p-ISSN 2301–4156 | e-ISSN 2460–5719

Optimizing the Accuracy of the Semantic-Based Compound Emotion Classifications using the XLM-RoBERTa

Aripin[1], Steven Adi Santoso[2], Hanny Haryanto[3]

[1] Department of Biomedical Engineering Universitas Dian Nuswantoro, Semarang, 50131 INDONESIA (tel.: 024-70793733; fax: 024-3569684; email: [1]arifin@dsn.dinus.ac.id)
[2,3] Department of Informatics Engineering Universitas Dian Nuswantoro, Semarang 50131 INDONESIA (tel.: 024-3517261; fax: 024-3560567; email: [2]sstevenadi@gmail.com, [3]hanny.haryanto@dsn.dinusl.ac.id)

[Received: 9 December 2022, Revised: 2 February 2023]
Corresponding Author: Aripin

ABSTRACT — There are six basic emotions; they are anger, sadness, happiness, disgust, surprise, and fear. A combination of basic emotions creates a new type of emotion called a compound emotion. The examples of applying these compound emotions are in chatbots, translations, and text summarization. Several research on classifying these emotions based on Indonesian texts have used traditional models such as multinomial naïve Bayes, support vector machine (SVM), k-nearest neighborhood, and term frequency–inverse document frequency (TF-IDF). The previous research have a massive drawback, primarily on their less optimized performances. The models used could only classify things with the available data; thus, the text processing is required that results in a longer training time for larger This research aims to solve the issue from the previous research by using cross-lingual language model-robustly optimized bidirectional encoder representations from transformers approach (XLM-RoBERTa) model to classify compound emotions based on the semantics or meaning in words and sentences. The XLM-RoBERTa is a transformer model that can identify the meaning of a word from its attention mechanism and represent it as a vector to know the usage and position in a sentence. It is also a method to understand the meaning of a specific word. Using the attention mechanism, the model used the word position to recognize the sentence pattern and classify them even further to know the pattern and sequence to understand the semantics. The experiment result showed that the model could classify Indonesian texts into basic and compound emotion classes with an accuracy of up to 95.56%. This result is much higher than using traditional models to classify the compound emotion classes.

KEYWORDS — Compound Emotion Classification, Indonesian Sentences, Multilabel, Semantics, XLM-RoBERTa.

I. INTRODUCTION

An emotion is a feeling that someone or something feels. It is also a reaction from someone to a situation [1] and can be manifested as happiness, sadness, or fear. Emotions are a unique element to all living creatures, which humans especially use to communicate their feeling. Emotions can also affect their behavior, such as when they are happy they may smile or laugh. Those who are sad, on the other hand, will likely build a wall and refuse to see anyone.

Behavior is not the only way to show emotions. Voices and texts can also convey emotions, such as in a story or poetry. Others can notice the expressions using emotional cues [2]. These cues cover vocal or verbal cues, nonverbal behavior, or facial cues. Verbal cues are apparent from the word choices and how speaker speaks. Previous research has demonstrated the relationship between languages and the speaker's social and cultural values. One example of verbal communication is through text.

The classification process will be applied to the text to determine its emotion content. This classification process is a supervised-learning process where the computer learns from the labeled data or data that is part of the predictive feature. Six labels in a text that correspond to the basic emotions are used in the classification process [3]. Various techniques, including support vector machine (SVM), naïve Bayes, random forest, convolutional neural networks, have been used in numerous prior research [4]-[7]. The cross-lingual language model-robustly optimized bidirectional encoder representations from transformers approach (XLM-RoBERTa) model could improve the classification performance of a hate speech text in Indonesian to 89.52%, compared with the previous research using long short term memory (LSTM), which only reached 77.36% optimization [1].

The XLM-RoBERTa is a transformer model created by the coders at Facebook in 2019 [8]. This model is superior to the previous multilingual bidirectional encoder representations from transformers (mBERT) model which is only a multilingual transformer models [9]. The mBERT model is an evolution from the early BERT model which is a transformer model that can identify the attention in a text, thus predicting the following words based on the attention from each word [10].

Several previous research can also improve the model performance using transformers on the text classifying process [11], for example, the offensive texts classification, Internet Movie Database (IMDb) review, and sentiment analysis. The transformer models outperformed other traditional models. In this research, the model yielded a good result in classifying the text even though it still made some errors. The model classified the text using a single label or data with two labels, 1 and 0.

A similar experiment used the naïve Bayes model to extract emotional data from a text [12]. This model could produce an accuracy of 75.47%. It classified the text using the multinominal naïve Bayes to produce the probability numbers from all six basic emotions. The next step was using the threshold number to find the dominant emotion classes. In the experiment, the researcher trained the model using text with 2,187 data with 364 data per label on average. The drawback of this reserach is that the model only has limited vocabularies

p-ISSN 2301–4156 | e-ISSN 2460–5719

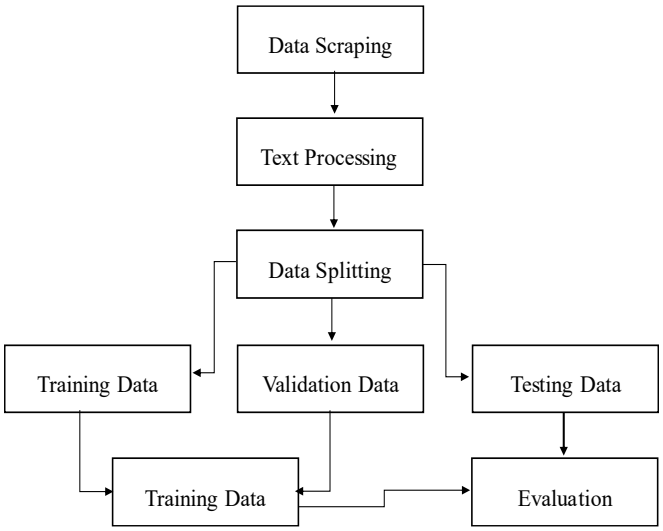

Figure 1. Overview of compound emotion classification.

and was not trained with larger data. The model performed well on data it had previously learned. However, it underperformed when it met data in sentences with unique words such as "*memakan*" (to eat) and "*makannya*" (therefore) that have different meanings, which, due to the word-stemming process, the machine interpreted as having the same meaning.

This proposed research aims to reduce the drawback from the previous research models [13], using the multilingual transformer model namely XLM-RoBERTa. The model was used to process the classification of an Indonesian text by learning the meaning and position of each word in a sentence. It will provide a significant improvement in classification accuracy. A traditional model usually only uses term frequency-inverse document frequency (TF-IDF) with different representations based on the frequency of a sentence or text. Several advantages of the XLM-RoBERTa model include 1) it can classify objects with a pattern or different words and still have correct predictions; 2) it has extensive Indonesian vocabularies, can continue training as well to speed up the learning process; and 3) it can perform multilabel classifications, where a word can have several labels to predict a compound emotion. Having a high accuracy in predicting a compound emotion can benefit multiple applications, such as facial recognition applications and social situation analysis.

The paper has a concise structure to highlight each section. Each section has a different purpose to explain the study. The introduction is to describe the research background, related research, and the purpose of this experiment. The evaluation of compound emotion describes the definition of basic and compound emotion in Indonesian. The methodology describes the steps within the experiment. And the result and discussion describe the result and analysis. This conclusion is to summarize the experiment and the conclusion.

II. COMPOUND EMOTION CLASSES

The six basic human emotions are happiness, sadness, anger, disgust, fear, and surprise [3]. Happiness is when one feels content about something. Sadness results from regret or failure to achieve what one wants. Anger is exhibited when one expresses their frustration about something. Disgust is an expression of a rejection. Fear arises when one feels worried. Last, surprise is when one encounters something unexpected.

TABLE I
EXAMPLE OF TOKENIZATION PROCESS

Text	Text Token	Token
Aku senang sekali bisa hidup di dunia ini tanpa ada nya halangan keuangan	[*'Aku', 'senang', 'sekali', 'bisa', 'hidup', 'di', 'dunia', 'ini', 'tanpa', 'adanya', 'halangan', 'keuangan'*]	[1, 101, 2, 18, 9, 10, 38, 49, 12, 34, 8, 103, 5]

The combination of these basic emotions creates a new compound emotion [14]. Identifying this compound emotion is possible by analyzing the muscle points on the face. Compound emotion can also happen when classifying Indonesian text. One of the latest compound emotions is a combination of happiness and sadness.

Compound emotion has several categories, namely 1) surprise happiness, 2) happiness disgust, 3) happiness sadness, 4) sadness fear, 5) sadness anger, 6) surprise sadness, 7) fear anger, 8) surprise fear, 9) fear disgust, 10) anger surprise, 11) anger disgust, 12) surprise disgust, 13) shock, 14) hatred, and 15) admiration. Identifying these compound emotions can work on several applications such as chat-bots, translation, text summarization, and more. And the advantage of this identification is to analyze the social situation.

III. METHODOLOGY

In general, the study has several steps such as data scraping, text processing, data splitting, the classifying model, and model evaluation. The overall steps in this study are seen in Figure 1. Data scraping is the data collection process. The next step was to process the data for the model to use. Then, the data would be divided into three parts; they were the data for training, data for validation that would partake in the training process, and the testing data used for the evaluation process. During the evaluation, the testing data would be tested against the training data.

A. DATA SCRAPING

This study used sentences in Indonesian as the dataset with texts with and without containing emotions. The dataset scraped the Twitter application programming interface (API) with the help of library tweeps on Phyton [15]-[17]. Twitter is a social media platform that allows its users to say something in a tweet, which is a short text with a maximum of 140 characters.

Twitter provides API access with an open-source license for research or study purposes by providing access to data such as a tweet, usernames, statuses, and others while protecting users' private data. Many researchers use this data to classify the sentiments [18]. In this study, linguists classify the dataset from Twitter with the six basic emotions, anger, happiness, sadness, disgust, surprise, and fear. The amount of data from this process was 40,000 Indonesian texts.

Later, the balance of these 40,000 texts in each emotion class was analyzed. The result from this analysis showed the dataset in label distribution lacks balance. Therefore, another scraping was conducted to obtain 25,100 texts in Indonesian. The overall 65,100 Indonesian texts went through a selection, aiming to balance the amount of each label. As a result, the final datasets consisted of 29,171 Indonesian texts.

The data with more than one label is referred to as multilabel data. The task is to classify the texts into one or several classes. One of the techniques to classify multilabel

p-ISSN 2301–4156 | e-ISSN 2460–5719

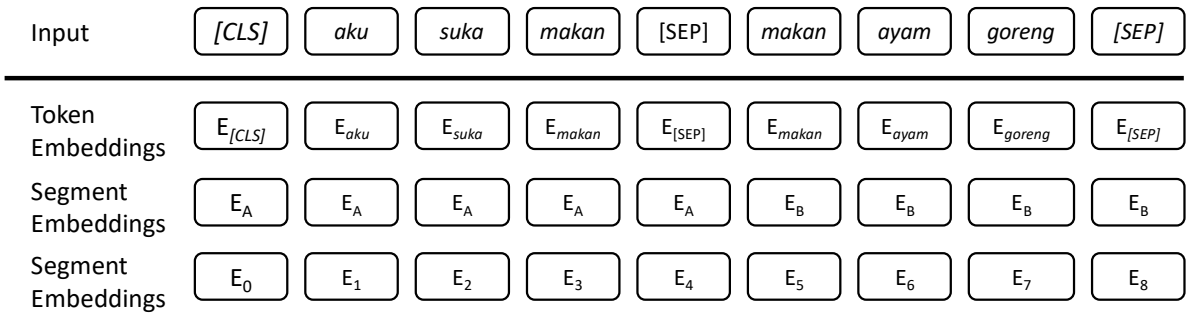

Figure 2. Ptokenization process with XLM-R model

texts is deep learning. There are, however, additional ways to perform the same classification. One of them is the binary classification on each existing label and rebuilding the sentence. There may be multiple emotion classes in the datasets used in this study. Therefore, each text could have more than one label. Finally, it sums up what multilable is.

B. TEXT PROCESSING

Text processing is necessary to help the model on classifying the texts [19]. The same study has stated that doing text processing can improve the classification by having better accuracy. In this study, the text processing included case folding and erasing several characters like emoji, entities, hashtags, punctuation, and overused spaces.

Case folding is the process of transforming the capital letter into a regular character. This process can help the tokenization process to prevent the creation of new tokens. Another example of preprocessing the text is filtering text such as erasing the emojis, entities in text like links, non-ASCII characters, punctuation, Twitter mention, hashtags, and retweet [20], as well as removing overused spaces.

The next step of text preprocessing was ro remove the data with fewer words or less than fourty characters. This text processing produced 29,171 data from 65,100 Indonesian texts.

C. TOKENIZATION

The tokenization process is a general process in natural language procession. It is a process to separate the sentence into several words and change them to a sequence of numbers for the next process [21]. An example of this tokenization process is presented in Table I. Each text needs to go through this process for a better understanding of each word without any other barriers [21].

The tokenization process using the XLM-RoBERTa model is different from the general tokenization process. The XLM-RoBERTa model is a transformer with an additional segment in the tokenization process. These additional segments are token embedding, segment embedding, and position embedding as in Figure 2. Token embedding is a process to transform the text into a specific number sequence to process them mathematically. The process started by splitting the texts into words notated as *E* (to represent embeddings) and having them as E_{aku}, E_{suka}, and so on. The *[CLS]* and *[SEP]* notation were the additional tokens in the sentence. The *[CLS]* acted as the initial token to signify the start of a sentence. Meanwhile, the *[SEP]* acted as the token to split or end the sentence. The next step was segment embeddings. This process transformed texts into different segments based on the text's delimiter. The E_A notation signifies the parts of the first sentence, E_B signifies the second sentence, and so on. This process is necessary to split the sentences into several sentences. Therefore, the machine can learn the correlation between each sentence on the next sentence prediction in the transformers. The position embedding step is to provide the numbers for each word based on the sequence in the text, therefore the model can learn the context in a sentence. Each word had the E annotation followed by its position in the text. The number sequence for the position started from Index 0.

The next text preprocessing steps were padding and truncating. Padding is the process to give additional data or 0 to the matrix tokens that have inadequate lengths, therefore they have a uniform matrix size. Meanwhile, truncation is the process of cutting the lengthy text to be the same as the other data.

Table I shows the tokenization process of splicing the words with spaces and turning them into a word list, which later turned into a number sequence based on the available word corpus. The example in Table I shows words transformed into numbers, such as "*aku*" to 1 and "*senang*" to 101. The number must remain consistent throughout the dataset.

D. DATASET

Data splitting is the process of dividing the dataset for the training process, validating data, and testing data. This division aims see the model's performance. Based on the analysis of the data scrapped from Twitter, there were 90% of training data, 5% of validating data, and 5% of testing data. Therefore, the overall testing data were 26,254 training data, 1,458 validating data, and 1,459 testing data.

E. XLM-ROBERTA

The XLM-RoBERTa is more accurate than its predecessor, XLM [8], with the accuracy improved by 14% on the cross-lingual natural language inference (XNLI) dataset, and a 2.4% FI-Score on the named entity recognition (NER). The researchers of this model trained the machine with 100 languages at large files to enhance the classification performance [8]. The XLM-RoBERTa transformer model could identify the meaning of a word using a self-attention mechanism which is an embedding method for words and use the position and relation with the next words to have different weights depending on its usage. The formula for the self-attention mechanism is expressed in (1).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \qquad (1)$$

The *Q* variable is the embedding vector for the word in a sentence, *K* is the embedding vector for the other words in the sentence, *V* is the vector that calculates the dot product from the word embedding with specific parameters, and *d* is the

p-ISSN 2301–4156 | e-ISSN 2460–5719

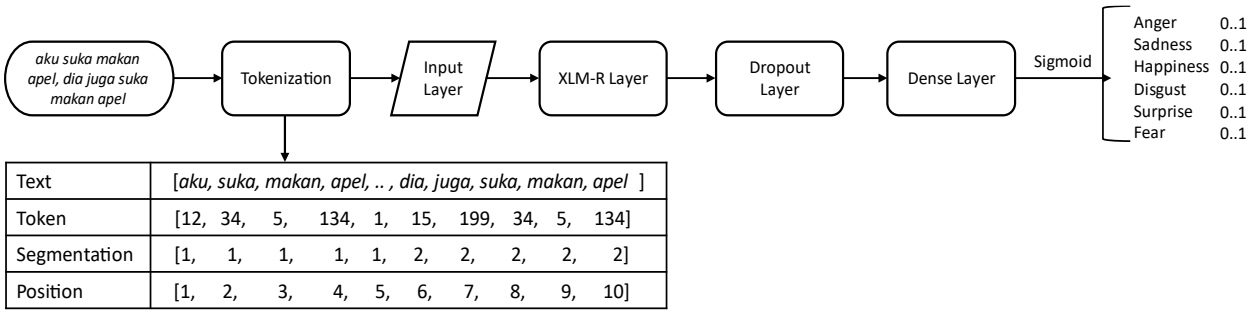

Figure 3. The architecture of the XLM-RoBERTa model.

dimension for the *Q*, *K*, and *V* vectors. Formula (2) uses the softmax function to calculate the result of $\left(\frac{QK^T}{\sqrt{d_k}}\right)$.

$$\sigma(\vec{z})_i = \frac{e^{zi}}{\sum_{j=1}^{K} e^{zj}} \quad (2)$$

The $\vec{z}$ notation is the input vector, e^{zi} is the exponent from the input vector, *K* is the total amount of classes, and the e^{zj} is the exponent of the output vector. The result of this calculation is a dynamic weight for each word with different embeddings at every position in the sentence.

To use the XLM-RoBERTa model in the classification process, the model architecture must be fine-tuned by adding multiple layers. These layers are the input layer, which is the first layer to input the data; the model layer, which is the primary layer in the model using the XLM-RoBERTa transformer model; the dropout layer, which is the layer that will deactivate several neurons inside the hidden layer; and the dense layer, which is the layer that has interconnected neurons.

In this study, the XLM-RoBERTa model architecture is in Figure 3. A text went through the tokenization process as input to the primary model using the input layer. The classification process was in the XLM-RoBERTa layer. The result must pass the dropout layer to ensure that several neurons from the training did not transform into an overfit, which is when the model records a good performance score on the training data but not on the validating data. Afterward, data moved to the dense layer, where the neuron extracts the output from the preceding layers by recalculating the sigmoid activation function so that the output becomes a number ranging from 0 to1. The function resulted in six numbers results that each representing the basic expression in a text, which is anger, sadness, happiness, surprise, disgust, and fear. The prediction result was the probability amount of each emotion class on each label. A combination of several emotion classes was the base for the compound emotion.

F. FINE-TUNING

Fine-tuning is the process to transform the model architecture by changing the input and output layers on a model so it can solve numerous problems such as classification, chatbot, and NER. This process includes adding several layers to the primary model and performing the hyper-turning on several layers. Therefore, this study used the *XLM-RoBERTa* with the fine-tuning process. There are not many changes to the model since the fine-tuning only adds the input and output layers based on necessity.

Adding these layers acted as the entry and exit points from the model. The model used the layer mask (LM) and next sentence prediction to provide the output which was a text with *[MASK]* token and text with a next sentence label to train the early model. It also used different input layer for the process in the question answering model by inputting texts that had start/end span, which was a text with sentence limitation that defined the question division and answers, and the output layer resulted in the output text with questions as the feature and answers as the question labels.

TABLE II
ACCURACY VALUE EACH ITERATION

Iteration	Training Accuracy	Validation Accuracy	Training loss value	Validation loss value
1	91.06%	95.48%	30.02%	16.27%
2	94.96%	95.48%	18.68%	15.40%
3	94.97%	95.50%	17.42%	14.25%
4	94.95%	95.50%	16.12%	13.46%
5	95.02%	95.56%	15.04%	13.34%

The input layer acted as the layer that received the data as numbers for the subsequent model's calculation process. The model was required to preprocess the received data so that they could be utilized by other layers. This layer is crucial because, without any input data, the subsequent layers will be useless. The XLM-RoBERTa layer is the layer from the primary model. This layer has multiple other layers that are also available in the primary model. A total of 559,896,432 layers were part of the XLM-RoBERTa layer.

The dropout layer is the layer that prevents a model from repeatedly learning the same thing and becoming overfitted [22]–[24]. A dropout occurs when several neurons are deactivated in a hidden layer. The parameter determines the percentage of a deactivated neuron. The dense layer was the layer used as the output layer in this architecture. This layer provided the result of the calculation done by the previous layers. The output was an array of six numbers to signify the text classes. Each number was between 0 and 1. If the number is higher, it signifies the model's confidence in labeling.

G. ACTIVATION FUNCTION

The activation function is the formula used to calculate the total input and weight of a neuron in order to determine if the neuron must remain active or inactive [25], [26]. One of the activation formulas in this study was the Sigmoid. This function is a nonlinear activation to transform the input to numbers within 0 to 1. the Sigmoid activation function uses the formula (3).

$$S(x) = \frac{1}{1+e^{-x}}. \quad (3)$$

The *S(x)* denotes the Sigmoid formula, *x* is the data input, and *e* is the exponent. The Sigmoid formula was used to create a model output consisting of numbers between 0 and 1. This

TABLE III
CLASSIFICATIONS OF COMPOUND EMOTION CLASSES

No.	Indonesian Texts	Anger	Sadness	Happiness	Disgust	Fear	Surprise	Actual Label	Prediction Result	Compound Emotion
1	*waktu smk pernah tuh jumpa sama cewe namanya dea lokasi tu di kolam renang yang family friendly lah harga masuknya awal nya saling lempar senyum aja agak salting sih pas si dea terus ngelitin terus pas akunya nunjukin gaya renang pake ngomongin lagi sama temannya*	0%	100%	84%	1%	0%	0%	Happiness, sadness	Happiness, sadness	Affected
2	*ini pak lieus lagi cari perhatian pak jokowi kita ketawain aja maaf pak bukan saya yang edit*	0%	0%	98%	0%	11%	0%	Happiness, Fear	Happiness, Fear	Excited
3	*saat konstantinopel takhluk ayasofya tetap dijaga bahkan saat perang pun rumah ibadah tak boleh diganggu perusakan rumah ibadah di kab sintang kalimantan barat sangat disayangkan saya turut mengecam dan berduka cita semoga pelakunya dapat diproses hukum dengan adil*	95%	0%	98%	0%	0%	0%	Anger, Sadness	Anger, Sadness	Dissappointed
4	*dia nih yang bikin umbrella academy tebel coba aja dia mau nurutin apa kata bapaknya buat gausah aneh aneh pasti umbrella academy tipis banget wkwk*	72%	0%	81%	0%	0%	2%	Anger, Happiness	Anger, Surprise	Resentful
5	*firasat alam jatuh dari jabatan dalam waktu dekat terperosok dalam lumpur yang menjijikan insya allah wallahu alam bishowab*	2%	100%	0%	11%	2%	0%	Sadness, Disgust	Sadness, Disgust	Depressed

activation function is perfect for a case involving biner classification.

Using the Sigmoid as an activation function, the model output that was consisted of six numbers between 0 to 1 determined the emotion classes in a text. The result from this classification process with a higher probability to have more than one label means the text has a compound emotion. While text with a lower probability for any label means it is neutral text.

IV. DISCUSSION AND RESULT

This section elaborates on the performance of the model in this research. In this study, the XLM-RoBERTa model boosted the accuracy to 95.56%. This model also performed better with

p-ISSN 2301–4156 | e-ISSN 2460–5719

unseen data, as the transformer model relied on probabilities based on the semantic (context) of the text rather than its frequency.

The result from this study showed the model suggested could predict the emotion classes from each Indonesian text. The prediction result showed that each Indonesian text could have one or more basic emotion classes that later combined to form a compound emotion. This part elaborates on further detail on the study of the training and prediction process.

A. TRAINING

The model only required five times training using the loss function binary cross-entropy (BCE) [27] using formula (4).

$$BCE = \frac{1}{n}\sum_{i=1}^{n} yi \cdot \log \hat{y}_i + (1-y) \cdot \log(1-\hat{y}_i). \quad (4)$$

The $\hat{y}_i$ variable is the predicted label, with y_i as the actual label and n as the label amount. The loss function functions to calculate the difference between the predicted label to the actual label. This function was used to classify binary data with only 1 and 0 as labels.

During the training process, the model had the validating data to measure its performance in recognizing data that it had never encountered during the training. The validating data was also used to help manage the hyperparameter by predicting the testing data and providing an evaluation as a reference on hyperparameter management, as an example adding a hidden layer when using another activation function or changing the model architecture. The validating data also can refer to data testing for the training process.

The study result in Table II demonstrates that the suggested model has very high accuracy. On the initial test, the accuracy for the validation and training data was 90% and it continued to grow on each test. Table II also exhibits the steady accuracy growth from each repetition, while the loss value for each training data and validating data steadily decreases for each iteration. A smaller loss value means the model performs better.

The analysis of this result showed that the model yielded a small loss value on the first repetition, which was 30.02% for the training data and 16.27% for the validation data. This results proves that the suggested model can do the classification process without too much training. Table II presents the loss value for each test's training and validating data.

A better description of the result for the five times repetition is as follows. On the first repetition, the training accuracy was 91.00%, and the validating accuracy was 95.46%. While on the second iteration, the accuracy rose by 3.90% to 94.96%, while the validating data value remained the same. During the training process, the accuracy continued to grow and on the fifth or final repetition, the training accuracy reached 95.02% and the validating accuracy was at 95.56%.

The result describing the change in the loss training value and the validating value is as follows. The loss value continued to decline steadily from the initial iteration. On the first test, the training loss value was at its highest, which was at 30.02%, and the validating loss was at 16.27%. For each iteration, the training and validating loss values continued to decline until they are at their lowest in the final test, which was 15.04% for the training loss and 13.34% for the validating value.

The result describes the basic emotion classes by classifying the Indonesian text is on Table III. Each sentence provides the primary emotion classes as the foundation to build a compound emotion class. As an example, the classification result on text number one provides the probability of the "sadness" basic emotion class at 100% and the happiness emotion class at 84%. The probability from these two emotion classes is the dominant value compared to other basic classes. Both of these dominant classes are the basis to build a compound emotion class that is "affected". The same goes for other classification results in the other Indonesian texts as shown in Table III.

Based on the classification in Table III, the compound facial expression was based on the probability value from two of the basic emotion classes with the highest probability values. Utilizing the facial action coding system (FACS), a complex facial emotion was implemented on a 3D animated face animation. It is a system to code the facial muscle activities that express emotion. Each one of the basic emotion class consists of several action unit (AU) in FACS that shows muscle movement. The combination of the related facial muscles can express a new compound emotion.

Upon implementing the 3D facial expression, the number between 0 and 1 represents the value of each *AU*. In this study, the machine implemented the value of each basic emotion using the probability from the classification process as in Table III.

The compound emotion on a 3D face occured by combining the *AU* from two basic emotion classes. The combination of each class's *AU* formed the facial muscles that showed a specific compound emotion. The probability value for each basic emotion class was the value for each AU. Therefore, if there is an imbalance in the probability value for each class (which tends to have an emotion class as the most dominant one), then the compound facial expression will also follow to express the specific emotion class. An example is text number 2 in Table III. The probability value of happiness was 98%, while the probability value for fear was 11%. Thus, the expressed facial expression leans toward happiness.

This study also showed that each dominant class from the classification process was the same as the actual emotion classes. It demosntrates the suggested model in this study is capable of expressing the compound emotion class which is the combination of dominant basic emotion classes. In addition, the model also has high accuracy in doing so.

B. PREDICTION

The prediction for the overall dataset of 29,171 texts had good accuracy with five times repetition. The XLM-RoBERTa model performed well on data it had never encountered due to the general knowledge possessed by this model. The performance still has room for improvement using bigger and better training data for the classification process.

The next experiment evaluated the testing data by distributing the training data, validating data, and testing data with different compositions. The experiment aims to understand the data composition to have the best performance. The first data composition had training data of 70%, validating data of 15%, and testing data of 15%. This composition resulted in a 94.42% accuracy and a loss value of 14.59%. These different data composition suggests the model provides better accuracy when it has much larger data. The best accuracy was at the composition of training data of 90%, validating data of 5%, and testing data of 5%, resulting in an accuracy of 95.56% and a loss value of 12.90%. This result exceeds the previous study that uses naïve Bayes and TF-IDF which was at 75.47% [13].

The naïve Bayes and TF-IDF model used the TF-IDF embedding process that counted the word frequency in a sentence or text. It is very different from the model XLM-RoBERTa in this study. Each word is represented based on the position and interword correlation in the text. Therefore, a word has a different representation based on the context, making the classification process easier.

The XLM-RoBERTa model can predict most of the text with less precise data and does not need detailed processing. It makes the XLM-RoBERTa model as far superior to other traditional models.

V. CONCLUSION

Based on several experiments, it is safe to state that the XLM-RoBERTa model can optimize the multi-label classification with only five training times. The optimized accuracy was 95.56%, with a 12.90% loss. This model was also able to predict the primary emotion in a text that had multiple emotion classes.

It is worth noting that the classification process still has room to grow by adding the amount of dataset and a balanced distribution for each class. The transformers model also can improve its performance by replacing or repairing the layer transformer architecture.

CONFLICT OF INTEREST

Hereby the author states that the data and analytical results in this paper have no conflict of interest with any party. The author also takes full responsibility if there is any issue with the data accuracy.

AUTHOR CONTRIBUTION

Conceptual, Aripin; methodology, Aripin; software, Steven Adi Santoso; validation, Aripin; formal analysis, Aripin, Hanny Haryanto; resource person, Aripin, Steven Adi Santoso; investigation, Aripin, Hanny Haryanto; data curation, Aripin, Steven Adi Santoso; writing—original drafting, Aripin, Steven Adi Santoso; writing—review and editing, Aripin, Hanny Haryanto; visualization, Aripin, Steven Adi Santoso; research administration, Aripin; funding acquisition, Aripin.

ACKNOWLEDGMENTS

The authors like to say our gratitude to the Directorate of Research and Public Service of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia that has helped fund this research through the grant program in the *Penelitian Terapan Unggulan Perguruan Tinggi* (PTUPT) based on the contract number 10/61031/PB/SP2H/AK.04/2022.

The authors also would like to express our gratitude to the rectorate of Dian Nuswantoro University for allowing us to use the Smart System Laboratory and helping us complete this research.

REFERENCES

[1] I.F. Putra and A. Purwarianti, "Improving Indonesian Text Classification Using Multilingual Language Model," *2020 7th Int. Conf. Adv. Inform.: Concepts, Theory, Appl. (ICAICTA)* 2020, pp. 1–5, doi: 10.1109/ICAICTA49861.2020.9429038.

[2] S. Du, Y. Tao, and A.M. Martinez, "Compound Facial Expressions of Emotion," *PNAS*, Vol. 111, No. 15, pp. E1454-E1462, Mar. 2014, doi: 10.1073/pnas.1322355111.

[3] P. Ekman, "An Argument for Basic Emotions," *Cogn., Emot.*, Vol. 6, No. 3–4, pp. 169-200, Jan. 2008, doi: 10.1080/02699939208411068.

[4] V. Dogra *et al*., "A Complete Process of Text Classification System Using State-of-the-Art NLP Models," *Comput. Intell., Neurosci.*, Vol. 2022, pp. 1–26, Jun. 2022, doi: 10.1155/2022/1883698.

[5] T.H. Saputro and A. Hermawan, "The Accuracy Improvement of Text Mining Classification on Hospital Review Through the Alteration in the Preprocessing Stage," *Int. J. Comput., Inf. Technol. (IJCIT)*, Vol. 10, No. 4, pp. 140–146, Jul. 2021, doi: 10.24203/ijcit.v10i4.138.

[6] W.-H. Khong, L.-K. Soon, and H.-N. Goh, "A Comparative Study of Statistical and Natural Language Processing Techniques for Sentiment Analysis," *J. Teknol.*, Vol. 77, No. 18, pp. 155–161, Nov. 2015, doi: 10.11113/jt.v77.6502.

[7] Aripin, H. Haryanto, and W. Agastya, "Synthesis of Compound Facial Expressions Based on Indonesian Sentences Using Multinomial Naïve Bayes Model and Dominance Threshold Equations," *Eng. Lett.*, Vol. 30, No. 1, pp. 1–10, Mar. 2022.

[8] A. Conneau *et l*., "Unsupervised Cross-lingual Representation Learning at Scale," *Proc. 58th Annu. Meeting Assoc. Comput. Linguist.*, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.

[9] H. Gonen, S. Ravfogel, Y. Elazar, and Y. Goldberg, "It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT," *Proc. Third BlackboxNLP Workshop Anal., Interpreting Neural Netw. NLP*, 2020, pp. 45–56, doi: 10.18653/v1/2020.blackboxnlp-1.5.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol.*, 2019, pp. 4171–4186, doi: 10.18653/v1/n19-1423.

[11] K. Taneja and J. Vashishtha, "Comparison of Transfer Learning and Traditional Machine Learning Approach for Text Classification," *2022 9th Int. Conf. Comput. Sustain. Glob. Development (INDIACom)*, 2022, pp. 195–200, doi: 10.23919/INDIACom54597.2022.9763279.

[12] Aripin, W. Agastya, and H. Haryanto, "Ekstraksi Emosi Majemuk Kalimat Bahasa Indonesia Menggunakan Convolution Neural Network," *J. Nas. Tek. Elekt., Teknol. Inf. (JNTETI)*, Vol. 10, No. 2, pp. 148–155, Mei 2021, doi:10.22146/jnteti.v10i2.1051.

[13] W. Agastya and Aripin, "Pemetaan Emosi Dominan pada Kalimat Majemuk Bahasa Indonesia Menggunakan Multinomial Naïve Bayes," *J. Nas. Tek. Elekt., Teknol. Inf. (JNTETI)*, Vol. 9, No. 2, pp. 171–179, Mei 2020, doi: 10.22146/jnteti.v9i2.157.

[14] S. Du and A.M. Martinez, "Compound Facial Expressions of Emotion: From Basic Research to Clinical Applications," *Dialogues Clin. Neurosci.*, Vol. 17, No. 4, pp. 443–455, 2015, doi: 10.31887/DCNS.2015.17.4/sdu.

[15] A. Wibowo and E. Winarko, "Paper Review: Data Mining Twitter," *Konf. Nas. Sist., Inform. (KNS&I)*, 2014, pp. 1–6.

[16] N. Azam, Jahiruddin, M. Abulaish, and N.A.H. Haldar, "Twitter Data Mining for Events Classification and Analysis," *2015 Second Int. Conf. Soft Comput., Mach. Intell. (ISCMI)*, 2015, pp. 79–83 doi: 10.1109/ISCMI.2015.33.

[17] R. Batool, A. Khattak, J. Hashmi, and S. Lee, "Precise Tweet Classification and Sentiment Analysis," *2013 IEEE/ACIS 12th Int. Conf. Comput., Inf. Sci. (ICIS)*, 2013, pp. 461–466. doi: 10.1109/ICIS.2013.6607883.

[18] S. Vosoughi, H. Zhou, and D. Roy, "Enhanced Twitter Sentiment Classification Using Contextual Information," *Proc. 6th Workshop Comput. Approaches Subjectivity Sentiment, Social Media Anal.*, 2016, pp. 16–24, doi: 10.18653/v1/W15-2904.

[19] M.D. Samad, N.D. Khounviengxay, and M.A. Witherow, "Effect of Text Processing Steps on Twitter Sentiment Classification using Word Embedding," 2020, *arXiv:2007.13027*.

[20] A.I. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification," *Int. J. Comput. Sci., Inf. Secur.*, Vol. 16, No. 6, pp. 22–32, Jun. 2018.

[21] J.J. Webster and C. Kit, "Tokenization as the Initial Phase in NLP," *Proc. 14th Conf. Comput. Linguist.*, 1992, Vol. 4, pp. 1106–1110, doi:10.3115/992424.992434.

[22] N. Srivastava *et al*., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, Vol. 15, No. 56, pp. 1929–1958, Jun. 2014.

[23] A. Labach, H. Salehinejad, and S. Valaee, "Survey of Dropout Methods for Deep Neural Networks," 2019, *arXiv:1904.13310*, doi: 10.48550/arXiv.1904.13310.

p-ISSN 2301–4156 | e-ISSN 2460–5719

[24] S. Cai *at al.*, "Effective and Efficient Dropout for Deep Convolutional Neural Networks," 2020, *arXiv:1904.03392*, doi: 10.48550/arXiv.1904.03392.

[25] S.R. Dubey, S.K. Singh, and B.B. Chaudhuri, "Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark," 2022, *arXiv:2109.14545*, doi: 10.48550/arXiv.2109.14545.

[26] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of Trends in Practice and Research for Deep Learning," 2018, *arXiv:1811.03378*, doi: 10.48550/arXiv.1811.03378.

[27] A.U. Ruby, P. Theerthagiri, I.J. Jacob, and Y. Vamsidhar, "Binary Cross Entropy with Deep Learning Technique for Image Classification," *Int. J. Adv. Trends Comput. Sci., Eng.*, Vol. 9, No. 4, pp. 5393–5397, Aug. 2020, doi: 10.30534/ijatcse/2020/ 175942020.