JOURNAL OF INFORMATICS DEVELOPMENT

E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Comparative Analysis of Transformer Models for Emotion Classification in Indonesian Text Data

Faishal Basbeth[1], Cahyasari Kartika Murni[2], Abdur Ro'uf[3]

Department of Information Technology, Faculty of Engineering and Informatics, Institut Teknologi dan Bisnis Widya Gama Lumajang, Indonesia[1,2,3]

Corresponding Author: Faisal Basbeth (faishalbasbeth@gmail.com)

ARTICLE INFO

Date of entry:
10 January 2026

Revision Date:
28 August 2026

Date Received:
01 October 2026

ABSTRACT

This study compares three transformer-based models, BERT, RoBERTa, and DistilBERT, for emotion classification on Indonesian text. The Indo4B dataset, which consists of five emotion labels (anger, fear, happiness, love, and sadness), is used as the benchmark corpus. All three models were fine-tuned and hyperparameter-tuned. The evaluation was conducted using accuracy, F1-score, model size, and training time as the primary performance metrics. The results indicate that RoBERTa achieves the best overall performance, with an accuracy of 90.83% and an F1-score of 91. Meanwhile, DistilBERT demonstrates approximately 60% faster training time compared to BERT, with only a 0.61% decrease in accuracy. These findings suggest that DistilBERT offers a highly efficient alternative while maintaining competitive classification performance.

Keywords: Emotion Classification; Transformer; BERT; RoBERTa; DistilBERT

Cite this as: Basbeth, F. (2026). Comparative Analysis of Transformer Models for Emotion Classification in Indonesian Text Data. *Journal of Informatics Development*, *5*(1), 28–40. https://doi.org/10.30741/jid.v5i1.2082

INTRODUCTION

Sentiment analysis has become widely adopted in the field of artificial intelligence, as it helps analyze digital text by integrating techniques such as Natural Language Processing (NLP) and deep learning (Huan et al., 2022). Sentiment analysis serves multiple important purposes, including evaluating products, company services (Lin & Nuha, 2023), and government programs, all assessed through digital text. It identifies the emotional meaning of sentences classified as positive or negative (Medhat et al., 2014). Positive sentences may correspond to emotions such as happiness or love, while negative sentences may correspond to emotions such as anger or sadness. However, this binary outcome, positive or negative, still has limitations in capturing the full range of emotional meaning (Aman & Szpakowicz, 2007). In contrast, text emotion classification offers advantages by recognizing a broader and more diverse spectrum of emotional expressions (Acheampong et al., 2021). Developing text emotion classification requires sufficient data consisting of sentences and their corresponding labels. The sentences represent human expressions obtained in digital text form, while the labels indicate the specific emotional category assigned to each sentence. Beyond data

E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

availability, this task also relies on various NLP and deep learning models explored in previous studies.

Most datasets used for emotion classification are based on English text, and studies by Wilie et al. (2020) and Yanuar & Shiramatsu (2020) report that research progress in Indonesian text emotion classification remains limited. Given this gap, researchers are encouraged to develop more comprehensive and effective approaches to Indonesian text emotion classification. Previous studies on Indonesian text emotion classification have employed several strong transformer-based models, including BERT (Wilie et al., 2020), RoBERTa (Wiciaputra et al., 2021), and DistilBERT (Nissa & Yulianti, 2023). The study by Wilie et al. (2020) applied the BERT model to Indonesian emotion classification by fine-tuning it with a Masked Language Modeling (MLM) approach that predicts up to 20 masked tokens. The resulting model was named IndoBERT BASE. The same study also introduced an Indonesian emotion dataset called Indo4B, which consists of five emotion labels and is divided into 3,521 training samples, 440 validation samples, and 442 test samples. The classification task was performed using a single-label classification approach. The reported accuracy results were 73.88% for IndoBERT-lite BASE and 75.48% for IndoBERT BASE. For the RoBERTa model, the study by Wiciaputra et al. (2021) performed Indonesian text classification using XLM-RoBERTa without applying any hyperparameter tuning. Instead, the authors utilized five different text datasets. They combined them in several configurations, including: English news dataset with 37,886 English samples, Mixed news dataset (75,772 samples) consisting of 37,886 English and 37,886 Indonesian samples, Mixed news dataset (108,190 samples) consisting of 37,886 English and 70,304 Indonesian samples, English news subset with 2,000 samples, and Mixed news subset (4,000 samples) consisting of 2,000 English and 2,000 Indonesian samples. The best performance was achieved on the mixed news dataset with 75,772 samples, achieving an F1 Score of 93.30.

For the DistilBERT model, Nissa & Yulianti (2023) applied both the m-BERT-base and DistilBERT-base multilingual architectures to Indonesian text classification. The research used the *Airy Rooms Hotel Reviews* dataset, which contains user comments describing specific hotel rooms or areas. The dataset consists of 10 labels and is divided into 2,283 training samples, 286 validation samples, and 285 test samples. A multi-label classification approach was employed in this study. The results showed that m-BERT-base achieved 90.9% accuracy, while DistilBERT-base achieved 87.8%. Several previous studies on Indonesian emotion classification, as described above, only compared limited model groups namely (BERT and RoBERTa), XLM-RoBERTa, and (m-BERT-base and DistilBERT-base). According to the authors, this still leaves gaps in terms of comprehensive model comparison. Therefore, this study aims to analyze three transformer-based models: BERT, RoBERTa, and DistilBERT. BERT offers the advantage of learning contextual information bidirectionally, enabling the model to capture meaning based on both preceding and following words (Devlin et al., 2018). RoBERTa provides similar strengths but is trained on significantly more data approximately 144 GB allowing it to surpass BERT on various downstream tasks (Liu et al., 2019). DistilBERT, on the other hand, offers a smaller and more efficient architecture while maintaining 97% of BERT's performance (Sanh et al., 2020). This research aims to evaluate the performance of these three models (BERT, RoBERTa, and DistilBERT) in Indonesian emotion classification by measuring model size, training time, accuracy, and F1-score to determine the strengths and effectiveness of each architecture.

The three models were developed through fine-tuning and hyperparameter tuning. During the fine-tuning process, an additional output layer was added to each model to generate classification predictions, along with a loss function to measure the error between predicted and actual labels. Hyperparameter tuning was also applied to optimize model performance. The dataset used for fine-tuning was Indo4B. Before training, the dataset was balanced using undersampling, adjusting each emotion class to match the class with the fewest samples. This step was necessary because data imbalance can negatively affect model performance, especially in multi-class emotion classification tasks. Based on the above explanation, this study aims to conduct a comparative performance

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

analysis of three transformer-based models, BERT, RoBERTa, and DistilBERT, for Indonesian text emotion classification. The experiment was carried out using fine-tuning and hyperparameter tuning on the Indo4B dataset, measuring and comparing model size, training time, accuracy, and F1-score to evaluate each model's effectiveness at understanding emotional expressions in Indonesian text. The findings of this study are expected to advance Indonesian-language NLP models, particularly in digital emotion classification tasks. Furthermore, the results may serve as a reference and foundation for future research in artificial intelligence, emotion analysis, and emotion-based recommendation systems.

THEORETICAL FRAMEWORK AND HYPOTHESES

Several previous studies on Indonesian emotion classification, as described above, only compared limited model groups, namely (BERT and RoBERTa), XLM-RoBERTa, and (m-BERT-base and DistilBERT-base). According to the authors, this still leaves gaps in terms of comprehensive model comparison. Therefore, this study aims to analyze three transformer-based models: BERT, RoBERTa, and DistilBERT. BERT offers the advantage of learning contextual information bidirectionally, enabling the model to capture meaning based on both preceding and following words (Devlin et al., 2018). RoBERTa provides similar strengths but is trained on significantly larger amounts of data, approximately 144 GB, allowing it to surpass BERT on various downstream tasks (Liu et al., 2019). DistilBERT, on the other hand, offers a smaller and more efficient architecture while maintaining 97% of BERT's performance (Sanh et al., 2020). This research aims to evaluate the performance of these three models (BERT, RoBERTa, and DistilBERT) in Indonesian emotion classification by measuring model size, training time, accuracy, and F1-score to determine the strengths and effectiveness of each architecture.

METHODS

This study employs a quantitative research approach, using textual data from Twitter posts as the primary research object. These tweets serve as the main source for analyzing user emotions, processed using the BERT, RoBERTa, and DistilBERT transformer models. The prediction results from each model are then evaluated using several evaluation metrics to obtain an optimal performance comparison. The research workflow is illustrated in Figure 1.

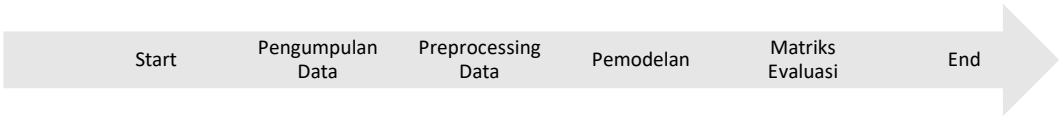

Figure 1. Research Workflow is Illustrated

Data Collection: This study uses a single text dataset, namely Indo4B, which was developed in the study by Wilie et al. (2020). The Indo4B dataset was collected from Twitter. Twitter is widely used as a real-time source of social information, including breaking news, celebrity posts, trending topics (Lee et al., 2011), expressions from daily life, and movie reviews (Li & Xu, 2014). Its richness and diversity make Twitter a valuable source of textual data. Users can post short text messages up to 140 characters, making it suitable for capturing concise emotional expressions. The Indo4B dataset contains five emotion labels: fear, anger, sadness, love, and happiness. These emotion categories are derived from Shaver's basic emotion taxonomy (Shaver, 1987). Most emotion-classification studies on textual data typically rely on lexicons, which represent the smallest meaningful units of language. However, emotion classification in the Indonesian language is generally not lexicon-based. While various emotion lexicons exist for English, only one Indonesian emotion lexicon has been

E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

developed, as presented in Shaver (1987). The studies conducted by Dong (2018) and Wilie et al. (2020) also adopted the same emotion labels proposed by Shaver's framework.

The Indo4B dataset was annotated by two linguistic experts (Dong, 2018) and consists of 3,521 training samples, 440 validation samples, and 442 test samples across five emotion labels. Unlike the dataset usage described in Wilie et al. (2020), this study employs the Indo4B dataset with an additional balancing process using undersampling, in which each class is adjusted to match the smallest class size. Undersampling is applied because class imbalance significantly affects model performance (Kaur et al., 2019); emotion classification models trained on imbalanced data tend to achieve higher accuracy on majority classes while performing poorly on minority classes. In this study, the dataset is split into three subsets: 80% for training, 10% for validation, and 10% for testing (Wilie et al., 2020).

Data Preprocessing: Data preprocessing is the process of transforming raw textual data into a more structured format that aligns with the requirements of data mining tasks (Huang et al., 2015). Figure 2 illustrates the sequence of preprocessing stages. The first stage is case folding, which converts all uppercase letters into lowercase. This step is crucial because the models used in this study are uncased models, which operate optimally on lowercase text. The second stage is filtering, which removes selected elements such as @username, retweets, hashtags, emoticons, and URLs, as these components carry minimal semantic relevance for emotion classification tasks. The third stage is tokenization, which splits textual data into smaller units called tokens, providing an important foundation for subsequent preprocessing procedures. The fourth stage is stopword removal, which eliminates non-essential words to retain only semantically meaningful terms. The fifth stage is stemming, which converts words with affixes into their root forms, helping preserve consistent lexical meaning across variations (Hidayatullah & Ma'Arif, 2017). Finally, the labels are one-hot encoded to prepare them for the classification models.

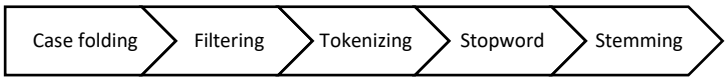

Figure 2. Sequence of Data Preprocessing Stages

Modeling: This study involves developing BERT, RoBERTa, and DistilBERT models through fine-tuning and hyperparameter tuning. Fine-tuning is a technique that uses a pre-trained model on a specific task and fine-tunes several of its final layers to solve a new task, thereby significantly improving the model's performance (Sun et al., 2019). In this study, the development of Indonesian text emotion classification using BERT, RoBERTa, and DistilBERT models employed pre-existing models that were processed through fine-tuning:

1. BERT Model: bert-base-uncased model, which uses 12 hidden layers, 12 multi-head attention, and a 768 hidden size. It supports a maximum token length of 512 and has a vocabulary size of 30522 (Devlin et al., 2018).
2. RoBERTa Model: The roberta-base model uses 12 hidden layers, 12 *multi-head attention heads*, and a 768-dimensional hidden size. It supports a maximum token length of 512 and has a vocabulary size of 50265 (Liu et al., 2019).
3. DistilBERT Model: distilbert-base-uncased, which uses 6 hidden layers, 12 *multi-head attention*, and a 768 hidden size. It supports a maximum token length of 512 and has a vocabulary size of 30522 (Sanh et al., 2020).

For the three models above, fine-tuning is performed as follows: adding an output layer and a loss function cross-entropy loss to BERT, RoBERTa, and DistilBERT. The output layer determines the classification prediction, using the softmax activation function. Cross-entropy loss is added to measure the difference between the model's predicted output and the expected output. An

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

illustration of the emotion classification process for Indonesian text using fine-tuning with BERT, RoBERTa, and DistilBERT is shown in Figure 3.

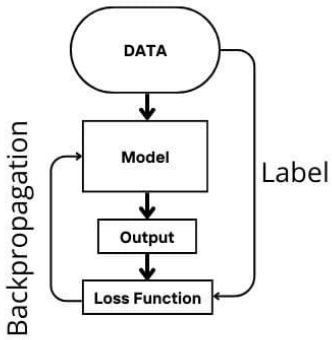

Figure 3. An illustration of emotion classification on text data using fine-tuning with BERT, RoBERTa, and DistilBERT

The optimal hyperparameters used in this study were determined based on gradient fluctuations, with batch sizes ranging from {16,…,320}, learning rates from {0.01,…,0.20}, and epochs from {10,…,25} (Hastomo et al., 2021). The optimizer used was the Adam optimizer with beta 1 (β_1 = 0.9) and beta 2 (β_2 = 0.999) (Sun et al., 2019), and adjustments were made accordingly (He et al., 2019).

Evaluation Matrix: The evaluation metric used in this study is the confusion matrix. A confusion matrix is a decision-measurement model commonly used in supervised learning and a visual evaluation tool often applied in deep learning. The confusion matrix visualizes the degree of misclassification made by a deep learning model across different classes. It contains the predicted and actual class results, capturing all classification outcomes for each label (Xu et al., 2020). The confusion matrix used in this study is multiclass. A multiclass confusion matrix is suitable for this research because the dataset contains more than two classes. A multiclass confusion matrix consists of four different combinations of predicted and actual values: true positive, true negative, false positive, and false negative (Rahmad et al., 2020). These four combinations are then used to compute accuracy, precision, recall, and F1-score (Markoulidakis et al., 2021).

RESULTS AND DISCUSSION

Based on the research procedures conducted, training and testing results were obtained, describing the performance of each architecture. The analysis focused on accuracy, F1-score, model size, and training time to assess the effectiveness of the three transformer models (BERT, RoBERTa, and DistilBERT) for Indonesian text emotion classification.

Data Preprocessing and Results: In this study, the model was developed using 3,401 text samples across five labels, divided into three subsets: 80% for training, 10% for validation, and 10% for testing. The 3,401 text samples were balanced using an undersampling technique to ensure an equal number of data points for each label. The data balance is shown in Figure 4. Examples and the distribution of data per label can be seen in Table 1.

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

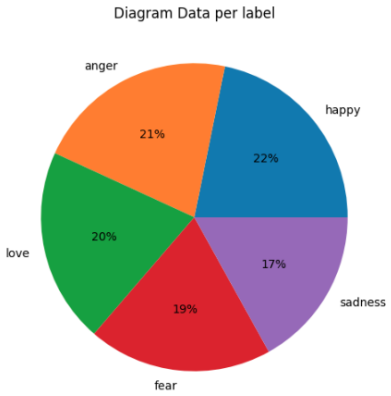

Figure 4. Data Balance per Label

Table 1. Examples and Data Distribution per Label

Label	Training & Validation	Test	Total
Fear	597	69	666
Anger	657	64	721
Sadness	520	69	589
Love	630	63	693
Happy	670	62	732

After processing the data and creating a balanced dataset, preprocessing was then performed. The results of the preprocessing steps are presented in Table 2. The preprocessing procedures include:

a. **Case folding, where all letters in the sentences have been converted to lowercase.**
b. **Filtering, where elements such as @username, retweets, hashtags, emoticons, and URLs have been removed.**
c. **Stopword removal, where additional affixed words have been reduced, such as *dipancang* → *pancang*, *tergesa* → *gesa*, and *ditanam* → *tanam*.**
d. **Stemming, where inflected or affixed words have been reduced to their root forms, such as *pangkalmu* → *pangkal* and *kutuki* → *kutuk*.**

The final preprocessing stage converts the labels into a one-hot encoded format, as shown in Table 3.

Table 2. Data Preprocessing Results

Text	Preprocessing Text
Kalo kamuuuuuu, kamuu itu specialll. Cintanya juga ga kalah kok sama yang lain. Karena sebenarnya bagi aku kalian itu satu. Jadi, cinta satu cinta semua. [URL]	kamu bahagia cinta kalah sebenernya cinta cinta
temen2 masa kecil yang turned out being asshole itu kalo cewe jadi ukhti2 hijrah kalo cowo jadi abdi negara, rasanya semua statement ku di sosmed pasti yang menyangkal orang2 kek gitu, parahnya lagi kalo mereka pacarana	teman makhluk asshole cewek ukhti hijrah pria abdi negara nyata Bahagi sangkal orang parah hubung
Ini adalah hal yang paling membahagiakan saat biasku foto Bahagia ELF #ReturnOfTheLittlePrince #HappyHeeChulDay	Bahagia foto

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Table 3. One-hot encoding format for label

Label	One-hot encoding
Anger [0]	[1.,0.,0.,0.,0.,]
Fear [1]	[0.,1.,0.,0.,0.,]
Happy [2]	[0.,0.,1.,0.,0.,]
Love [3]	[0.,0.,0.,1.,0.,]
Sadness [4]	[0.,0.,0.,0.,1.,]

Model Training Results: In the subsection on model training results, we present the architectures of each model developed through hyperparameter tuning and fine-tuning, along with a comparison of parameters/sizes and training times for BERT, RoBERTa, and DistilBERT. Table 4 shows the architectures used for hyperparameter tuning and fine-tuning, along with the total number of parameters.

Table 4. Hyperparameter tuning, fine-tuning, and total parameter architecture

No	Model	Layer	Parameter	Batch Size	Drop out	Att Dropout	Epoch	Learning Rate
1.	BERT	Input_ids = 256 Attention_mask = 256 4 Bert = 768 Output_layer = 5 T	0 0 109482240 3845 109486085	16	0.1	0.1	10	1e-5
2.	RoBERTa	Input_ids = 256 Attention_mask = 256 4 Roberta = 768 Output_layer = 5 T	0 0 124645632 3845 124649477	16	0.1	0.1	10	1e-5
3.	DistilBERT	Input_ids = 256 Attention_mask = 256 4 Distilbert = 768 Output_layer = 5 T	0 0 66362880 3845 66366725	16	0.1	0.1	10	1e-5

The layer column in Table 4 presents the structural components that have been constructed, as illustrated in Figure 3:

1. Input_ids is the input layer containing text data that has been converted into word embeddings with a sequence length of 256.
2. Attention_mask is the input layer containing the text data that has also been converted into word embeddings and padded with the value 1, with a sequence length of 256.
3. BERT, RoBERTa, and DistilBERT are layers that receive input via *Input_ids* and *Attention_mask* and process it using their respective architectures, with a hidden size of 768.
4. Output_layer is the layer that receives the model's processed output and applies a neural network to produce the final classification across five labels.

The parameter column in Table 4 produces the total parameter/size for each model:

1. The BERT model has a parameter count/size of 109486085.
2. The RoBERTa Model has a parameter count/size of 124649477.
3. The DistilBERT Model has a parameter count/size of 66366725.

The parameter size results show that the DistilBERT model achieves a 40% reduction (43.119.360 parameters) compared to the BERT model. Meanwhile, the RoBERTa model is shown to have more parameters than BERT, with an increase of 15.163.392. The models that underwent hyperparameter tuning and fine-tuning were then trained using the constructed data splits. From this training process, the training speed was measured. Table 5 presents the training time results for each model.

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Table 5. Training Time Results from Architecture

No	Model	Time Training
1.	BERT	24 min 00 s
2.	RoBERTa	25 min 39 s
3.	DistilBERT	13 min 34 s

Training Time column in Table 5 presents the training duration for each model:
1. The BERT model has a training time of 24 min 00 s.
2. The RoBERTa model has a training time of 25 min 39 s.
3. The DistilBERT model has a training time of 13 min 34 s.

Results show that DistilBERT achieved training 60% faster than BERT, saving 10 min 26 s, while the RoBERTa model required a longer training time than BERT, adding 1 min 39 s. The training time was measured using the "%%time" command in Google Colab. **Model Evaluation Results:** The model evaluation in this study is based on accuracy, F1-score, and prediction results for the models BERT, RoBERTa, and DistilBERT. The accuracy results are shown in Table 6.

Table 6. Accuracy Results from Architecture

No	Model	Accuracy
1.	BERT	89.60
2.	RoBERTa	90.83
3.	DistilBERT	88.99

The *Accuracy* column in Table 6 presents the accuracy results for each model:
1. The BERT model achieved an accuracy of 89.60.
2. The RoBERTa model achieved an accuracy of 90.83.
3. The DistilBERT model achieved an accuracy of 88.99.

The accuracy of the DistilBERT model (88.99) is nearly equivalent to that of the BERT model (89.60), indicating approximately 97% similarity, despite a 40% reduction in parameter count and 60% faster training time. Meanwhile, the RoBERTa model achieved the highest accuracy (90.83), outperforming the BERT model. Based on the series of model development stages carried out, the next step is to classify the emotions in the constructed test data, as shown in Table 1. Figure 5 presents the prediction results of the BERT model, Figure 6 presents the prediction results of the RoBERTa model, and Figure 7 presents the prediction results of the DistilBERT model. The label codes can be found in Table 3.

Prediction errors in the BERT model:

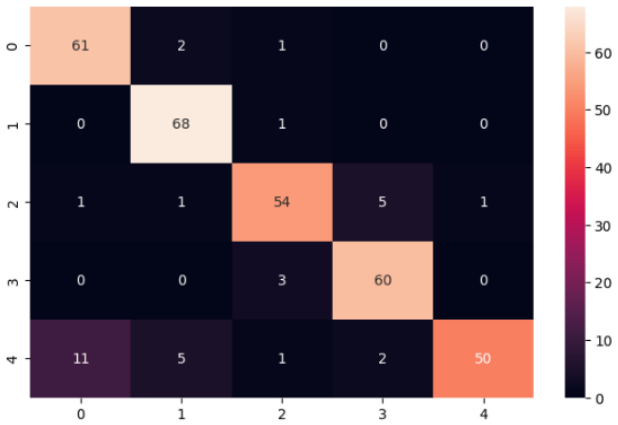

Figure 5. BERT Model Classification Prediction Results

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

a. Anger: misclassified as *fear* = 2, *happy* = 1, *love* = 0, and *sadness* = 0 (total = 3).
b. Fear: misclassified as *anger* = 0, *happy* = 1, *love* = 0, and *sadness* = 0 (total = 1).
c. Happy: misclassified as *anger* = 1, *fear* = 1, *love* = 5, and *sadness* = 1 (total = 8).
d. Love: misclassified as *anger* = 0, *fear* = 0, *happy* = 3, and *sadness* = 0 (total = 3).
e. Sadness: misclassified as *anger* = 11, *fear* = 5, *happy* = 1, and *love* = 2 (total = 19).

Prediction errors in the RoBERTa model:

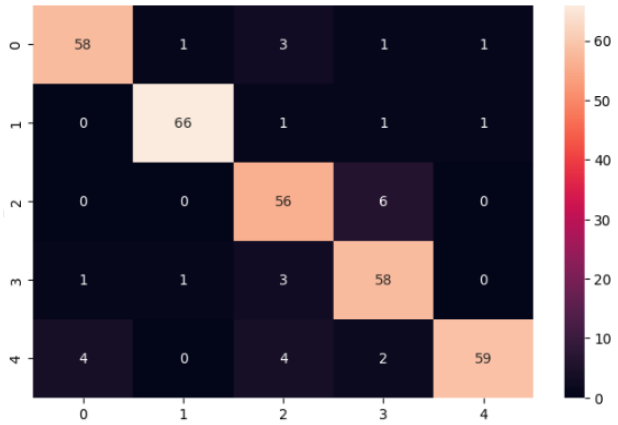

Figure 6. RoBERTa Model Classification Prediction Results

a. Anger: misclassified as *fear* = 1, *happy* = 3, *love* = 1, and *sadness* = 1 (total = 6).
b. Fear: misclassified as *anger* = 0, *happy* = 1, *love* = 1, and *sadness* = 1 (total = 3).
c. Happy: misclassified as *anger* = 0, *fear* = 0, *love* = 6, and *sadness* = 0 (total = 6).
d. Love: misclassified as *anger* = 1, *fear* = 1, *happy* = 3, and *sadness* = 0 (total = 5).
e. Sadness: misclassified as *anger* = 4, *fear* = 0, *happy* = 4, and *love* = 2 (total = 10).

Prediction errors in the DistilBERT model:

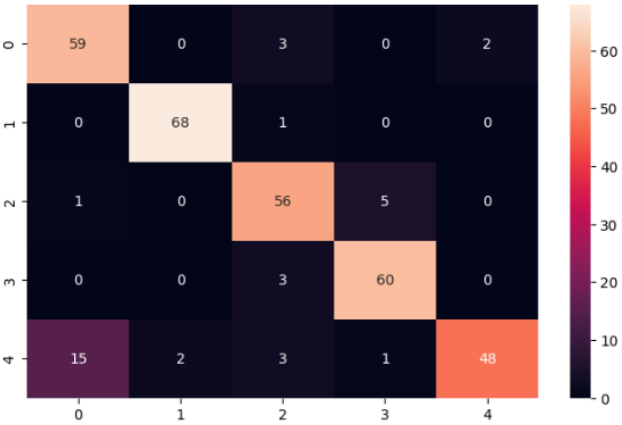

Figure 7. DistilBERT Model Classification Prediction Results

a. Anger: misclassified as *fear* = 0, *happy* = 3, *love* = 0, and *sadness* = 2 (total = 5).
b. Fear: misclassified as *anger* = 0, *happy* = 1, *love* = 0, and *sadness* = 0 (total = 1).
c. Happy: misclassified as *anger* = 1, *fear* = 0, *love* = 5, and *sadness* = 0 (total = 6).
d. Love: misclassified as *anger* = 0, *fear* = 0, *happy* = 3, and *sadness* = 0 (total = 3).
e. Sadness: misclassified as *anger* = 15, *fear* = 2, *happy* = 3, and *love* = 1 (total = 21).

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Based on the prediction error information above, the total number of errors across all labels is **34** for the BERT model, **30** for the RoBERTa model, and **36** for the DistilBERT model. Across the label-wise prediction errors for BERT, RoBERTa, and DistilBERT, the fewest errors are the same across all models, occurring for the **fear** label. This may be because sentences expressing fear typically contain only a few indicative words, such as *"takut"* (afraid). Conversely, the largest number of prediction errors is also consistent across all models, occurring in the **sadness** label. This may be caused by the wide variety of words that can be used to express sadness, such as *crying*, *disappointment*, *hurt*, and sadness. The final stage of this research involves evaluating the model performance using a confusion matrix, which produces the *F1-score* metric. The BERT model achieved an *F1-score* of 90, the RoBERTa model achieved 91, and the DistilBERT model achieved 89. The detailed *F1-score* results for BERT are presented in Figure 8, for RoBERTa in Figure 9, and for DistilBERT in Figure 10.

	precision	recall	f1-score	support
0	0.84	0.95	0.89	64
1	0.89	0.99	0.94	69
2	0.90	0.87	0.89	62
3	0.90	0.95	0.92	63
4	0.98	0.72	0.83	69
accuracy			0.90	327
macro avg	0.90	0.90	0.89	327
weighted avg	0.90	0.90	0.89	327

Figure 8. BERT Model F1-score Results

	precision	recall	f1-score	support
0	0.92	0.91	0.91	64
1	0.97	0.96	0.96	69
2	0.84	0.90	0.87	62
3	0.85	0.92	0.89	63
4	0.97	0.86	0.91	69
accuracy			0.91	327
macro avg	0.91	0.91	0.91	327
weighted avg	0.91	0.91	0.91	327

Figure 9. RoBERTa Model F1-score Results

	precision	recall	f1-score	support
0	0.79	0.92	0.85	64
1	0.97	0.99	0.98	69
2	0.85	0.90	0.88	62
3	0.91	0.95	0.93	63
4	0.96	0.70	0.81	69
accuracy			0.89	327
macro avg	0.90	0.89	0.89	327
weighted avg	0.90	0.89	0.89	327

Figure 10. DistilBERT Model F1-score Results

E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Overall Comparison and Analysis: The resulting parameters, training time, accuracy, and *F1-score* are compared across all developed models to evaluate each model's performance on the Indonesian text emotion classification task using fine-tuning and hyperparameter tuning. Table 7 presents the parameter settings, training time, accuracy, and *F1-score* for each hyperparameter-tuning configuration and fine-tuning process across all models.

Table 7. Parameter Results, Training Time, Accuracy, F1-score of Hyperparameter-tuning and Fine-tuning Architecture

No	Model	Parameter	Training Time	Accuracy	F1-score
1.	BERT	109486085	24 min 00 s	89,60	90
2.	RoBERTa	124649477	25 min 39 s	90,83	91
3.	DistilBERT	66366725	13 min 34 s	88,99	89

This study aims to develop BERT, RoBERTa, and DistilBERT models and evaluate their performance on Indonesian text emotion classification through fine-tuning and hyperparameter tuning. The findings of this research demonstrate the comparative quality of the three models under the same experimental setting. The results are summarized as follows:

a. Model size/parameters: DistilBERT shows a 40% parameter reduction (43,119,360 parameters) compared to BERT, while RoBERTa has 15,163,392 more parameters than BERT, making it the largest model among the three.
b. Training time: DistilBERT trains 60% faster than BERT, saving 10 min 26 s, whereas RoBERTa takes 1 min 39 s longer than BERT, indicating a slower training process.
c. Accuracy: DistilBERT achieves an accuracy of 88.99, which is 97% of BERT's accuracy (89.60), despite having 40% fewer parameters and significantly faster training time. RoBERTa achieves the highest accuracy of 90.83, outperforming both BERT and DistilBERT.

Across all three models, BERT, RoBERTa, and DistilBERT, the highest prediction errors occur in the sadness label. This is likely due to the wide range of vocabulary and expressions associated with sadness ("menangis", "kecewa", "sakit", "sedih"). Conversely, the lowest prediction errors occur in the fear label, which generally contains only a limited number of explicit fear-related expressions ("takut").

CONCLUSION

The BERT, RoBERTa, and DistilBERT models are employed to classify emotions in Indonesian text through a series of well-defined stages. These stages include data collection, data preprocessing, hyperparameter tuning and fine-tuning, evaluation of total parameters, training time measurement, accuracy assessment, analysis of performance improvement during training, examination of per-label prediction results, and evaluation using a confusion matrix and F1-score. The models classify emotions into five labels: anger, fear, happy, love, and sadness using a single-label classification approach. The dataset is processed using an undersampling strategy to balance the classes. Emotion classification is implemented by fine-tuning BERT, RoBERTa, and DistilBERT. Fine-tuning involves adding an output layer and applying the cross-entropy loss function. The output layer generates the predicted emotion labels, while cross-entropy loss computes the difference between the model's predicted output and the expected target. Hyperparameter tuning determines optimal values for batch size = 16, dropout = 0.1, attention dropout = 0.1, epochs = 10, and learning rate = 1e-5. This study successfully produces models that achieve strong performance in emotion classification. Under the applied experimental scenario, BERT achieves 109,486,085 parameters, 24 min 00 s training time, 89.60 accuracy, and an F1-score of 90. RoBERTa achieves 124,649,477 parameters, 25 min 39 s training time, 90.83 accuracy, and an F1-score of 91. DistilBERT achieves 66,366,725 parameters, 13 min 34 s training time, 88.99% accuracy, and an F1-score of 89.

E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

REFERENCES

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer Models for Text-based Emotion Detection: A Review of BERT-based Approaches. *Springer*. https://doi.org/https://doi.org/10.1007/s10462-021-09958-2

Aman, S., & Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. *LNAI*, *4629*, 196–205. https://doi.org/10.1007/978-3-540-74628-7_27

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. https://doi.org/http://arxiv.org/abs/1810.04805

Dong, M. (2018). Emotion Classification on Indonesian Twitter Dataset. *IEEE*. https://doi.org/10.1109/IALP.2018.8629262

Hastomo, W., Bayangkari Karno, A. S., Kalbuana, N., Meiriki, A., & Sutarno. (2021). Characteristic Parameters of Epoch Deep Learning to Predict Covid-19 Data in Indonesia. *Journal of Physics: Conference Series*, *1933*(1). https://doi.org/10.1088/1742-6596/1933/1/012050

He, F., Liu, T., & Tao, D. (2019). *Control Batch Size and Learning Rate to Generalize Well: Theoretical and Empirical Evidence*.

Hidayatullah, A. F., & Ma'Arif, M. R. (2017). Pre-processing Tasks in Indonesian Twitter Messages. *Journal of Physics: Conference Series*, *801*(1). https://doi.org/10.1088/1742-6596/801/1/012072

Huan, H., Guo, Z., Cai, T., & He, Z. (2022). A text classification method based on a convolutional and bidirectional long short-term memory model. *Connection Science*, *34*(1), 2108–2124. https://doi.org/10.1080/09540091.2022.2098926

Huang, J., Li, Y.-F., Xie, M., Huang, J., Xie, M., & Li, Y. F. (2015). A Systematic Analysis of Data Preprocessing for Machine Learning- based Software Cost Estimation. *HAL*, *67*. https://doi.org/10.1016/j.infsof.2015.07.004ï

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, *52*(4). https://doi.org/10.1145/3343440

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011). Twitter trending topic classification. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 251–258. https://doi.org/10.1109/ICDMW.2011.171

Li, W., & Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, *41*(4 PART 2), 1742–1749. https://doi.org/10.1016/j.eswa.2013.08.073

Lin, C. H., & Nuha, U. (2023). Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00782-9

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.1907.11692

Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., & Doulamis, N. (2021). Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies*, *9*(4), 81. https://doi.org/10.3390/technologies9040081

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Nissa, N. K., & Yulianti, E. (2023). Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model. *International Journal of Electrical and Computer Engineering*, *13*(5), 5641–5652. https://doi.org/10.11591/ijece.v13i5.pp5641-5652

E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Rahmad, F., Suryanto, Y., & Ramli, K. (2020). Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification. *IOP Conference Series: Materials Science and Engineering*, *879*(1). https://doi.org/10.1088/1757-899X/879/1/012076

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.1910.01108

Shaver, P. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*. https://doi.org/https://doi.org/10.1037/0022-3514.52.6.1061

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *Springer*. https://doi.org/10.1007/978-3-030-32381-3_16

Wiciaputra, Y. K., Young, J. C., & Rusli, A. (2021). Bilingual text classification in english and indonesian via transfer learning using XLM-RoBERTa. *International Journal of Advances in Soft Computing and Its Applications*, *13*(3), 72–87. https://doi.org/10.15849/ijasca.211128.06

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2009.05387

Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, *507*, 772–794. https://doi.org/10.1016/j.ins.2019.06.064

Yanuar, M., & Shiramatsu, S. (2020). *Aspect Ectractionfor Tourist Spot Review in Indonesian Language Using BERT*.