Multilabel classification sentiment analysis on Indonesian mobile app reviews

Riccosan, Karen Etania Saputra
Computer Science Department, School of Computer Science, Bina Nusantara University Bandung Campus, Jakarta, Indonesia

Article Info

Article history:

Received Oct 17, 2024
Revised Jul 2, 2025
Accepted Aug 6, 2025

Keywords:

Indonesia
Mobile application review
Multiclass
Multilabel
Sentiment analysis
Text classification

ABSTRACT

Mobile applications continue to evolve to satisfy the users. For that, the developers need to understand user feedback for improvements. Indonesia, one of the countries with the most mobile app users, has many textual mobile app reviews that may be processed and analyzed. Understanding the value of mobile app reviews requires understanding the value of sentiments and emotions to create more appropriate features to satisfy the users. To acquire a more accurate analysis of user reviews, it is important to detect sentiments that are closely associated with human emotion values due to the nature of multilabeled data. This research classifies the sentiments and emotions in Indonesian textual mobile app reviews, which are multilabel and multiclass in the form of 3 sentiments, namely positive, negative, and neutral, paired with 6 emotions, namely anger, sad, fear, happy, love, and neutral. We employ the Transformers architecture model, which includes two monolingual (a generic English and an Indonesian) and a multilingual pre-trained models with the results: bidirectional encoder representations from transformers (BERT) base uncased (micro avg. F1-score=0.69, precision=0.68, recall=0.70, receiver operating characteristic-area under the curve (ROC-AUC)=0.78), IndoBERT base uncased as best result (micro avg. F1-score=0.77, precision=0.78, recall=0.76, ROC-AUC=0.85), and multilingual BERT (M-BERT) base uncased (micro avg. F1-score=0.72, precision=0.73, recall=0.71, ROC-AUC=0.82).

Corresponding Author:

Riccosan
Computer Science Department, School of Computer Science, Bina Nusantara University Bandung Campus
Jakarta, 11480 Indonesia
Email: riccosan@binus.ac.id

1. INTRODUCTION

As technology advances, more mobile applications/apps are created and accessible globally [1]. These applications can now be accessed for free or through subscription [2] and Indonesia is no exception for this event. User reviews for mobile applications are usually obtained through public posts and performance or feature ratings on social media or comments on the application download page [3] in the mobile apps store (like the Google Play Store or iOS App Store). Mobile app users voluntarily provide reviews for the mobile apps [4], which developers can use as data to understand general user needs preferences, criticism, and suggestions [5]. Mobile app reviews hold great significance for developers or companies as they offer valuable insights [6] in the form of user sentiment and emotional values related to the application, including ratings and suggestions for desired or undesirable features as well as dissatisfaction with performance [7]. To design and develop good mobile apps that are appropriate or fulfill their user's needs and desires, the developers or company must identify and evaluate the sentiments and emotions of mobile app users [8].

Indonesia, being one of the highest countries of smartphone device users in the world, with a total of 249.22 million users [9] since 2024, has a significant market opportunity for mobile applications and has the potential to become a data source for mobile app reviews. An important thing is the textual data of mobile app reviews provide information about user needs and demands [10]. Those data will be enormous, with variations of information in the form of words and phrases. To obtain the sentiment and emotional value required by mobile app developers from those data, the process will be very time-consuming and the result is potentially biased if it is done manually [11] by humans. For that reason, this research applied sentiment analysis task as part of natural language processing (NLP), in the form of text classification [12]. In our research, the text classification task becomes a multi-label and multi-class process based on the associated sentiment and emotion values contained in the mobile application reviews. Multi-label classification is a classification task on data that has more than one type of class [13] for the grouping process and attaches more than a class for every data. Where in this research sentiment is the first-class type and emotion is the second-class type. On the other hand, multi-class classification is a classification task that uses more than 2 classes for the process but only one label attached to the data [14]. For our experiment, the text classification task is no longer done manually by humans but carried out by utilizing a computerized systematic learning method [12], namely machine learning (ML) with its deeper level, namely deep learning (DL).

There are already some earlier works on sentiment analysis, especially by utilizing text classification. This research provides helpful conceptual groundwork for our work. Start with the implementation of k-nearest neighbor (KNN) for Indonesia news article topic classification with multi-label characteristics [15]. However, the research discovered challenges with a lack of research data, resulting in unsatisfactory model performance. Another research implemented the multilingual DL model to improve the quality of Indonesian text classification [16]. That research proves by utilizing the DL multilingual model, the Indonesian text classification and analysis is becoming better than using the base monolingual model that is trained with English corpus. The third research [17] tried to combine two DL models from different architectures, namely bidirectional long-short term memory (Bi-LSTM) combined with Transformers [18] pre-trained model, namely bidirectional encoder representative from transformers (BERT) [19] for hate-speech classification from the Indonesian tweet dataset. The research succeeded in obtaining a fairly high level of accuracy for the text classification task but still had a drawback because the model was trained with a general textual dataset and not for a specific task. Therefore Hendrawan *et al.* [17] suggested using specially created datasets provided by researchers to train the model from the DL architecture for a particular task. Next, textual hate-speech classification on Indonesian tweet data [20], with the main model support vector machine (SVM) compared with convolutional neural network (CNN) [21], and DistilBERT [22]. This research utilized a dataset from previous work [23] which contained three base languages, namely Indonesian, English, and Hindi. The dataset was translated into Indonesian and supplemented with new data by Hana *et al.* [20]. According to the research, SVM is the most effective model for the task and the classification can perform better when textual data is pre-processed without stemming stage and stopword removal.

Based on the recent work, this research utilizes the experiment by applying several related suggestions obtained. First, select a few models from DL architecture for the sentiment analysis task, and second, utilize the self-generated dataset specified for sentiment and emotion classification. This research dataset was from previous research about Indonesian mobile application reviews containing multi-label multi-class sentiment and emotion class [24]. For this research model, we utilized the Transformers [18]: a multilingual type, namely multilingual BERT (M-BERT) [19] and two monolingual models, the IndoBERT [25], a specified model for Indonesian NLP and the BERT base [19] as a comparison. Transformers architecture has advantages, that are simple model structure with good performance [26], the ability to do parallel data processing, and the ability to extract dense information from textual data with attention mechanisms [18] which provide relation values between terms. Another advantage of Transformers is the ease of fine-tuning models [27] through layer and parameter configuration, transfer learning, or embedding with other architectures. Last, BERT is a pre-trained model that has been trained with combined data from more than 100 languages [26]. This proves that the BERT model has been designed specially to carry out the NLP task.

This research contributed to the most recent sentiment analysis experiment, specifically a multi-label multi-class classification task through model training for extracting sentiments and emotional values from Indonesian textual data, which are still limited. This research also presents a brief understanding of multi-label multi-class classification tasks and presents useful recommendations for future research in the area of sentiment analysis through text classification. Next part, we describe the dataset used in training and evaluating the utilized DL model, the method and research flow, and the model training configuration. For the rest, this article presents the experiment result in section 3 result and discussion, and the research conclusion with a few suggestions for future research in section 4 conclusion.

2. METHOD

This research consists of 5 stages, namely data preparation, parameter tuning, data tokenization, model training, and evaluation as shown in Figure 1. The first is data preparation, where raw data is converted into a dataframe with separated columns of text and classes. Next, the data is shuffled to ensure there are learning variations [28] in the DL model's algorithm. After that, the data is split according to its function in the experiment, namely train, valid, and test. In the first stage, the classes are tokenized with one-hot encoding since the classes and their data have been separated. The utilization of one-hot encoding may guarantee that each pair of classes is unique [29], which is important for this research that employs multi-label multi-class classification. With one-hot encoding, the pair of classes will be converted into a binary block and the method sets only true value where the classes exist in the sequence as shown in Figure 2.

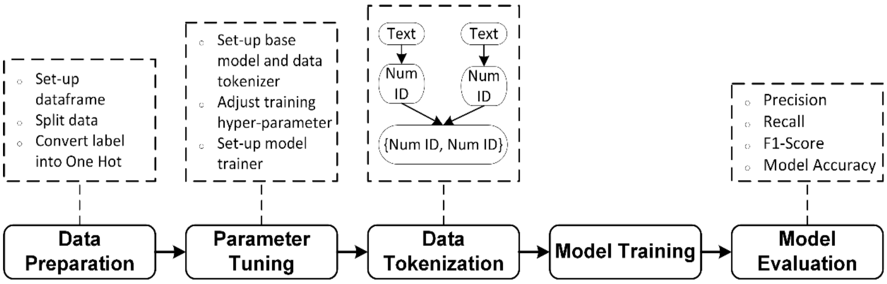

Figure 1. Research flow

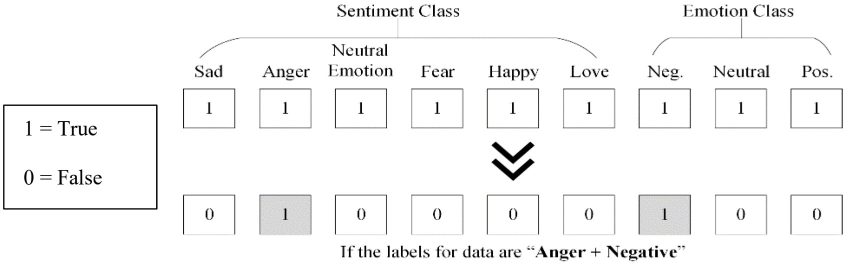

Figure 2. An example of one-hot encoding output

The second stage is the implementation of parameter tuning which will be used in the DL model training process. This stage starts with the selection of base models along with its textual data tokenizer. In this research, we utilized the Transformers architecture [18] of several base models, namely BERT [19] with monolingual and multilingual types, as well as the IndoBERT [25], a monolingual model specified for Indonesian NLP tasks. For the data tokenizer, we utilized the BERT tokenizer because its data conversion can be implemented for other Transformers models with the base structure same as the BERT model. After the base models and tokenizer set-up, a model trainer function was created to configure the training parameters, namely epochs, training batch-size, learning-rate, weight decay, logging steps, and, optimizer. The third is data tokenization, the process of textual data conversion into a block of vector. In the tokenization, each word gets different values because of the attention mechanism from the Transformers [18] which shows how closely the word is related to another one. The fourth stage is model training which implements the parameter tuning and model trainer function that has been made from the second stage.

In the fifth stage, experiment results are collected and model evaluation is carried out using test data to obtain evaluation results in the form of precision, recall, F1-score, and receiver operating characteristic-area under the curve (ROC-AUC) as shown in calculations (1)-(4). The calculation results will be separated for every class and then averaged with the micro averaging so that the weight distribution is balanced for each class with the division based on the total amount of test data [30]. Data for precision, recall, F1-score, and ROC-AUC calculations are the results obtained from the confusion matrix process in the form of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). After the final evaluation, results are obtained and the analysis is carried out. To get the ROC-AUC value, first need to calculate the recall of true positive rate (TPR) as in (2) and the false positive rate (FPR) as in (4) for every single class. Next, calculate the ROC-AUC as in (5) with TPR and FPR values from every single class in the test data.

$$Precision_{micro\_avg.} = \frac{TP}{TP+FP} \tag{1}$$

$$Recall\ (or\ TPR)\ _{micro\_avg.} = \frac{TP}{TP+FN} \tag{2}$$

$$F1 - Score_{micro\_avg.} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

$$FPR_{micro\_avg.} = 2 \times \frac{FP}{FP + TN} \tag{4}$$

$$ROC - AUC_{micro\_avg.} = \int_0^1 TPR_{micro\_avg.}\ (FPR_{micro\_avg.})\ dFPR_{micro\_avg.} \tag{5}$$

Implementing ROC-AUC in model evaluation is better, especially for multilabel classification tasks and imbalanced datasets, than the traditional accuracy. The ROC method can distinguish between the true and false prediction rates [31] by evaluating each class as a single class, even though they are paired as multilabel. Next point, the calculation combined with the AUC to get a deeper insight by calculating the TPR against the FPR to help distinguish positive and negative rates for all test data in a paired form compared to the real label. The ROC-AUC is better on imbalanced data than the accuracy method that only focuses on a single threshold of label-to-label comparison to get the positive and negative prediction values [32].

3. RESULTS AND DISCUSSION

3.1. Dataset

This research implemented a dataset that was created from our previous research about Indonesian mobile application reviews [24] that contain 21,697 user reviews. The utilized dataset is a multi-label multi-class type because of two types of classes, namely sentiment class consisting of positive, negative, and neutral; and emotion class consisting of sad, anger, fear, happy, love, and neutral. This research dataset is ready-to-use because the dataset has been through the cleaning step and pre-processed in the previous research [24]. In the experiment, the dataset was only pre-processed to be converted into data frames, split based on its function, randomization of the data order, and data tokenization into a numeric to be used in the training and model evaluation process. The data distribution based on its class can be seen inside Figure 3.

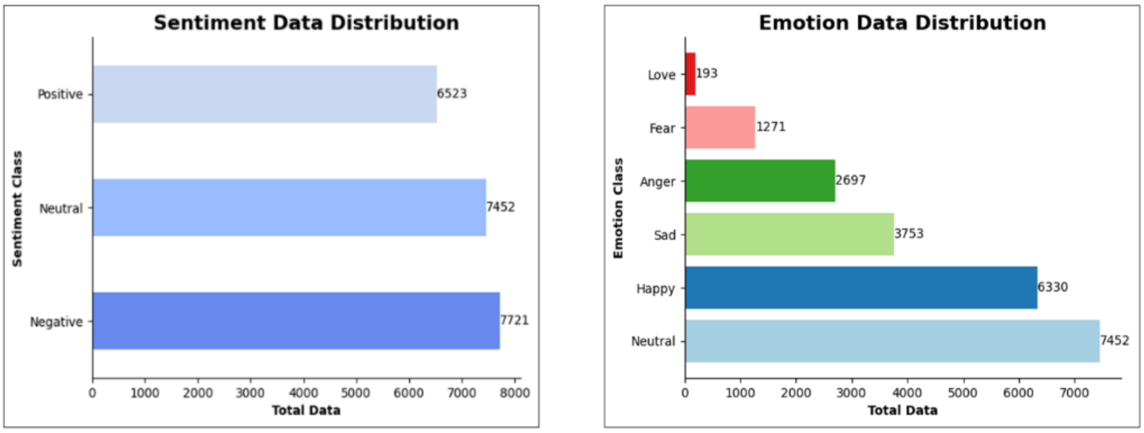

Figure 3. Data distribution based on sentiment and emotion class

Based on Figure 3, there is diversity in the data distribution. From the sentiment class, there is a range of data differences around 900 to 1,200 data. In the emotion class, the data is dominated by happy and neutral, whereas the data distribution gap looks significant in love, fear, anger, and sad labels. This research utilized the original amount of data to gain intrinsic value about model performance on the imbalanced dataset. To overcome the data distribution diversity in the dataset, we combine the data label for every class, with a pattern positive sentiment label with love/happy; negative sentiment label with fear/anger/sad; and neutral sentiment label with neutral emotion (i.e. [positive, love] or [negative, sad]), as shown in Figure 2.

Next, we split the data into 3 parts, namely train, valid, and test. The data splitting has the purpose of decreasing the potential of model over-fit [33], especially at the model evaluation stage. By splitting the data in train, valid, and testing, the utilized model will not read the same data twice. The data splitting process in this research was done using a single data sampling method [34] for every data. The first ratio is 75% from 21,696 data as 16,272 train data and the second ratio is around 25% (this ratio then restarted as 100% for residual data) is split again with a ratio of 90% for 4,882 valid data and 10% for 542 test data.

3.2. Parameter tuning and experiment

There are three base models utilized in this research that come from transformers architecture [18] with the uncased type, namely BERT base, M-BERT [19], and IndoBERT [25]. We used the uncased type with the insight that there are differences in the form of upper-case and lower-case characters that might have some emphasis on sentiment or emotion [35] in Indonesian mobile app reviews. In the experiment stage, a function was created that served as the DL model trainer. Few training parameters have been determined in that function, namely epochs=10, because the utilized base models are pre-trained, the training process is only focussing on gaining data patterns for the related task; training batch-size=16, to split total training data processed per epoch into smaller groups in forming more structured and denser data; learning-rate=3e-5, is the optimum value for the speed of the learning process by the model and this value has been determined as a recommendation from the original research of BERT model [19]; weight decay=0.01, is a parameter that normalizes the training process by giving a penalty point to the excessive training weight, so that the potential of model over-fit can be reduced; logging-steps=100, is a parameter to determine the model checkpoint steps in validating training performance; training evaluation_metric=eval_f1 (F1-score) [36] and ROC-AUC [37], [38], to get validation score from model performance during the training process; and training optimizer=AdamW_Torch (Adam Weighted) [39], [40], a function that optimize the model training process by limiting the range of data score that taken as learning value to no less than 0 and no more than 1, by implement the score from weight decay calculation that added separately so it does not directly affect the value of training momentum. After the training parameters are determined, the training and evaluation for every model are carried out alternately.

The following are comparisons of the model training process results in the form of training and validation loss, training accuracy, and training F1-score and ROC-AUC. The first comparison is training and validation loss, every model utilized in this research successfully in decreasing the training loss value. However, this is the opposite of the validation loss value that keeps increasing until the last step as seen in Figure 4. The lowest training loss value is from BERT base with a value of 0.0363, while M-BERT is in the second position, and IndoBERT is in the highest of all. From the validation loss value, M-BERT obtained the lowest result with a value of 0.4393, while IndoBERT is in the second position and BERT Base is in the highest position. From this experiment's results, it is known that all utilized models are experiencing over-fitting because of imbalanced data distribution. This experiment’s results also show that the base model (BERT base) can learn slightly better than the derived models (M-BERT and IndoBERT). From the general side, the multilingual model has a better general understanding of the training data than other models.

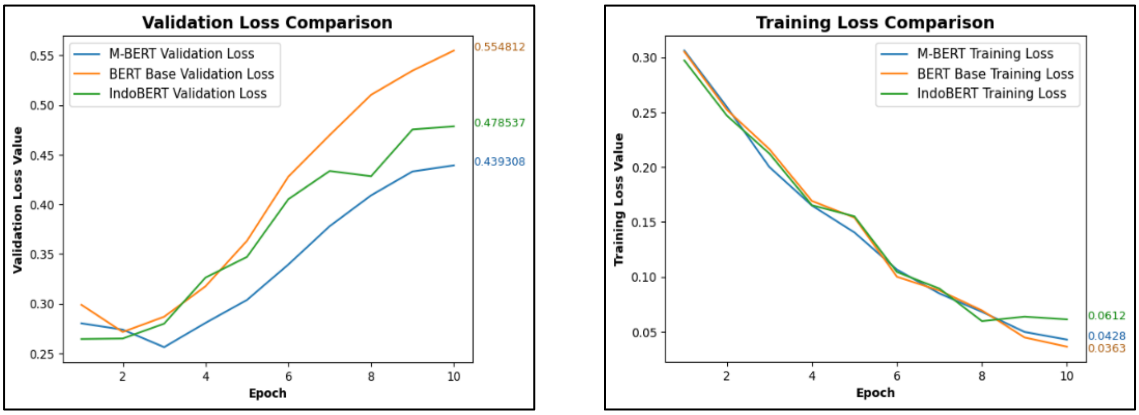

Figure 4. Training and validation loss comparison

The next comparison is training accuracy and F1-score (Figure 5). From the training accuracy value, IndoBERT occupied the top position with a value of 0.6988, while BERT Base is in the second position with a value of 0.6743, and M-BERT is in the third position with a value of 0.6511. In the F1-score value comparison, IndoBERT once again comes out in the first position with a value of 0.7589, while M-BERT is in the second position with a value of 0.7495, and followed by BERT base with a value of 0.7360. These results show that a pre-trained model that is built specifically for a single language can influence the performance of NLP tasks, especially for classification on multi-label multi-class textual data. From the results obtained, BERT base and M-BERT still have good performance, but the data generalization performance is not as good as the monolingual model, IndoBERT, which was specifically built for the Indonesian NLP task like in this research that utilized a specific Indonesian textual dataset.

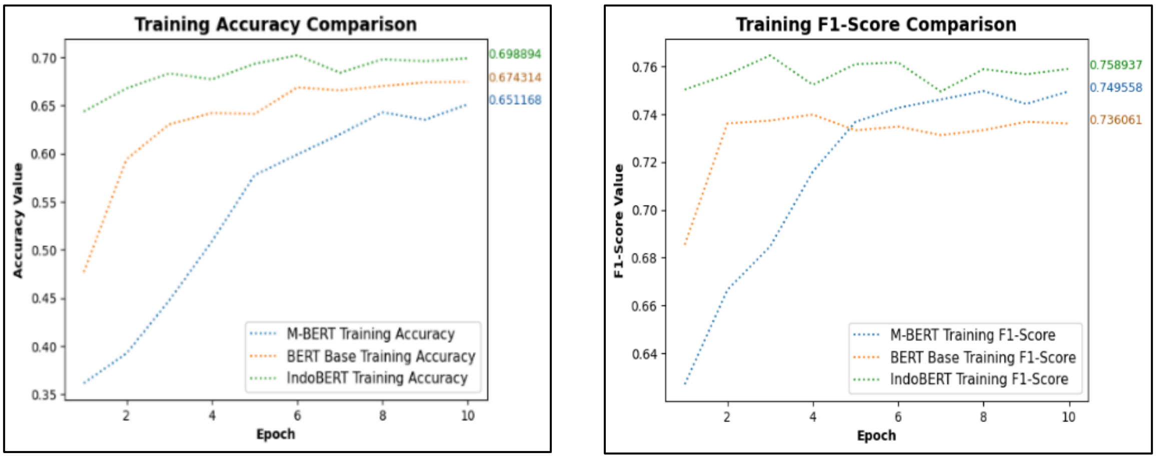

Figure 5. Training accuracy and F1-score comparison

The following is Table 1 which compares all results from the model training stage. IndoBERT obtained the best performance based on training accuracy and F1-score, even though the training and validation loss values are slightly higher than those of BERT Base and M-BERT. The model training process results show that the implementation of the textual data classification task depends on the model's type that is utilized, especially if the task is monolingual or specific for some language. To implement the classification task on monolingual data, this research recommends establishing a model or implementing a pre-trained model that also has a monolingual characteristic that is specific to a language. Another thing that needs attention is the data splitting process only is not enough to encounter the potential of over-fitting from every model. Applying a more balanced data distribution or data with more guaranteed quality will certainly give more significant results for the performance related to the Indonesian textual data classification task.

Table 1. Model training results

Model	Training loss	Validation loss	Training accuracy	Training F1-score
BERT base	0.0363	0.5548	0.6743	0.7360
Multilingual BERT (M-BERT)	0.0428	0.4393	0.6511	0.7495
IndoBERT	0.0612	0.4785	0.6988	0.7589

3.3. Final results and evaluation

The last stage is the prediction test utilizing test data to evaluate the model performance. In this stage, the trained model predicts the class pair from every test data totalling 542 sentences. After the prediction test, a comparison process is carried out on the true and predicted labels, followed by calculation to obtain the average values in the form of TP, TN, FP, and FN. Those values are then used to obtain the final evaluation results for each model in the form of precision, recall, F1-score, and ROC-AUC. As shown in the Figure 6, the best result of this research from the evaluation stage is the IndoBERT model with the final result being precision=0.78, recall=0.76, F1-score=0.77, and ROC-AUC=0.85. The M-BERT obtained the second position with the final result being precision=0.73, recall=0.71, F1-score=0.72, and ROC-AUC=0.82, followed by the BERT base model in the third position with the final result being precision=0.68, recall=0.70, F1-score=0.69, and ROC-AUC=0.78.

From this experiment evaluation results, IndoBERT still gets the best performance with a slightly higher F1-score value of 0.05 points compared to M-BERT and 0.08 points compared to BERT base. These results prove that a monolingual model trained specifically with Indonesian data obtained more optimal performance in carrying out the classification task on Indonesian textual data in extracting sentiment and emotion values. Although M-BERT is a pre-trained model that has been trained with the multilingual textual dataset, the ability to pay attention in terms of the Indonesian language remains fragmented and does not have a focus as well as a monolingual model. As for the BERT base, still has a disadvantage because this model only applies attention scheme contextually for every word in the data, but the language terms still need to be translated into the base language used to train the model. This lack of ability could be the reason for the decrease in model performance on a textual data classification task because the data that is translated into another language could have synonyms and potentially change the true value from its original form.

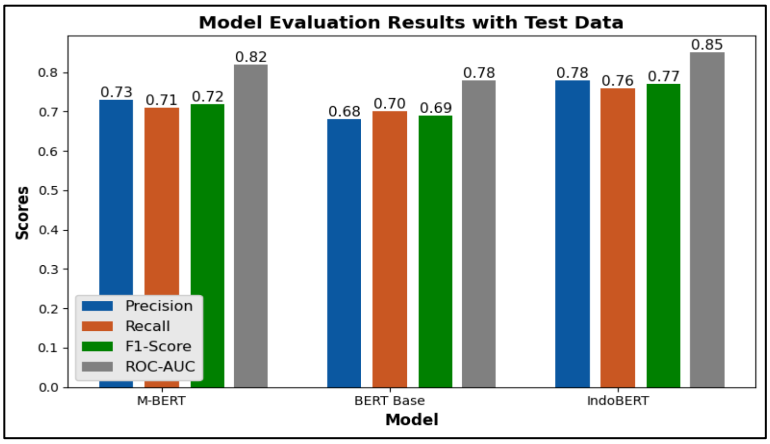

Figure 6. Model evaluation results

4. CONCLUSION

This research has successfully experimented on a multi-label and multi-class classification with three models from transformers architecture in obtaining the sentiment and emotion value from Indonesian textual data, especially from mobile application reviews that could be used as a reference for similar future research. The implementation of classification task on mobile application review is beneficial in knowing the reputation of a feature or the mobile application itself in the middle of the user community which can lead to strategic decision-making in mobile application development in the future. Multi-label and multi-class classification could be carried out if there is a combination of contexts between two or more types of classes in the utilized data. This research has shown the three pre-trained models utilized from transformers architecture, namely BERT, IndoBERT, and M-BERT have provided a fairly good performance on the classification of multi-label Indonesian mobile app reviews. The best evaluation result is from IndoBERT, in the form of precision=0.78, recall=0.76, F1-score=0.77, and ROC-AUC=0.85. Behind the good results, still there are problems, especially with model over-fit and imbalanced data distribution that cause the decreasing of performance in the implementation of the classification task. On that basis, the experiment that was carried out through this research still has potential for future development. On the other side, the improvement of the dataset for multi-label multi-class classification on mobile application review must continue because the data will continue to grow and become more diverse according to the market demand. In the future, our research will do experiments in model developments which are smaller, but with a target of obtaining a better ability to carry out the multi-label multi-class classification on textual Indonesian mobile application review in sentiment analysis segments as part of NLP.

FUNDING INFORMATION

No funding involved in the research process of this research article.

AUTHOR CONTRIBUTION STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Riccosan	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Karen Etania Saputra		✓	✓	✓		✓	✓	✓	✓	✓	✓			

C :	**C**onceptualization	I :	**I**nvestigation	Vi :	**Vi**sualization
M :	**M**ethodology	R :	**R**esources	Su :	**Su**pervision
So :	**So**ftware	D :	**D**ata Curation	P :	**P**roject administration
Va :	**Va**lidation	O :	Writing - **O**riginal Draft	Fu :	**Fu**nding acquisition
Fo :	**Fo**rmal analysis	E :	Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

The authors in this research article declare that they have no known competing financial interests or personal relationships that could have appeared to influence the research work reported in this paper.

DATA AVAILABILITY

The data used for this research is open access and can be found in this research article: “Multilabel multiclass sentiment and emotion dataset from indonesian mobile application review” [24] or through this link: https://doi.org/10.1016/j.dib.2023.109576.

REFERENCES

[1] A. Alzhrani, A. Alatawi, B. Alsharari, and U. Albalawi, “Towards security awareness of mobile applications using semantic-based sentiment analysis,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 800–809, 2022, doi: 10.14569/IJACSA.2022.0130493.

[2] S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, “Discrepancy detection between actual user reviews and numeric ratings of Google App store using deep learning,” *Expert Systems with Applications*, vol. 181, 2021, doi: 10.1016/j.eswa.2021.115111.

[3] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, “Analysing app reviews for software engineering: a systematic literature review,” *Empirical Software Engineering*, vol. 27, no. 2, 2022, doi: 10.1007/s10664-021-10065-7.

[4] M. Umer, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, “Predicting numeric ratings for Google Apps using text features and ensemble learning,” *ETRI Journal*, vol. 43, no. 1, pp. 95–108, 2021, doi: 10.4218/etrij.2019-0443.

[5] G. P. Wiratama and A. Rusli, “Sentiment analysis of application user feedback in Bahasa Indonesia using multinomial naive Bayes,” *Proceedings of 2019 5th International Conference on New Media Studies, CONMEDIA 2019*, pp. 223–227, 2019, doi: 10.1109/CONMEDIA46929.2019.8981850.

[6] Y. Wang, J. Wang, H. Zhang, X. Ming, L. Shi, and Q. Wang, “Where is your app frustrating users?,” *Proceedings - International Conference on Software Engineering*, vol. 2022-May, pp. 2427–2439, 2022, doi: 10.1145/3510003.3510189.

[7] M. Pandey, R. Litoriya, and P. Pandey, “Perception-based classification of mobile apps: a critical review,” *Smart Computational Strategies: Theoretical and Practical Aspects*, pp. 121–133, 2019, doi: 10.1007/978-981-13-6295-8_11.

[8] V. Balakrishnan, P. K. Selvanayagam, and L. P. Yin, “Sentiment and emotion analyses for Malaysian mobile digital payment applications,” *ACM International Conference Proceeding Series*, pp. 67–71, 2020, doi: 10.1145/3388142.3388144.

[9] H. N.-Woff, “Number of mobile internet users in Indonesia 2020-2029,” *Statista*, 2024. Accessed: Jul. 07, 2024. [Online]. Available: https://www.statista.com/statistics/558642/number-of-mobile-internet-user-in-indonesia/

[10] W. Chen, Z. Xu, X. Zheng, Q. Yu, and Y. Luo, “Research on sentiment classification of online travel review text,” *Applied Sciences*, vol. 10, no. 15, 2020, doi: 10.3390/APP10155275.

[11] C. Yang, L. Wu, C. Yu, and Y. Zhou, “A phrase-level user requests mining approach in mobile application reviews: concept, framework, and operation,” *Information*, vol. 12, no. 5, 2021, doi: 10.3390/info12050177.

[12] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: a survey,” *Information*, vol. 10, no. 4, 2019, doi: 10.3390/info10040150.

[13] M. L. Zhang, Y. K. Li, H. Yang, and X. Y. Liu, “Towards class-imbalance aware multi-label learning,” *IEEE Transactions on Cybernetics*, vol. 52, no. 6, pp. 4459–4471, 2022, doi: 10.1109/TCYB.2020.3027509.

[14] R. M. Mathew and R. Gunasundari, “A review on handling multiclass imbalanced data classification in education domain,” *2021 International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2021*, pp. 752–755, 2021, doi: 10.1109/ICACITE51222.2021.9404626.

[15] N. Isnaini, Adiwijaya, M. S. Mubarok, and M. Y. A. Bakar, “A multi-label classification on topics of Indonesian news using k-nearest neighbor,” *Journal of Physics: Conference Series*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012027.

[16] I. F. Putra and A. Purwarianti, “Improving Indonesian text classification using multilingual language model,” *2020 7th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2020*, 2020, doi: 10.1109/ICAICTA49861.2020.9429038.

[17] R. Hendrawan, Adiwijaya, and S. Al Faraby, “Multilabel classification of hate speech and abusive words on Indonesian Twitter social media,” *2020 International Conference on Data Science and Its Applications, ICoDSA 2020*, 2020, doi: 10.1109/ICoDSA50139.2020.9212962.

[18] A. Vaswani *et al.*, “Attention is all you need,” 31st Conference on Neural Information Processing Systems (NIPS 2017), Aug. 2023,pp. 6000-6010, doi: 10.5555/3295222.3295349.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *Proceedings of NAACL-HLT 2019*, May 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.

[20] K. M. Hana, Adiwijaya, S. Al Faraby, and A. Bramantoro, “Multi-label classification of Indonesian hate speech on Twitter using support vector machines,” *2020 International Conference on Data Science and Its Applications, ICoDSA 2020*, 2020, doi: 10.1109/ICoDSA50139.2020.9212992.

[21] L. D. L. C. Jackel, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, B. L. Cun, J. Denker, and D. Henderson, “Handwritten digit recognition with a back-propagation network,” *Advances in Neural Information Processing Systems*, pp. 396–404, 1990.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv-Computer Science*, pp. 1-5, Mar. 2020.

[23] M. O. Ibrohim and I. Budi, “Multi-label hate speech and abusive language detection in Indonesian Twitter,” *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 46–57, doi: 10.18653/v1/w19-3506.

[24] Riccosan and K. E. Saputra, “Multilabel multiclass sentiment and emotion dataset from Indonesian mobile application review,” *Data in Brief*, vol. 50, 2023, doi: 10.1016/j.dib.2023.109576.

[25] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: a benchmark dataset and pre-trained language model for Indonesian NLP,” *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[26] Riccosan, R. Sutoyo, and A. Chowanda, “Sentiment classification for Indonesian sentences using multilingual transformers model,” *ICIC Express Letters*, vol. 16, no. 10, pp. 1047–1055, 2022, doi: 10.24507/icicel.16.10.1047.

[27] A. N. Azhar and M. L. Khodra, “Fine-tuning pretrained multilingual BERT model for indonesian aspect-based sentiment analysis,” 2020, doi: 10.1109/ICAICTA49861.2020.9428882.

[28] P. Ganesh, H. Chang, M. Strobel, and R. Shokri, “On the impact of machine learning randomness on group fairness,” *ACM International Conference Proceeding Series*, pp. 1789–1800, 2023, doi: 10.1145/3593013.3594116.

[29] J. T. Hancock and T. M. Khoshgoftaar, “Survey on categorical data for neural networks,” *Journal of Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00305-w.

[30] M. Heydarian, T. E. Doyle, and R. Samavi, “MLCM: multi-label confusion matrix,” *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[31] A. J. Bowers and X. Zhou, “Receiver operating characteristic (ROC) area under the curve (AUC): a diagnostic measure for evaluating the accuracy of predictors of education outcomes,” *Journal of Education for Students Placed at Risk*, vol. 24, no. 1, pp. 20–46, 2019, doi: 10.1080/10824669.2018.1523734.

[32] A. M. Carrington *et al.*, “Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 329–341, 2023, doi: 10.1109/TPAMI.2022.3145392.

[33] V. R. Joseph, “Optimal ratio for data splitting,” *Statistical Analysis and Data Mining*, vol. 15, no. 4, pp. 531–538, 2022, doi: 10.1002/sam.11583.

[34] Z. Meng, R. McCreadie, C. MacDonald, and I. Ounis, “Exploring data splitting strategies for the evaluation of recommendation models,” *RecSys 2020 - 14th ACM Conference on Recommender Systems*, pp. 681–686, 2020, doi: 10.1145/3383313.3418479.

[35] Z. Halim, M. Waqar, and M. Tahir, “A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email,” *Knowledge-Based Systems*, vol. 208, 2020, doi: 10.1016/j.knosys.2020.106443.

[36] P. Christen, D. J. Hand, and N. Kirielle, “A review of the F-measure: its history, properties, criticism, and alternatives,” *ACM Computing Surveys*, vol. 56, no. 3, 2024, doi: 10.1145/3606367.

[37] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997, doi: 10.1016/S0031-3203(96)00142-2.

[38] A. Tharwat, “Classification assessment methods,” *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.

[39] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[40] J. Bjorck, K. Q. Weinberger, and C. P. Gomes, “Understanding decoupled and early weight decay,” *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, vol. 8A, pp. 6777–6785, 2021, doi: 10.1609/aaai.v35i8.16837.

BIOGRAPHIES OF AUTHORS

Riccosan obtained his master’s degree in Computer Science from Bina Nusantara University. He currently serves as a full-time lecturer in the computer science program at Bina Nusantara University’s Bandung Campus. His primary research interest revolves around emotions and sentiment analysis, delving into the fascinating world of understanding human emotions through technology. He can be contacted at email: riccosan@binus.ac.id.

Karen Etania Saputra received her master’s degree in Computer Science from Bina Nusantara University. She is a full-time lecturer in the Computer Science Program at Bina Nusantara University’s Bandung Campus. Her primary research focus lies in the field of deep learning. She can be contacted at email: karen.saputra@binus.ac.id.