PERFORMANCE ANALYSIS OF MULTILINGUAL AND MONOLINGUAL MODELS IN PREDICTING INDONESIAN LANGUAGE EMOTION USING TWITTER DATASET

Muhammad Magistra Apta Paramarta[1], Ramaditia Dwiyansaputra[2], Regania Pasca Rassy[3]
[1,2,3]Dept Informatics Engineering, Mataram University
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA
Email: magistraagis@gmail.com, [rama, ganiarachsy]@unram.ac.id

Abstract

Although Indonesia has the third largest population in the world, the number of datasets available in the field of text processing in Indonesian is still very limited. Therefore, this research utilizes the ability of multilingual models that can be trained with multiple languages to predict emotions based on low-resource language such as Indonesian. Several training scenarios were conducted to evaluate the transferability and performance of these multilingual models compared to the monolingual IndoBERT model. The experimental results show that XLM-R outperforms mBERT and achieves competitive performance to IndoBERT, with XLM-R and IndoBERT achieving F1-score of 0.7793 and 0.7733 respectively. XLM-R also demonstrates competitive results on other evaluation metrics. These findings suggest that XLM-RoBERTa could be a promising alternative for emotion detection in languages with limited resources, such as Indonesian.

Key words: *Emotion Detection, Text Classification, Monolingual Model, Multilingual Model*

I. INTRODUCTION

Emotion detection aims to recognize a person's emotional state. Emotions can be identified through several methods, such as voice, facial expressions, and text [1]. Emotion detection has various applications, one of which is in the field of education, where teachers can understand the psychological state of students, thereby helping to improve the learning process, prevent stress, and create a better learning environment [2]. In the business field, emotion detection can be used to analyze customer reviews of products. Emotion detection can help companies understand customer satisfaction levels with products, evaluate products against market needs, and adjust marketing strategies to be more targeted [3]. Since emotions are a component of a person's personality, emotion detection can also be used in employee recruitment and counseling processes. In the employee recruitment process, emotion prediction can be used to select employees who align with the company's needs and help the company find candidates who not only possess technical skills but also have good emotional intelligence. This demonstrates that emotion prediction is highly useful and important for gaining a deeper understanding of an individual's emotional state [4].

There are several types of physiological characteristics that can be used to detect emotions, such as voice intonation, facial expressions, hand movements, body movements, and information from textual data [5]. One of the most commonly used methods in emotion detection is through textual data. This is not only because it is easier, but also because emotions can be expressed implicitly or explicitly in written form.

Currently, social media is often used to express emotions. Twitter is one of the social media platforms with millions of users. Indonesia ranks third in the number of Twitter users from 2012 to 2018 [6]. By using text data from Twitter, emotion detection can be utilized by various groups, such as governments, to detect public satisfaction regarding government policies or political situations. Additionally, emotion detection can also be utilized by companies to monitor public responses regarding products or services they use.

Several studies have conducted emotion detection using datasets sourced from Twitter social media. One study compiled an Indonesian-language dataset from Twitter for emotion classification tasks, with an F1 score of 69.73% [6]. Another study performed emotion detection using a Sundanese-language dataset obtained from Twitter and employed several methods such as K-Nearest Neighbor (KNN), Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM). The results of this study showed that the SVM method had the highest accuracy at 95%, followed by the RF, NB, and KNN methods with accuracies of 91%, 75%, and 69% [7]. Although they have relatively high accuracy, traditional machine learning methods such as SVM, RF, NB, and KNN have several limitations, such as an inability to understand the context within the text and a reliance on feature extraction for predicting emotions using text.

With the development of Natural Language Processing (NLP) technology, several studies have developed a Transformer-based model such as BERT [8] that performs better than traditional machine learning methods in performing NLP tasks [9]. The BERT model itself has been developed and has several variants, such as monolingual models that are trained with only one language and multilingual models that are trained with various languages [10]. Several studies have compared the performance of monolingual and multilingual models in performing NLP tasks such as text classification. Some studies have developed several monolingual models trained with a specific language and multilingual models trained with multiple languages.

One study compared several multilingual models such as mBERT, XLM-RoBERTa, and IndicBERT with several monolingual models such as MahaBERT, MahaALBERT, and MahaRoBERTa. The results of this study show that monolingual models still have superior performance in text classification and hate speech detection [11]. In another study, BERT and ALBERT (A Lite BERT) models were trained in Portuguese and tested for performance in several scenarios, such as sentiment analysis, fake news detection, and others [12].

In predicting emotions in Indonesian, there are several limitations in previous studies that constitute research gaps. One of these is the lack of research focusing on the issue of limited dataset size, which impacts model performance, especially for models with large architectures. For example, study [13] suggests using larger datasets to improve model performance in sentiment analysis tasks.

Indonesian language datasets, especially those used in emotion detection tasks, are still limited compared to datasets in other languages. For example, in the context of emotions, Indonesian language datasets consist of around 4,000 data points [6], which is a relatively small number, especially for use in training deep learning and transformer-based models. For example, in one literature on hate speech classification, such a dataset size is clearly categorized as a low-resource language [14].

This study utilizes Cross-lingual Representation [15] in multilingual models to map data in English to data in Indonesian in a shared vector space. This allows the model to transfer knowledge from English to Indonesian so that multilingual models can recognize words even when written in different languages. The goal is for the model to understand texts in different languages consistently. Additionally, while previous studies have focused on evaluating monolingual models [16], there are still very few studies comparing the performance of multilingual models such as XLM-RoBERTa and mBERT with monolingual models like IndoBERT in Indonesian emotion detection.

This study aims to evaluate the performance of monolingual and multilingual models in detecting emotions based on text through three scenarios: (1) multilingual, which involves training a multilingual model with English and Indonesian, then testing it with Indonesian data. (2) zero-shot, which involves training a multilingual model with English data only, then testing it on Indonesian data; and (3) monolingual, which involves training and testing the model with Indonesian data only. The models used include IndoBERT, mBERT, and XLM-RoBERTa.

By conducting this scenario, this study aims to determine whether multilingual models trained with English data can transfer knowledge and provide competitive performance in detecting emotions for an low-resource language, such as Indonesian.

II. Related Works

A. Pre-Trained Language

Pre-trained Language Models (PLMs) are often used in the field of NLP, which employs transfer learning methods in the training process. These models are trained using large, general datasets and then used to perform more specific tasks through fine-tuning. This approach offers advantages such as reducing the computational costs of training a model [14]. The development of PLM in the field of NLP is marked by the release of an architecture called Transformers, which has superior performance [8]. However, the development of PLM has only focused on languages that are rich in resources. This has caused PLM to work optimally on languages that have larger datasets, such as English and Chinese, for training PLM. Multilingual models were developed to address this issue by modeling the semantic representation of resource-rich languages with resource-poor languages [17]. Additionally, PLMs also show improved performance when trained with larger and more diverse datasets [15].

B. Monolingual Language Model

A Monolingual Language Model is a PLM that is trained and tested in several NLP tasks such as text generation and text classification using a single language. In its development, there are several Monolingual Language Models developed based on the BERT model. IndoBERT is one of the Monolingual Language Models trained using the Indonesian language and performs well when tested in several NLP tasks such as Named Entity Recognition (NER), sentiment analysis, and emotion detection [18].

C. Multilingual Language Model

A multilingual language model is a model trained with multiple languages using a large dataset. Several studies have used multilingual models to perform various NLP tasks. XLM-RoBERTa is one such multilingual model developed with 100 languages [15]. The XLM-RoBERTa model has been proven to perform better than other multilingual models in cross-language testing. Meanwhile, mBERT is a multilingual model based on the BERT model, trained with 104 different languages. Research on mBERT indicates that the mBERT model performs exceptionally well in zero-shot and cross-lingual model transfer tasks [19].

One study developed a model called XLM-T, which is an extension of the XLM-RoBERTa model trained using a dataset from Twitter to perform various tasks such as sentiment and emotion analysis in eight different languages. The XLM-T model outperforms the XLM-RoBERTa model with an average accuracy of 69.35% in each trial [20]. In another study, the XLM-EMO model [18] was specifically developed to detect emotions based on text. This model was developed based on the XLM-RoBERTa model but was specifically trained using a dataset with 19 different languages. The performance of XLM-EMO is superior to several other models in emotion detection, particularly when detecting emotions in low-resource languages using the zero-shot method. The XLM-EMO model achieves an average F1-score of 0.84, outperforming both the XLM-RoBERTa and XLM-Twitter models [21].

Based on several previous studies, multilingual models (such as XLM-RoBERTa, XLM-EMO, and XLM-T) have performed quite well in emotion detection in low-resource languages [21][22]. However, unlike previous studies, this research does not use specific models for emotion detection, such as XLM-EMO and XLM-T, but instead focuses on evaluating multilingual models trained with general datasets like XLM-RoBERTa and mBERT. The purpose of this approach is to measure the transfer learning capability of multilingual models that have not been optimized for emotion classification tasks, particularly under resource-constrained conditions in the Indonesian language. Additionally, this study aims to evaluate the capability of baseline models models trained using general datasets and not specifically developed for a particular task to transfer knowledge from English to Indonesian. The English language dataset was chosen because English is one of the languages with large resources.

D. Cross-lingual Transfer Learning

Cross-lingual transfer learning is a method that utilizes high-resource languages to overcome the lack of datasets in low-resource languages [15]. Several previous studies have used the cross-lingual transfer learning method in performing various NLP tasks such as Named Entity Recognition [23] and sentiment analysis [22]. These studies demonstrate that cross-lingual transfer learning can help address the lack of dataset size, particularly in low-resource languages. This research focuses on the application of cross-lingual transfer learning in emotion classification tasks for the Indonesian language. Emotion classification has distinct characteristics compared to other NLP tasks such as sentiment analysis, text summarization, or Named Entity Recognition.

Unlike sentiment analysis, which generally categorizes text based on labels such as positive, negative, and neutral, emotion classification involves detecting basic emotions such as joy, anger, sadness, fear, and love. This makes emotion classification a more complex multi-class classification task because emotions have semantic dimensions that are often ambiguous, overlapping, and dependent on cultural and linguistic contexts. For example, the emotions of love and joy are often expressed in similar linguistic forms, making them difficult to distinguish using traditional feature-based models. Therefore, while transformer-based multilingual models have been successfully applied to other NLP tasks, it is important to evaluate their performance and adaptability specifically for emotion detection tasks, particularly in the context of low-resource languages with limited data sources.

III. Methodology

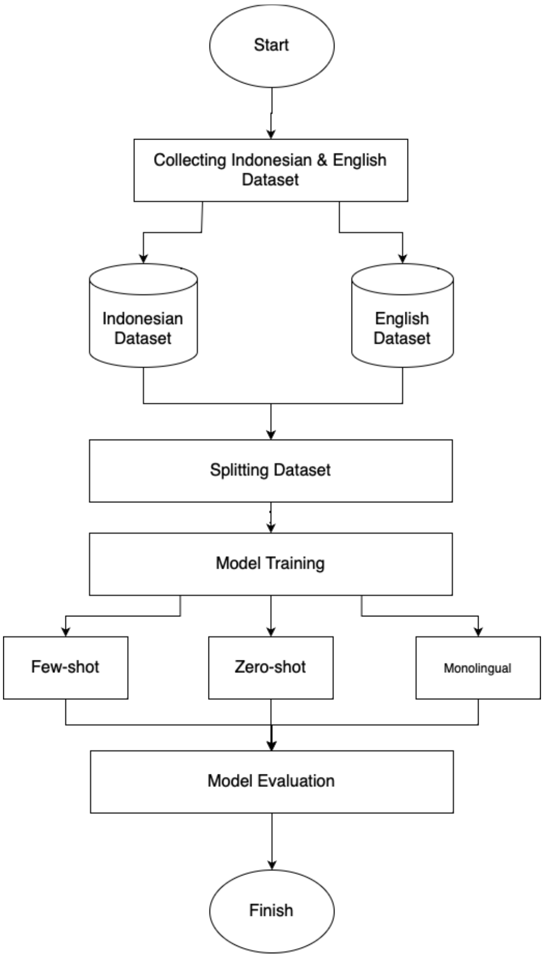

Fig. 1. *Research flow.*

In this study, there are several stages of activities carried out in the research process, as follows.

A. Collecting Dataset

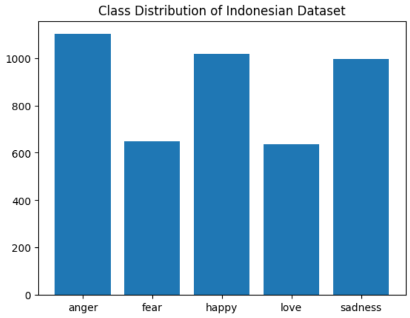

Fig. 2. *Class Distribution in Indonesian Dataset.*

This study used two datasets in two different languages, Indonesian and English. The Indonesian dataset was obtained from research conducted by Saputri [6] with the aim of creating a dataset for predicting emotions in Indonesian sourced from Twitter. The dataset collection process was carried out in two stages, namely collecting data from Twitter for two weeks and producing 4,043 lines of data. After the data was collected, a data annotation process was carried out to label each dataset with five basic emotion categories based on the emotion model created by Parrot [24], such as *anger*, *happy*, *sadness*, *fear*, and *love*.

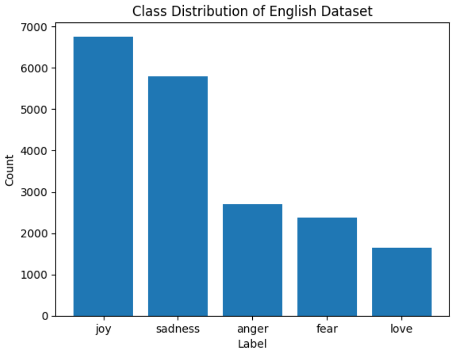

Fig. 3. *Class Distribution in English Dataset.*

For the English dataset, this study used the Saravia dataset [25], which aims to create an emotion dataset using a method called distant supervision. The dataset collection process was carried out on the Twitter platform using several hashtags related to basic emotions. The published dataset consists of six emotion labels: *anger*, *joy*, *sadness*, *fear*, *love*, and *surprise*. To compare the performance of multilingual and monolingual models, the number of emotion labels in English was adjusted to match the Indonesian dataset by dropping the *surprise* emotion label from the English dataset and we mapped *joy* emotion to *happy*.

TABLE I. DATASETS EXAMPLE IN INDONESIAN

No	Text	Label
1	Soal jln Jatibaru,polisi tdk bs GERTAK gubernur .Emangny polisi tdk ikut pmbhasan? Jgn berpolitik. Pengaturan wilayah,hak gubernur. Persoalan Tn Abang soal turun temurun.Pelik.Perlu kesabaran. [USERNAME] [USERNAME] [URL]	Anger
2	Kepingin gudeg mbarek Bu hj. Amad Foto dari google, sengaja, biar teman-teman jg membayangkannya. Berbagi itu indah.	Happy
3	Orang lain kalau pake ponco itu buat jas hujan, nah dia pake buat kasur. Ya tadi gara2 saking gak punya apa2. Mamak bilang, kami tuh di awal pernikan gak ada ngalamin bulan madu kayak skrg2. Org tidur nya aja pake ponco. Gimane mau bulan madu.	Sadness

TABLE II. DATASETS EXAMPLE IN ENGLISH

No	Text	Label
1	i am ever feeling nostalgic about the fireplace i will know that it is still on the property	Love
2	i lost my special mind but don t worry i m still sane i just wanted you to feel what i felt while reading this book i don t know how many times it was said that sam was special but i can guarantee you it was many more times than what i used in that paragraph did i tell you she was special	Happy
3	i hate living under my dads roof because it gives him an excuse to be an asshole to me because hes providing for me to live here i think he feels that he needs to make me feel as unwelcome as possible so ill leave	Sadness

B. Splitting Dataset

TABLE III. DATA SPLITTING DISTRIBUTION

No	Dataset	Train	Validation	Test
1	Indonesian Dataset	3.082	660	660
2	English Dataset	15.425	3.856	-

Table III shows the number of dataset divisions in the two datasets used. The Indonesian language dataset is divided into 3.082 training data, 660 test data, and 660 validation data. Meanwhile, the English language dataset is only divided into 15.425 training data and 3.856 validation data, without test data. This is because the English language dataset is only used as a source for knowledge transfer to the Indonesian language during the training of the multilingual model. This division aims to ensure that each model is trained with sufficient data in each scenario and can be objectively evaluated on data that has never been seen before during the training process. We shuffle the dataset during the splitting process to prevent order bias and ensure that both the training and test sets are representative of the overall data distribution.

C. Model Training

In the training process, this study used the IndoBERT model as a monolingual model, and mBERT and XLM-R as multilingual models.

- IndoBERT-base-p2 (IndoBERT) is a model specifically developed for the Indonesian language, so it has a deeper understanding of the structure and context of the Indonesian language.

- multilingual-bert-base-uncased (mBERT) is a multilingual model based on BERT, the model is trained with more than 100 languages.
- XLM-RoBERTa-base (XLM-R) is a multilingual model developed from the RoBERTa architecture and trained with larger and more diverse data, designed to perform well across various languages.

Each model is trained for 10 epochs, a number chosen based on initial evaluations that showed the model's performance stagnated after the 10th epoch. The purpose of this training is to compare the performance of monolingual and multilingual models in classifying Indonesian emotions. The entire training process is carried out using Google Colaboratory with TPUV2 as the accelerator. The model training process for each scenario was performed only once due to computational limitations.

TABLE IV. MODEL SPECIFICATION

No	Model	Parameter
1	Multilingual-BERT-uncased	179 Million
2	XLM-RoBERTa-base	279 Million
3	IndoBERT-base-p2	124.5 Million

The three models used in the model training process in predicting emotions in Indonesian have different specifications on parameter size. The mBERT model has about 179 million parameters, while the XLM-R model has 279 million parameters. Meanwhile, the IndoBERT model has a parameter size of about 124.5 million parameters. The XLM-R model has the largest number of parameters which allows the model to have better text representation capabilities compared to other models. Despite the difference in parameter size, each model is trained with the same hyperparameters with a learning-rate of 2e-5 and a batch-size of 16. This is done to maintain consistency during the training process and the process of comparing models in predicting emotions in Indonesian is done fairly without differences in training configuration.

In the multilingual scenario, the mBERT and XLM-R models were trained using a combination of Indonesian and English datasets, then validated by predicting emotions using an Indonesian dataset. Each model was trained using an English dataset divided into training and validation data with (80:20) ratio. After being trained with the English-language dataset, the model was retrained using the Indonesian dataset divided into training, validation and test set with a ratio of (70:15:15). This scenario aims to evaluate how the addition of multilingual data can improve the model's performance in predicting emotions in unseen languages, such as Indonesian.

In the zero-shot scenario, we conduct two experiment to test multilingual performance on unseen language with Indonesian and Sundanese dataset. The mBERT and XLM-R models were trained with an English language dataset. The dataset was divided into training data and validation data with (80:20) ratio. The trained model was then used to predict emotions in Indonesian and Sundanese, aiming to evaluate its transfer learning capabilities across languages. This scenario tested the model’s ability to generalize to unseen languages in a multilingual setting.

In the monolingual scenario, models such as mBERT, XLM-R, and IndoBERT were trained using Indonesian language datasets and tested to predict emotions in Indonesian. The dataset used is divided into training, validation, and test data with a ratio of (70:15:15). This scenario aims to evaluate the performance of multilingual models such as mBERT and XLM-R with monolingual models such as IndoBERT in predicting emotions in Indonesian.

D. Evaluation Metrics

To measure the performance of the model, several metrics such as precision, recall, and F1-score are used to provide an overview of the model's performance in predicting emotions. Precision indicates the proportion of correct emotion predictions out of all predictions made for a class. Recall measures how well the model can find all correct examples for an emotion. F1-score is the harmonic mean of precision and recall, and provides a more balanced assessment especially when the data is not balanced.

IV. RESULTS

Table VI and VII shows the performance of each model in predicting emotions using the Indonesian and English datasets. In comparing the performance of each model, this study uses the accuracy, F1-Socre, Recall, and Precision metrics as a reference to assess the performance of each model in predicting emotions.

TABLE V. PREDICTION RESULTS IN ZERO-SHOT AND MULTILINGUAL SCENARIO

Model	Zero-shot Indonesian			Zero-shot Sundanese			Multilingual		
	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision
mBERT	0.2853	0.3109	0.3557	0.2193	0.2799	0.3077	0.6445	0.6393	0.6515
XLM-R	**0.4419**	**0.4495**	**0.5174**	**0.3576**	**0.3962**	**0.4586**	**0.7373**	**0.7428**	**0.7357**

TABLE VI. PREDICTION RESULT IN MONOLINGUAL SCENARIO

Model	Monolingual		
	F1-Score	Recall	Precision
mBERT	0.6632	0.6610	0.6689
XLM-R	**0.7793**	**0.7887**	**0.7742**
IndoBERT	0.7733	0.7749	0.7739

A. Performance of Multilingual Models in Zero-shot and Multilingual Scenario

In the zero-shot scenario, The mBERT model achieved an an F1-score of 0.2853, recall of 0.3109, and precision of 0.3557. The XLM-R model showed higher results in this scenario, with an an F1-score of 0.4419, recall of 0.4495, and precision of 0.4495. XLM-R consistently outperformed mBERT across all metrics (F1-score, recall, and precision) in both languages, demonstrating stronger cross-lingual transfer capabilities.

However, In the multilingual scenario, both model have better performance compared to zero-shot scenario with mBERT model achieved an F1-score of 0.6445, recall of 0.6393, and precision of 0.6515. Meanwhile, the XLM-R model showed better performance with an F1-score 0.7373, recall of 0.7428, and a precision of 0.7357. This improvement in performance highlights the positive impact of fine-tuning on English and Indonesian, demonstrating that such an approach can significantly enhance the effectiveness of multilingual models.

B. Model Performance in the Monolingual Scenario

Meanwhile, in the monolingual scenario, XLM-R outperformed mBERT and slightly outperformed IndoBERT, achieving an F1-score of 0.7793 compared to IndoBERT's 0.7733. Although IndoBERT is a model specifically pre-trained on Indonesian, XLM-R—despite being a general multilingual model—demonstrated superior performance across all metrics (F1-score, recall, and precision). This suggests that XLM-R's broader multilingual training may contribute to its robustness, even in single-language tasks like Indonesian emotion prediction.

C. Discussion

The overall results in Tables V and VI indicate that the XLM-R model consistently outperforms the mBERT model in various scenarios, especially in monolingual and zero-shot scenarios.

A significant decline in performance occurred in the zero-shot scenario, demonstrating the challenges of model generalization to languages that have not been trained before. Nevertheless, the XLM-R model outperformed the mBERT model. Architecturally, XLM-R has larger parameters and model capacity than mBERT, and was trained with a larger and more diverse dataset. This enables XLM-R to have better generalization capabilities and better understand cross-language context, including in cross-language knowledge transfer.

Then, adding an English dataset to the multilingual scenario improved the performance of the mBERT and XLM-R models. The XLM-R model had higher metric values for precision, recall, and F1-score compared to mBERT. This indicates that the XLM-R model is better at identifying emotions evenly across all emotion classes. This shows that even though English datasets were added to both models, the number of parameters and architecture influence both models in predicting emotions.

In the monolingual scenario, the XLM-R model showed competitive performance compared to the other two models, including IndoBERT, which was specifically trained for Indonesian. This shows that even though IndoBERT has a lighter architecture and XLM-R is a multilingual model, both are capable of effectively capturing the linguistic representation of Indonesian.

D. Error Analysis

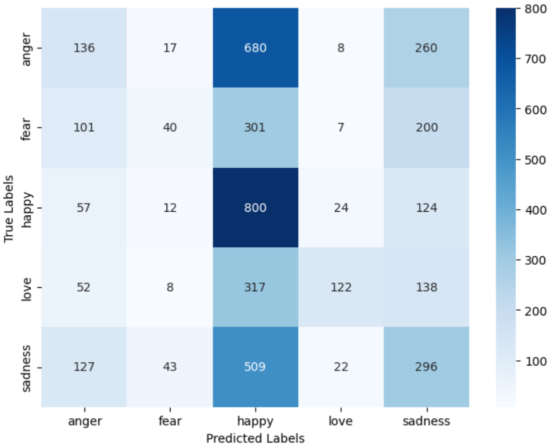

Fig. 4. *Confusion Matrix mBERT in Zero-shot Scenario.*

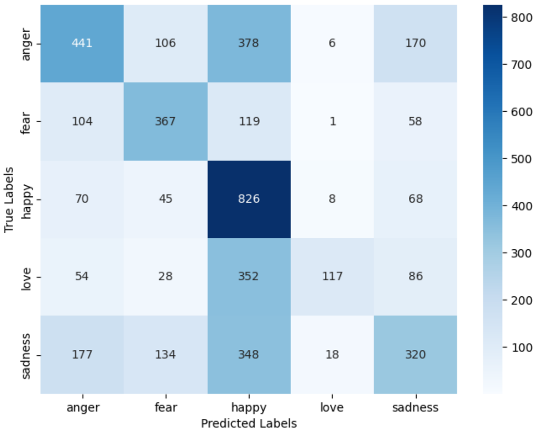

Fig. 5. *Confusion Matrix XLM-R in Zero-shot Scenario.*

In the zero-shot scenario, the model was trained using an English-language dataset and evaluated with an emotion prediction task in Indonesian. Based on the confusion matrix in Figure 2, the mBERT model shows a tendency to predict most samples as the *happy* emotion class. This bias may be due to the imbalance in the English training dataset, where the happy class has the highest number of samples. Although the *happy* class has the highest number of correct predictions in this model, its performance on other emotion classes such as *anger* and *love* is quite low, indicating a prediction bias toward a dominant class.

Meanwhile, in Figure 3, the XLM-R model also showed the best performance in the *happy* class, but overall provided more balanced results compared to mBERT, with more accurate predictions in the *anger*, *sadness,* and *fear* classes.

Both models tend to classify the emotion *happy* more easily. This tendency is likely due to the unbalanced class distribution in the dataset, where the *happy* class has a dominant proportion, causing the model to prioritize that class in its predictions.

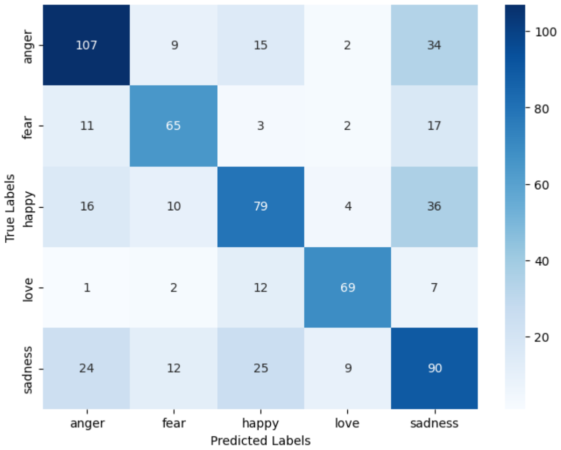

Fig. 6. *Confusion Matrix mBERT in Multilingual Scenario.*

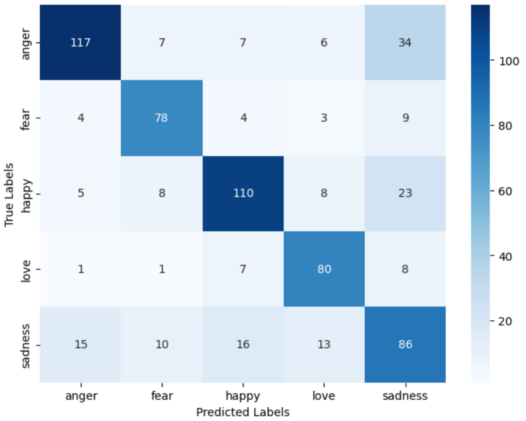

Fig. 7. *Confusion Matrix mBERT in Multilingual Scenario.*

In a multilingual scenario, the model was trained using a combination of English and Indonesian datasets, then evaluated by predicting emotions in Indonesian data. Based on the analysis of the confusion matrix in Figure 3, the mBERT model is able to classify several emotion classes well. However, there are still significant prediction errors, such as 25 data points in the *sadness* class being predicted as *happy* and 34 data points in the *anger* class being predicted as *sadness*.

In Figure 4, the XLM-R model shows better performance than mBERT in the same scenario, but still produces some classification errors. For example, 34 data points in the *anger* class and 23 data points in the *happy* class were classified as *sadness,* indicating ambiguity in distinguishing certain emotional expressions.

Overall, the addition of English data in the multilingual scenario improved the emotion classification performance of both models compared to the zero-shot scenario. However, these results also show that both models still face challenges in distinguishing between emotion classes that have contextual similarities or unbalanced data sets.

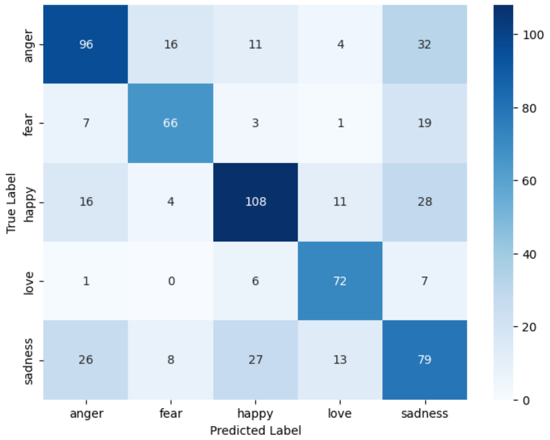

Fig. 8. *Confusion Matrix mBERT in Monolingual Scenario.*

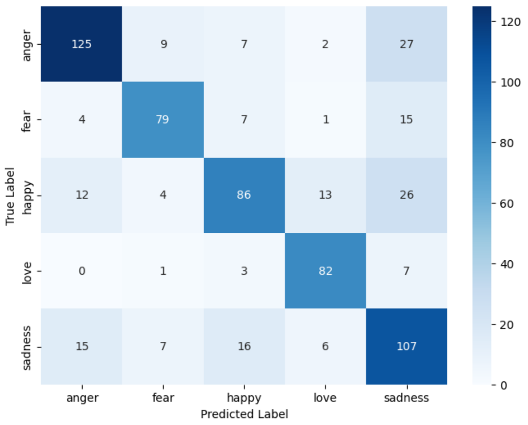

Fig. 9. *Confusion Matrix XLM-R in Monolingual Scenario.*

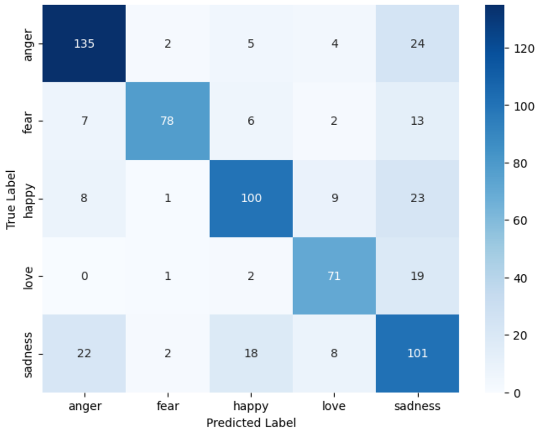

Fig. 10. *Confusion Matrix IndoBERT in Monolingual Scenario.*

In the monolingual scenario, each model was trained with an Indonesian language dataset and then evaluated for classifying emotions in Indonesian. Based on the confusion matrix results in Figure 6, the mBERT model correctly predicted the emotion class *happy* the most, but other emotion classes such as *anger* and *fear* were predicted as *sadness*.

Figure 7 shows that the XLM-R model produces more balanced results than the mBERT model. However, there are still some errors in several emotion classes, such as *anger* and *fear*. Meanwhile, the IndoBERT model has some errors in predicting emotions but has better generalization capabilities than the two previous models.

Then, Figure 8 shows the performance results of the IndoBERT model in a monolingual scenario. The IndoBERT model successfully predicted several emotion classes better than the mBERT and XLM-R models. However, there were several emotion classes that were mispredicted, such as the emotion classes *anger*, *happy*, and *love*, which were predicted as *sadness*.

TABLE VII. EXAMPLES OF TEXTS WITH MISCLASSIFIED EMOTIONS

No	Text	Predicted Label	Actual Label
1	RT [USERNAME] Dia suka mengopi, tapi tidak bersamamu. Dia suka jalan-jalan, tapi bukan jalan denganmu. Dia suka apa saja yang juga kau suka, tapi dia tidak menyukaimu..	Love	Sadness
2	Nona tidak perlu takut, toh ini hanya gerimis kan? Sudah ku siapkan rumah untuk berteduh barangkali nanti berubah menjadi hujan angin. #ceritakemarinsore #spoiler	Fear	Love
3	Aku termasuk orang yang tidak habis pikir dengan orang yang menganggap wajar hal tersebut dan justru menyalahkan Via Vallen karena dianggap lebay, namanya pelecehan dalam bentuk apapun itu pasti menyakitkan. Tidak peduli jenis pelecehan apa itu.	Sadness	Anger

Based on the error analysis, each model still tends to make errors in classifying certain emotions even though the models have been trained with two different languages. As shown in Table VII, some texts are ambiguous and make it difficult for the model to accurately predict the intended emotion, as the emotion is not always expressed explicitly. Linguistic cultural differences and the ways of expressing emotions in each language also pose challenges for the models, thereby affecting their performance in predicting emotions [26][27].

V. CONCLUSION

This study aims to evaluate the performance of multilingual and monolingual models in classifying emotions in Indonesian through three training scenarios: multilingual, zero-shot, and monolingual. Based on experiments conducted on the three scenarios multilingual, zero-shot, and monolingual it was found that the XLM-R model performed better than the other multilingual models.

In multilingual and zero-shot scenarios, XLM-R consistently outperforms the mBERT model across all evaluation metrics. In monolingual scenarios, the performance of XLM-R and IndoBERT is relatively competitive. This is evident in the metrics between XLM-R and IndoBERT, which are almost equivalent.

The results of the experiment show that the multilingual XLM-R model is capable of producing competitive results in Indonesian emotion classification tasks, even when compared to the monolingual IndoBERT model. This shows that multilingual models can be an effective alternative, especially in cases where data is limited or the language has a low data source.

However, each model still has errors in some emotion labels, which may be caused by linguistic cultural differences, especially in expressing emotions.

Suggestions for further research include increasing the amount of data, especially Indonesian language data, and using languages that are linguistically close to Indonesian to improve model performance.

REFERENCES

[1] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding Emotions in Text Using Deep Learning and Big Data," Comput. Hum. Behav., vol. 93, pp. 309–317, Apr. 2019, doi: 10.1016/j.chb.2018.12.029.

[2] A. O. R. Vistorte, A. Deroncele-Acosta, J. L. M. Ayala, A. Barrasa, C. López-Granero, and M. Martí-González, "Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review," Front. Psychol., vol. 15, p. 1387089, Jun. 2024, doi: 10.3389/fpsyg.2024.1387089.

[3] A. Basuki, "Sentiment Analysis of Customers' Review on Delivery Service Provider on Twitter Using Naive Bayes Classification," vol. 9, no. 2, 2023, doi: https://doi.org/10.26555/jiteki.v9i2.26327.

[4] P. S. Dandannavar, S. R. Mangalwede, and P. M. Kulkarni, "Social Media Text - A Source for Personality Prediction," in 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India: IEEE, Dec. 2018, pp. 62–65. doi: 10.1109/CTEMS.2018.8769304.

[5] C. Domenico, "Emotions That Influence Purchase Decisions And Their Electronic Processing ," Ann. Univ. Apulensis Ser. Oeconomica, vol. 2, no. 11, pp. 996–1008, Dec. 2009, doi: 10.29302/oeconomica.2009.11.2.45.

[6] M. S. Saputri, R. Mahendra, and M. Adriani, "Emotion Classification on Indonesian Twitter Dataset," in 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia: IEEE, Nov. 2018, pp. 90–95. doi: 10.1109/IALP.2018.8629262.

[7] O. V. Putra, F. M. Wasmanson, T. Harmini, and S. N. Utama, "Sundanese Twitter Dataset for Emotion Classification," in 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), Surabaya, Indonesia: IEEE, Nov. 2020, pp. 391–395. doi: 10.1109/CENIM51130.2020.9297929.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, arXiv: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[9] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," J. Comput. Cogn. Eng., vol. 2, no. 4, pp. 352–356, Apr. 2023, doi: 10.47852/bonviewJCCE3202838.

[10] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych, "How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online: Association for Computational Linguistics, 2021, pp. 3118–3135. doi: 10.18653/v1/2021.acl-long.243.

[11] A. Velankar, H. Patil, and R. Joshi, "Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi," vol. 13739, 2023, pp. 121–128. doi: 10.1007/978-3-031-20650-4_10.

[12] D. de V. Feijo and V. P. Moreira, "Mono vs Multilingual Transformer-based Models: a Comparison across Several Language Tasks," Jul. 19, 2020, arXiv: arXiv:2007.09757. doi: 10.48550/arXiv.2007.09757.

[13] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 8, 2023, doi: 10.14569/IJACSA.2023.0140813.

[14] M. Weißenbacher and U. Kruschwitz, "Steps towards Addressing Text Classification in Low-Resource Languages".

[15] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics, 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[16] R. Pramana, M. Jonathan, H. S. Yani, and R. Sutoyo, "A Comparison of BiLSTM, BERT, and Ensemble Method for Emotion Recognition on Indonesian Product Reviews," Procedia Comput. Sci., vol. 245, pp. 399–408, 2024, doi: 10.1016/j.procs.2024.10.266.

[17] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-Trained Language Models and Their Applications," Engineering, vol. 25, pp. 51–65, Jun. 2023, doi: 10.1016/j.eng.2022.04.024.

[18] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Oct. 08, 2020, arXiv: arXiv:2009.05387. doi: 10.48550/arXiv.2009.05387.

[19] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics, 2019, pp. 4996–5001. doi: 10.18653/v1/P19-1493.

[20] F. Barbieri, L. E. Anke, and J. Camacho-Collados, "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond," 2021, arXiv. doi: 10.48550/ARXIV.2104.12250.

[21] F. Bianchi, D. Nozza, and D. Hovy, "XLM-EMO: Multilingual Emotion Prediction in Social Media Text," in Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 195–203. doi: 10.18653/v1/2022.wassa-1.18.

[22] A. Kumar and V. H. C. Albuquerque, "Sentiment Analysis Using XLM-R Transformer and Zero-shot Transfer Learning on Resource-poor Indian Language," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 20, no. 5, pp. 1–13, Sep. 2021, doi: 10.1145/3461764.

[23] S. O. Khairunnisa, Z. Chen, and M. Komachi, "Dataset Enhancement and Multilingual Transfer for Named Entity Recognition in the Indonesian Language," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 22, no. 6, pp. 1–21, Jun. 2023, doi: 10.1145/3592854.

[24] A. Bandhakavi, N. Wiratunga, S. Massie, and D. P., "Emotion-aware polarity lexicons for Twitter sentiment analysis," Expert Syst., vol. 38, no. 7, p. e12332, Nov. 2021, doi: 10.1111/exsy.12332.

[25] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, “CARER: Contextualized Affect Representations for Emotion Recognition,” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3687–3697. doi: 10.18653/v1/D18-1404.

[26] S. Havaldar, B. Singhal, S. Rai, L. Liu, S. C. Guntuku, and L. Ungar, “Multilingual Language Models are not Multicultural: A Case Study in Emotion,” in Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 202–214. doi: 10.18653/v1/2023.wassa-1.19.

[27] S. Hassan, S. Shaar, and K. Darwish, “Cross-lingual Emotion Detection,” May 04, 2022, arXiv: arXiv:2106.06017. doi: 10.48550/arXiv.2106.06017.