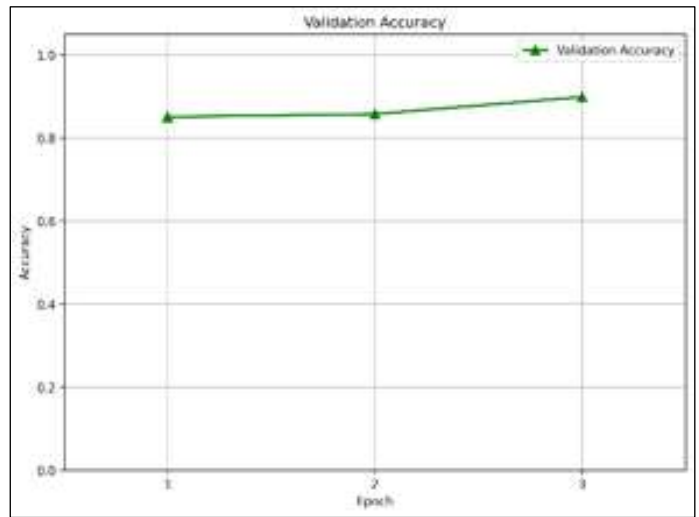


Figure 3: Validation accuracy.
Source : Authors, (2026).

On the other hand, Figure 3 shows the development of the *validation accuracy* value. In the first *epoch*, the model already achieved quite good results, indicating that from the beginning, the model was able to capture the basic patterns of the data. Entering the second *epoch*, the improvement was not too significant, showing that the model continued to learn and improve itself. Then, in the third epoch, there is a small increase in performance, indicating that the model has improved. This graph line shows stable and consistent progress. There are no signs of performance decline or imbalance, which usually occur when the model memorizes the data too much.

III.6 EMOTION CLASSIFICATION RESULTS AND MODEL PERFORMANCE EVALUATION

The emotion classification results on *RoBERTa* can be seen in Figure 4 (*confusion matrix*), with the following explanations:

- Anger**
 A total of 43 data points correctly showed the label *anger* (*True Positive*), but there were 17 data points that should have been classified as *anger* but were classified as other emotions, namely 3 *fear*, 5 *love*, and 9 *sadness* (*False Negative*). Meanwhile, 2 data from other labels were classified as *anger* (*False Positive*), namely 2 *love*, and 234 data that were not *anger* were predicted correctly (*True Negative*).
- Fear**
 A total of 43 data correctly showed the emotion of *fear* (*True Positive*), while 7 data that should have been *fear* were misclassified into other labels, namely 5 *love* and 2 *sadness* (*False Negative*). Six data points that were not *fear*, consisting of three *anger*, one *love*, and two *sadness*, were incorrectly predicted as *fear* (*False Positive*), and 240 data points that were not *fear* were correctly predicted as not *fear* (*TN*).
- Happy**
 A total of 50 data correctly showed the emotion of *happy* (*True Positive*), with no incorrect data that should have been classified as another emotion (*False Negative*). No data from other labels were incorrectly predicted as *happy* (*False Positive*), and 246 data that were not *happy* were successfully classified as not *happy* (*True Negative*).
- Love**
 A total of 58 data correctly showed the label *love* (*True Positive*), there were 4 data labels *love* that were classified as other emotions, namely 2 *anger*, 1 *fear*, 1 *sadness* (*False Negative*). There were 10 data points from other labels that were incorrectly predicted as the “love” label, namely 5 “*anger*” and 5 “*fear*” (*False Positive*), while 224 data points that were not the ‘love’ label were successfully classified correctly as not the “love” label (*True Negative*).
- Sadness**
 A total of 72 data points correctly showed the “sadness” label (*True Positive*), while 2 data points that were actually ‘sadness’ were incorrectly classified as other emotions, namely 2 “*fear*” (*False Negative*). There were 12 data points from other labels: 9 *anger*, 2 *fear*, 1 *love* that were misclassified as *sadness* (*False Positive*), and 210 data points that were not *sadness* were correctly classified as not *sadness* (*True Negative*).

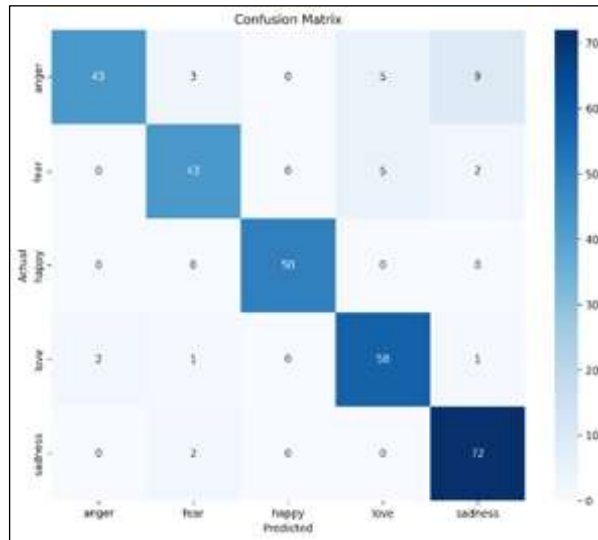


Figure 4: *Confusion matrix.*
Source : Authors, (2026).

Figure 5 presents a *classification report* that includes *precision*, *recall*, and *f1-score* for each label. The *happy* label achieved a perfect score in all metrics, namely *precision*, *recall*, and *f1-score* of 1.00, indicating that the model is very effective in recognizing the *happy* label. The “sadness” label also showed excellent performance with an F1-score of 0.91. The average macro values for precision, recall, and F1-score reached 0.91, 0.90, and 0.90, respectively, while the overall accuracy of the model was 90%. This shows that the model works optimally in recognizing emotions in Indonesian political speech texts.

	precision	recall	f1-score	support
anger	0.96	0.72	0.82	60
fear	0.88	0.86	0.87	50
happy	1.00	1.00	1.00	50
love	0.85	0.94	0.89	62
sadness	0.86	0.97	0.91	74
accuracy			0.90	296
macro avg	0.91	0.90	0.90	296
weighted avg	0.90	0.90	0.90	296

Figure 5: *Classification report.*
Source : Authors, (2026).

Table 6 shows various emotion pairs that often experience errors in the emotion recognition process by the model. The number of errors recorded comes from the analysis results shown in Figure 4 *confusion matrix*. The most common error occurs between the emotion labels *anger* and *sadness*, as shown in the following table.

Table 6: Emotional misalignment.

Emotion Pair	Error	Percentage	Main Cause
Anger↔Sadness	9 (FN)	15%	Model lacks sensitivity to assertive tones; absence of explicit keywords such as “marah” (angry)
Anger↔Love	5 (FN)	8,3%	- The model is less sensitive to sarcasm - Veiled criticism is often misclassified
Anger↔Fear	3 (FN)	5%	- The model is more sensitive to words that reflect threats or risks - It is closer to fear than to recognizing subtly implied anger
Fear↔Love	5 (FN)	10%	- The model tends to be mistaken because it is unable to recognize implicit emotions in diplomatic, technocratic, and solution-oriented language - Over-reliance on positive or negative keywords
Fear↔Sadness	2 (FN)	4%	The model has difficulty distinguishing between worry and sadness
Love↔Sadness	1 (FN)	1,6%	The model is not sufficiently trained on the emotion of <i>love</i> in formal language, so it fails to recognize <i>love</i> in institutional contexts such as political speeches.
Love↔Anger	2 (FN)	3,2%	Many emotion models, especially if not trained contextually, misinterpret assertive sentences as angry.
Sadness↔Fear	2 (FN)	2,7%	The model has difficulty recognizing the implicit context of sadness and looks for other alternatives such as <i>fear</i> because the sentences are constructed in a formal narrative style rather than an emotional one.

Source : Authors, (2026).

III.7 ERROR ANALYSIS BASED ON LINGUISTIC CHARACTERISTICS

Error analysis shows that the model has difficulty classifying emotions expressed through implicit linguistic structures. Examples of errors are shown in the following paragraphs:

“Rapat terbatas pada siang hari ini akan kita bahas lagi mengenai LRT di Palembang, di Bandung, dan di Jabodetabek. Ini adalah rapat yang ke-6. Artinya perlu harus ratas dan harus saya putuskan.”

The emotion actually points to the category of *anger* because there are indications of emotional pressure through the phrases “sixth meeting” and “I have to make a decision,” which indicate decisiveness and urgency. However, the model predicts the emotion of sadness. This indicates that the model associates the phrases “discussed again” and “sixth meeting” with a form of fatigue or despair that is closer to the emotion of sadness. This phenomenon illustrates the model's tendency to base predictions on explicit lexical cues rather than on pragmatic context and sentence structure.

III.8 IMPLICIT EMOTION AND FORMAL REGISTER

Prediction errors also appear in emotions that are conveyed implicitly. In the domain of political speeches, emotions are rarely conveyed through direct emotional vocabulary and appear more in the form of decisions, sarcasm, or discursive pressure. Formal, neutral, and bureaucracy-oriented language makes it difficult for models to recognize latent emotions. This is consistent with findings in NLP research which state that implicit emotions are more difficult to classify than explicit emotions because they depend on pragmatic and discourse context.

III.9 LONG-CONTEXT EFFECT

Long token analysis shows a correlation between prediction errors and context complexity. The average token length for correctly classified samples is 61.50 tokens, while incorrectly classified samples reach 66.89 tokens. This difference indicates that the longer the paragraph, the more likely the model is to lose emotional coherence. In addition, political texts tend to contain more than one emotion in a single paragraph, creating ambiguity for the model. This finding is in line with research reporting that Transformers still face challenges in maintaining long contexts and handling multi-emotion segments in a single discourse.

III.10 MODEL PERFORMANCE

The RoBERTa model performed well for the emotion classification task with 90% accuracy, 91% average precision, 90% recall, and 90% F1-score. The *happy* label achieved perfect scores on all metrics, while other labels also showed consistent scores. There were no indications of overfitting during training, as shown by the stability of the loss curve and the minimal performance gap between training and validation.

III.11 DISCUSSION

Overall, the results show that the Transformer approach, particularly RoBERTa, is effective in classifying emotions in formal Indonesian political speeches. This effectiveness is an important contribution given the limitations of previous models, which were dominated by lexicon-based approaches and conventional statistical methods in Indonesian language NLP research. The results of this study also reveal specific challenges in the domain of political speeches, namely: (i) implicitly conveyed emotions, (ii) the use of formal registers, and (iii) long contexts that may contain more than one emotion. These three factors are important considerations in the development of NLP models for the domain of political discourse in non-English languages.

IV. CONCLUSIONS

In conclusion, the RoBERTa model demonstrates high capability in classifying emotions in Indonesian political speech texts with high performance. This finding expands the application of Transformer in the field of Indonesian NLP and opens up opportunities for research on formal political discourse, which has rarely been studied. From an academic perspective, this study contributes to the study of emotion analysis, political discourse, and Transformer-based NLP. From a practical perspective, this approach has the potential to be used for mapping public opinion, analyzing political rhetoric, and formulating data-based policies. For further research, it is recommended to: (i) increase the number of expert annotators for inter-rater reliability analysis, (ii) expand the data domain to public debates, interviews, or policy speeches, (iii) add more diverse emotion categories, and (iv) explore other Transformer models such as IndoBERT, XLM-RoBERTa, or DistilBERT to obtain the optimal architecture configuration.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Daryanto, Triawan Adi Cahyanto.

Methodology: Daryanto.

Investigation: Triawan Adi Cahyanto.

Discussion of results: Daryanto, Triawan Adi Cahyanto.

Writing – Original Draft: Daryanto, Triawan Adi Cahyanto.

Writing – Review and Editing: Daryanto, Triawan Adi Cahyanto.

Resources: Daryanto, Triawan Adi Cahyanto.

Supervision: Daryanto, Triawan Adi Cahyanto.

Approval of the final text: Daryanto, Triawan Adi Cahyanto.

VI. REFERENCES

- [1] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," Sebastopol, California, Jun. 2009.
- [2] A. Seyeditabari, N. Tabari, and W. Zadrozny, "Emotion Detection in Text: a Review," Preprint ArXiv, 2018.
- [3] T. Brader, "Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions," *Am. J. Pol. Sci.*, vol. 49, no. 2, pp. 388–405, 2005, doi: 10.1111/j.0092-5853.2005.00130.x.
- [4] C. Mason, S. Hameau, and L. Nickels, "Emotion words in lexical decision and reading aloud," *Lang. Cogn. Neurosci.*, vol. 40, no. 4, pp. 527–546, Apr. 2025, doi: 10.1080/23273798.2024.2444582.
- [5] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, p. 15, 2017, doi: 10.48550/arXiv.1706.03762.
- [6] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, 2019, doi: 10.48550/arXiv.1907.11692.
- [7] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Naacl-Hlt* 2019, no. Mlm, pp. 4171–4186, 2018, doi: 10.48550/arXiv.1810.04805.
- [8] Putri Aisyiyah Rachma Dewi, Awang Dharmawan, Gilang Gusti Aji, Muhammad Danu Winata, and Jauhar Wahyuni, "Mapping Hoaxes, Disinformation, and Hate Speeches in Indonesia," *Technium Social Sciences Journal*, vol. 50, pp. 559–565, Nov. 2023, doi: 10.47577/tssj.v50i1.9943.
- [9] N. Ihsan, "indonesian-sbert-large." [Online]. Available: <https://huggingface.co/naufalihsan/indonesian-sbert-large>
- [10] N. Stylianou, D. Chatzakou, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "Domain-Aligned Data Augmentation for Low-Resource and Imbalanced Text Classification," 2023, pp. 172–187. doi: 10.1007/978-3-031-28238-6_12.
- [11] S. Limcorn, "Indo RoBERTa Emotion Classifier." [Online]. Available: <https://huggingface.co/StevenLimcorn/indonesian-roberta-base-emotion-classifier>
- [12] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 1–17. doi: 10.18653/v1/S18-1001.
- [13] H. Matthew, I. Montani, S. Van Landeghem, and B. Adriane, "spaCy Industrial-strength Natural Language Processing in Python," *spaCy*. [Online]. Available: <https://spacy.io/>
- [14] K. Park, J. Lee, S. Jang, and D. Jung, "An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 133–142. doi: 10.18653/v1/2020.aacl-main.17.
- [15] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.
- [16] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," 2019.
- [17] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization." [Online]. Available: <https://github.com/loshchil/AdamW-and-SGDW>
- [18] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," 2019, pp. 194–206. doi: 10.1007/978-3-030-32381-3_16.
- [19] P. Vickers, L. Barrault, M. A. France, E. Monti, and N. Aletras, "We Need to Talk About Classification Evaluation Metrics in NLP," *Long Papers*. [Online]. Available: <https://github.com/petervickers/>
- [20] I. Markoulidakis, G. Kopsiaftis, I. Rallis, and I. Georgoulas, "Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem," in *Proceedings of the 14th Pervasive Technologies Related to Assistive Environments Conference*, New York, NY, USA: ACM, Jun. 2021, pp. 412–419. doi: 10.1145/3453892.3461323.