Indonesian Twitter Emotion Text Detection Using BERT

Vito Rihaldijiran
Departement of Computer Engineering
Faculty of Intelligent Electrical
and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia 60111
email: vito.18072@mhs.its.ac.id

Reza Fuad Rachmadi
Departement of Computer Engineering
Faculty of Intelligent Electrical
and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia 60111
email: fuad@its.ac.id

Arief Kurniawan
Departement of Computer Engineering
Faculty of Intelligent Electrical
and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia 60111
email: arifku@ee.its.ac.id

***Abstract*—Social Media is a Digital Platform that facilitates the use to socialize with each other, be it communicating or share content in the form of writing, photos, and videos. All content that shared will be open to the public. Social Media too has many impacts, ranging from positive and negative impacts negative. One example of its negative impact is the utterance of cian in behaving on social media, one of the containers is via Twitter. Indonesian people in using social media have bad behavior and have the possibility to spread hate on Twitter. Thus, the creation of research to detect emotions from existing tweets to minimize bad behavior in using Social Media. researcher- This tian utilizes BERT as the algorithm used to detect Text Emotions whether they are angry, happy, dih, and more automatically. Prior to detection by BERT, the text will enter the tokenization stage. Output from classification using BERT is the probability of whether the tweet text has ki the tendency to have emotions in accordance with what has been classified. The aim of this research is to make at a model that can be used to classify a text to have an emotional tendency according to the model have been trained to know the text has the possibility of have certain emotions. The results of this study are models that can detect twitter text emotions with an accuracy level of above 80%.**

***Keywords*—BERT, Emotion, Classification, Probability.**

I. INTRODUCTION

Social media is a medium to socialize with each other and is done online which allows humans to interact with each other without being limited by space and time [5]. Social media removes human boundaries to socialize, space and time limits, with this social media humans are allowed to communicate with each other wherever they are and at any time, no matter how far apart they are, and no matter day or night. Social media has a huge impact on our lives today. Someone who is originally "small" can instantly become big with social media, and vice versa "big" people in a second can become "small" with social media.

But along with the positive impact, social media also has a negative impact in the form of rampant Negative Emotion Tweet on social media. This is very worrying because social media in today's era can be said to be included as a primary human need. In addition, from year to year or even day to day, the amount of Negative Emotion Tweet on social media shows absolutely no signs of disappearing or being resolved. One of the causes of this is the rise of social media users who just follow along, either spreading or making the same upload without knowing the original intent/message/type of an upload because it is being discussed[2].

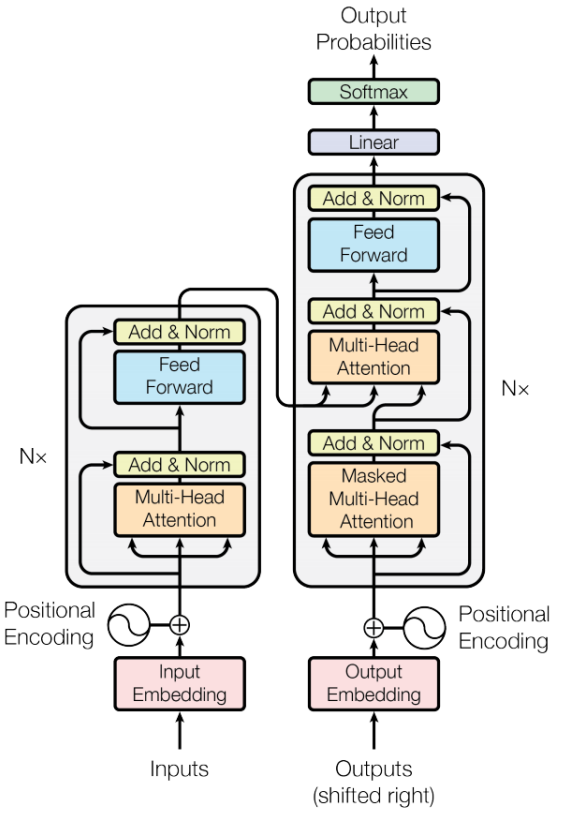
Based on a survey conducted by the company *Microsoft* entitled **Digital Civility Index**, which is a survey on the behavior of citizens who have been surveyed in using *Platform* Technology, Indonesia has a relatively low score compared to other Asean countries in behaving on Social Media with a ranking of 29 out of 32 countries. This survey was conducted annually in the last five years and had 16,000 respondents in 32 countries. Of course this is very bad for Indonesia because the survey reflects the behavior of its citizens in using social media[4].

The impact given by the rise of Negative Emotion Tweet will not only be felt by each individual, but the impact of this can also be felt by the international community. The existence of Negative Emotion Tweet directed at citizens of other countries can cause conflict and disrupt international relations between the two countries. Based on Mai Elsherief's research, the majority of Negative Emotion Tweet spreaders on social media use pseudonyms for their accounts in order to avoid knowing their real identities. In addition they generally target accounts that have a large number of followers or accounts that have a high level of activity[3].

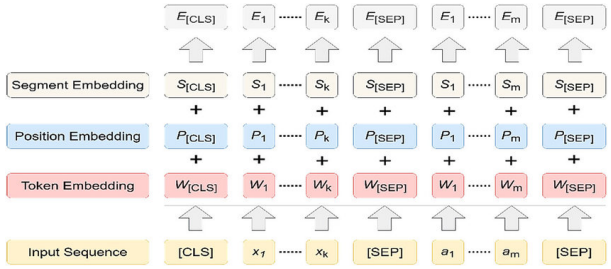
There have been several government efforts to overcome this problem by taking two approaches, namely preventive and repressive. The preventive approach can start from handling content related to Negative Emotion Tweet, while the repressive approach is more towards law enforcement related to legal efforts or processes.[1]. However, this step has not been effective because until now there are still many Negative Emotion Tweet on social media.

Neural Networks is a branch of machine learning that uses *neurons* like the structure of the human brain to process data and generate output. One of the relatively new *neural network* methods is *Bi-directional Encoder Representations from Transformers* or BERT for short. BERT is a method used to get context in an entered text, this makes BERT very suitable for performing tasks based on NLP (*Natural Language*

Processing). Even so, in Indonesia there are still not many implementations of BERT itself.



Gambar 1. Transformers Architecture That will be use in the Research



Gambar 2. BERT Architecture That will be use in the Research

II. PREVIOUS RESEARCH

In the research entitled *Emotion Detection and Analysis on Social Media*, the problem of detecting, classifying and quantifying text emotions in any form is discussed. This study uses English text collected from social media such as *Twitter*, which can provide useful information in various ways, especially opinion mining. Social media like *Twitter* and *Facebook* are full of emotions, feelings and opinions of people all over the world. However, analyzing and classifying text on the basis of emotion is a big challenge and can be considered as an advanced form of *Sentiment Analysis*. This *Paper* proposes a method for classifying text into six different Emotion-Categories: Happiness, Sadness, Fear, Anger, Shock, and Disgust. The model used is, the author uses two different approaches and combines them to effectively extract these emotions from the text. The first approach is based on *Natural Language Processing*, and uses some textual features such as *Emoticons*, word degrees and negation, Part Of Speech and other grammatical analysis. The second approach is based on the *Machine Learning* classification algorithm. We have also successfully devised a method to automate the creation of the *Training Set* itself, eliminating the need for manual annotation of large data sets. In addition, this research has succeeded in creating *Bag of Words* emotional words, along with their emotional intensity. On testing, it appears that this research model provides significant accuracy in classifying *tweet* taken from *Twitter*.

A similar study entitled *Emotion Detection Framework for Twitter Data Using Supervised Classifiers* has more or less the same goal. The research has the objective of emotion detection which involves text analysis. Humans show universal consistency in identifying emotions but show a great deal of variation between individuals in their abilities. This study has detected emotions for *Twitter* messages because they provide an ensemble of human emotions. This research has used machine learning algorithms or *Machine Learning* namely *Naive Bayes (NB)* and *k-nearest neighbor (KNN)* algorithms to detect the emotion of *Twitter* messages and then classify messages *Twitter* into four emotional categories. We also made a comparative study of two supervised machine learning algorithms; *Classifier* NB performs well when compared to *Classfier* KNN .

Another research entitled *Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD)* has a background problem, namely the proliferation of user-generated content on social media has made opinion mining a difficult task. As a *microblogging* platform, *Twitter* is used to gather views on products, trends and politics. Sentiment analysis is a technique used to analyze different people's attitudes, emotions and opinions towards anything, and can be done on *tweet* to analyze public opinion about news, policies, social movements and personalities. By using the *Machine Learning* model, opinion mining can be done without reading *tweet* manually. Their results can help governments and businesses launch policies, products and events. The *Seven Machine Learning* model is implemented for emotion recognition by classifying *tweet* as happy or unhappy. With an in-depth comparative performance analysis, it was observed that the proposed voting classifier *(LR-SGD)* with *TF-IDF* yielded the most optimal results with an accuracy of 79% and an F1 score of 81%. To further validate the stability of the proposed approach on two more datasets, one binary dataset and another *multiclass* dataset and achieve robust results.

Of the three studies, the *Supervised Learning* and *Natural Language Processing* methods were successful in classifying the *Emotion* type in *Twitter* Text. textbfBERT to apply to text detection *Emotion TWitter* in Indonesian. Based on the

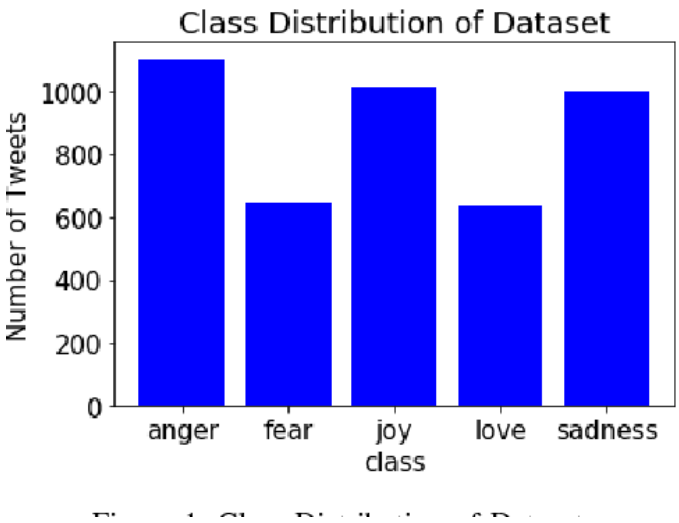
data and information presented in the previous section, it is necessary to conduct research to classify the types of *Emotion Twitter* Indonesia. In the future, the system can be developed into something useful, such as preventing Negative Emotion Tweet on social media.

III. DESIGN AND SYSTEM IMPLEMENTATION

This research is an application in the field of *Natural Language Processing* by using *Deep Learning* in order to detect emotions/*Emotion* from social media texts obtained from *Twitter* automatically. In the data training process, data is obtained from the source *Emotion Classification on Indonesian Twitter Dataset* [6] containing 4401 *tweet* which have been labeled with 5 emotions, namely *love*, *anger* , *sadness*, *joy*, *fear*. The personal data in the *Dataset* has been cleaned, for example, the *username* of each user has been replaced with the word *[USERNAME]*, the related link has been replaced with *[URL]*, and *Sensitive Number* has been replaced with *[SENSITIVE-NO]*.

A. Data Acquisition

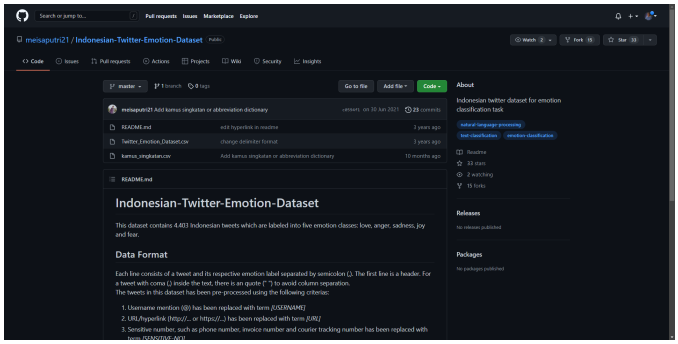
In the data acquisition stage, the data is taken from the *dataset* that has been created. Inside it is labeled which will later be classified using the **BERT** method. The related *Dataset* contains 4401 *tweet* in Indonesian which has 5 labels *emotion*, namely *love*, *anger*, *sadness*, *joy*, *fear*. For now, the entire *dataset* can be accessed and downloaded at the link https://github.com/meisaputri21/Indonesian-Twitter-Emotion-Dataset and related publications can be accessed via the link https://www.researchgate.net/publication/330674171_Emotion_Classification_on_Indonesian_Twitter_Dataset. The division of each number of *Tweet* in each *Emotion* is as follows:



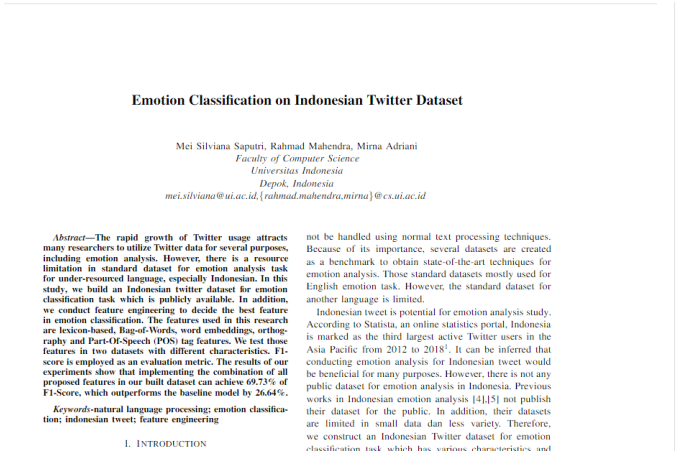
Gambar 3. The Division of each *Emotion* in the Dataset

B. Preprocessing

In the *Preprocessing* process, the data will be cleaned before the data is processed using the **BERT** model. There are two



Gambar 4. Retrieved source *Dataset*



Emotion Classification on Indonesian Twitter Dataset

Mei Silviana Saputri, Rahmad Mahendra, Mirna Adriani
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
mei.silviana@ui.ac.id,{rahmad.mahendra,mirna}@cs.ui.ac.id

***Abstract*—The rapid growth of Twitter usage attracts many researchers to utilize Twitter data for several purposes, including emotion analysis. However, there is a resource limitation in standard dataset for emotion analysis task for under-resourced language, especially Indonesian. In this study, we build an Indonesian twitter dataset for emotion classification task which is publicly available. In addition, we conduct feature engineering to decide the best feature in emotion classification. The features used in this research are lexicon-based, Bag-of-Words, word embeddings, orthography and Part-Of-Speech (POS) tag features. We test those features in two datasets with different characteristics. F1-score is employed as an evaluation metric. The results of our experiments show that implementing the combination of all proposed features in our built dataset can achieve 69.73% of F1-Score, which outperforms the baseline model by 26.64%.**

***Keywords*-natural language processing; emotion classification; indonesian tweet; feature engineering**

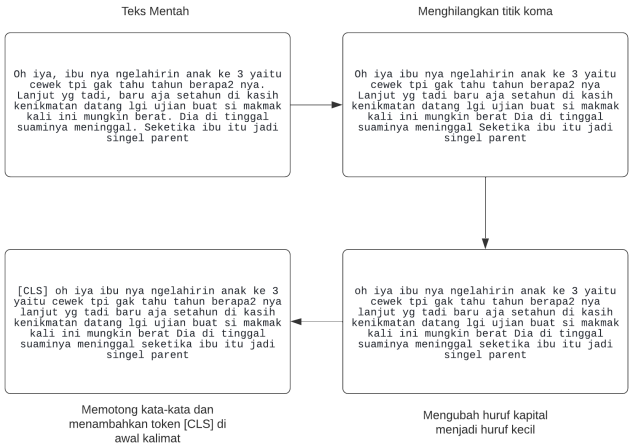
I. INTRODUCTION

not be handled using normal text processing techniques. Because of its importance, several datasets are created as a benchmark to obtain state-of-the-art techniques for emotion analysis. Those standard datasets mostly used for English emotion task. However, the standard dataset for another language is limited.

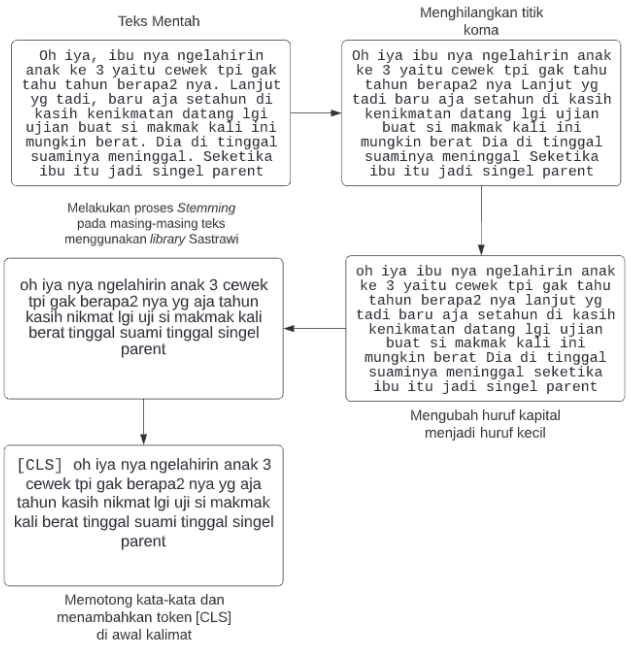
Indonesian tweet is potential for emotion analysis study. According to Statista, an online statistics portal, Indonesia is marked as the third largest active Twitter users in the Asia Pacific from 2012 to 2018[1]. It can be inferred that conducting emotion analysis for Indonesian tweet would be beneficial for many purposes. However, there is not any public dataset for emotion analysis in Indonesia. Previous works in Indonesian emotion analysis [4],[5] not publish their dataset for the public. In addition, their datasets are limited in small data dan less variety. Therefore, we construct an Indonesian Twitter dataset for emotion

Gambar 5. Publication Source of retrieved *Dataset*

methods *Preprocessing* that divide into two types *Dataset*. Besides wanting to test the performance of the model on **BERT**, the reason the *Preprocessing* method is divided into two is to find out whether the process of changing the basic structure of a sentence changes the accuracy of the model



Gambar 6. Preprocessing method with type *non-stemming*



Gambar 7. Preprocessing method with type *stemming*

or not. The first method is to use the *non-stemming* method, which is depicted in the image **??**. In this section, the contents of each *Feature* in *Dataset* will be changed several parts, but do not remove the meaning/value of the contents. The *Preprocessing* process includes removing periods and commas and changing the capital letters contained in each text.

Then the second method depicted in the image is the *Stemming* process. *Stemming* is a method to reduce inflectional/affixed words in a language to their basic form (*stem*). In this process, we use *Library* named **Sastrawi** which reduces the inflected words in Indonesian so that they can become sentences according to the basic form/close to the standard form. The process of the *Stemming* method is actually the same as the *non-stemming* method, starting with removing the semicolon, then changing the capital letters to lowercase letters. However, before adding the **[CLS]** token, the text will be processed using **Literary** to be converted into text with the basic form. The *Stemming* process ends by adding the **[CLS** token at the beginning of the sentence.

BERT has a maximum of 512 words or tokens that can be processed at one time, so text abbreviations must be shortened by taking the first 280 characters (according to the maximum number of characters when the user *Twitter* sends one *Tweet*), the last or a combination of the two forms. Chisun et al. found that retrieving text by taking the first 128 words in the middle and retrieving 382 words at the end resulted in better accuracy in some tasks [**?**].

In addition, the dataset will also be divided which initially amounted to 4401 which will be divided into 3 parts with the provisions that 70% will be used during the *training* process, 10% will be used for the validation process, and the remaining

Tabel I
DATASET SHARING DETAILS

Bagian	**Anger**	**Fear**	**Happy**	**Love**	**Sadness**	**Total Data**
Training	760	466	719	445	690	3080
Validasi	139	34	93	73	102	441
Pengujian	202	149	205	119	205	880
Total						**4401**

20% will be used at the time of testing.

For more details, please see the I table. From the table it can be seen that the division and total of the dataset are appropriate.

C. Training Process

Gambar 8. Training Method

At this stage, first the *Preprocessing* process will be carried out by **BERT Tokenizer**. This tokenizer is a process in which a word in a sentence becomes a token according to the *Word Embedding* provided by BERT. After that, the BERT Model will train the data from the *Preprocessing* Tokenizer from the *Word Embedding*. The *Output* of the BERT Model is the **[CLS]** token which will be entered into the text. This token will be entered into the classification algorithm to determine between *Feature* and its *Target*. The image **??** can be used as an explanation.

At the *Training* stage, adjustments were also made to settings such as *Batch*, *Learning Rate*, *Epoch*, *Hidden Dropout*, and *Epsilon*. *Batch* or *batch size* is *hyperparameter* which specifies the number of samples to be worked on before updating the internal model parameters. *Batch* in this case iterates over one or more samples and makes predictions.

The II table is further information on the *Hyperparameter* value used in this study.

Tabel II
BERT CONFIGURATION

Jenis Konfigurasi	**Keterangan**
epoch	3
batch size	8
learning rate	2e-5
hidden dropout	0.1
epsilon	1e-4 dan 1e-8

D. Testing Method

The testing process in this study is divided into 2 parts, the first part is carried out when the model has finished doing the *training* process but still has an unfinished *epoch* iteration. This process is called validation. The next part of the testing process is to do *test*. Just like during the validation process, the dataset used is completely different from the data used at the time of *training* and at the time of *validation*. From

this process, it can be concluded whether a model can be improved again by means of *re-training* and changing some parameters and configurations that have been set during the *training* process, or whether the model is considered good enough and will continue to the next process.

E. Performance Analysis

After carrying out the testing process, the Performance Analysis process is carried out. In this process, a calculation is made of how well the performance of the model that has been tested is calculated. Calculations in the Performance Analysis process use several references called *Metrics*. *Metrics* used is *Metrics* which is used for classification algorithms such as *Confusion Matrix*, *Accuracy Score*, *Precision*, *Recall*, and *F1 Score*. The set of *Metrics* can be calculated in one iteration using a feature named *Classification Report*.

IV. TESTING AND ANALYSIS

In this study, the results of the tests and analyzes carried out are presented in accordance with the system design that has been designed in the previous chapter. For testing and analysis, the *Dataset* used is the dataset from https://github.com/meisaputri21/Indonesian-Twitter-Emotion-Dataset. The test is carried out in several parts as follows:

nolistsep

1) Performance Testing based on Extracted Words
2) Performance Testing based on the BERT model used
3) Performance Testing based on the BERT model used
4) Performance Testing based on the *Training* Approach.

A. Model Performance Test Configuration

The current capacity of *BERT* can only process 512 words of Tokens, however, because the *Dataset* used has less text characteristics than normal text datasets, 280 words of Tokens are taken. These 280 words are taken because they match the maximum character *Twitter* if the user wants to send a *tweet*.

Performance testing on words that have been taken. The tokens that have been taken aim to determine the accuracy of the *BERT* model in the text. Because tokenized sentences have shorter words, there is no need to cut the word fragments because they are still in the capacity of *BERT* to process tokens, which is 512 words. From the tokenized *Dataset*, several *Parameters* were selected which were considered the most accurate among the other *Parameters* to be entered into the *BERT* model, including *epoch, batch , learning rate, hidden dropout, and epsilon.*

For some *Pretrained Model*, they generally use the same parameters, but there are conditions where different parameters must be used because the *Pretrained Model* directive specification requires that one of the models must use a parameter with a certain value, for example *epsilon* which requires using the values 1e-8.

The output of the model will be compared with the labels in the dataset, which will then be calculated to produce *confusion matrix*, *recall, precision, accuracy* and *f1-score* according to the formula described in the literature review chapter. summarized by a tool, namely *Classification Report*.

B. Model Performance Test Sharing Scheme

In the *Training* process, it is divided into 4 parts. The first is a test scheme using *Dataset* which does not go through the *Stemming* process or is called *non-stemming* and does not use *Freeze Parameter*, the second is a test scheme using *The dataset* which does not go through the *Stemming* process or is called *non-stemming* but uses *Freeze Parameter*, the third is *Dataset* which goes through the *Stemming* process and does not use *Freeze Parameter*, and the last one is *Dataset* which goes through the *Stemming* process and uses *Freeze Parameter*.

This scheme was chosen because during testing we wanted to know how accurate a model *BERT* would be if *Dataset* had different treatments as already mentioned. The *Stemming* process as described in the previous chapter is the process of converting affixed words into basic words. Examples such as holding back is holding, and reciprocating to be reciprocated. Here the test is carried out to find out whether the model will produce different accuracy because there are some parts of the affix that are omitted. *Stemming* is done considering that the *Dataset* used is a sentence taken from an Indonesian user *Twitter* where there are many sentences or words that are not standard.

Next is *Freeze Parameters*. *Freeze Parameter* is a condition where the layer or *Layer* is "frozen" or disabled for the purpose of *Training* so that the model created becomes simpler. The *Freeze Parameter* process is carried out whether there is a performance difference between the *BERT* model which has a more complex model and a simpler model. For the *Pretrained Model* used in the *Training* process this time there are 3 pieces sourced from *website* https://huggingface.co/, where these models are the models used devoted to training the Indonesian language *Twitter* text classification based on the resulting *Emotion*. The following are the models used in the **??** table below:

Tabel III
Pretrained Model USED IN THE TEST PROCESS

Model COde	**List of *Pretrained Model***
Model 1	akahana/indonesia-emotion-roberta
Model 2	indolem/indobertweet-base-uncased
Model 3	StevenLimcorn/indonesian-roberta-base-emotion-classifier

C. Testing Dataset Non-Stemming and Without Freeze Parameter

Testing with the first schema is done with *Dataset* which does not go through the *Stemming* process and without using *Freeze Parameter*. This test does not go through any *Preprocessing* which automatically makes it go straight through the *Training* process. From the three models used, the model performance is as follows:

From the experiments carried out, there were several results in the form of the *Metrics* value of the three *Pretrained Model* used. Experiments using the *akahana/indonesia-emotion-roberta* model showed very good results with an accuracy of 90% and the values *Precision*, *Recall*, and *F1-score* were not

		Model 1	Model 2	Model 3
Precision	Sadness	88%	64%	83%
	Anger	94%	87%	92%
	Love	90%	82%	90%
	Fear	86%	70%	81%
	Happy	91%	84%	94%
Recall	Sadness	86%	68%	84%
	Anger	92%	79%	90%
	Love	95%	81%	93%
	Fear	88%	86%	88%
	Happy	91%	79%	89%
F1-Score	Sadness	87%	66%	83%
	Anger	93%	83%	91%
	Love	93%	82%	91%
	Fear	87%	77%	84%
	Happy	91%	81%	91%
Accuracy		90%	78%	88%

Gambar 9. Experimental results of the *Non-Stemming* method and without the *Freeze Parameter* using 3 *Pretrained Model*

		Model 1	Model 2	Model 3
Precision	Sadness	85%	64%	82%
	Anger	95%	86%	92%
	Love	91%	82%	91%
	Fear	86%	71%	81%
	Happy	91%	83%	94%
Recall	Sadness	88%	69%	84%
	Anger	90%	79%	89%
	Love	95%	82%	93%
	Fear	89%	85%	88%
	Happy	90%	76%	88%
F1-Score	Sadness	86%	66%	83%
	Anger	92%	82%	90%
	Love	93%	82%	92%
	Fear	87%	77%	84%
	Happy	91%	79%	91%
Accuracy		90%	77%	88%

Gambar 10. Experiment results using *Non-Stemming* method and using *Freeze Parameter* using 3 *Pretrained Model*

there is less than 80% where this result is classified as very good, but in calculating the difference in *Loss* between *Evaluation* and *Training*, the distance is quite far from *Training loss* always decreases with with increasing *epoch*, but *Evaluation Loss* tends to increase as *epoch* increases.

Experiments using the *indolem/indobertweet-base-uncased* model showed results with an accuracy of 78% where each value between *Precision*, *Recall*, and *F1-Score* showed values that are varied but are in the range of 64% the lowest and 87% the highest. The difference in *Loss* in this model is that *Training Loss* always decreases with increasing *epoch* but for *Evaluation Loss* the value from one *epoch* to another *epoch* tends to stagnate .

Experiments using the *StevenLimcorn/indonesian-roberta-base-emotion-classifier* model more or less produce values that are not much different from those used with the *akahana/indonesia-emotion-roberta* model where the accuracy is 88%. The value between *Precision*, *Recall*, and *F1-Score* also shows good results where the lowest value is at 81%, namely the *Precision* value of the *label Fear* and the highest value is in *Precision* of the label *Happy* with a value of 94%. Meanwhile, the difference in loss is not much different from the results produced by the *indolem/indobertweet-base-uncased* model where the difference in *Loss* in this model is that *Training Loss* always decreases with increasing *epoch* but for *Evaluation Loss* the value from one *epoch* to another *epoch* tends to stagnate.

D. Testing Dataset Non-Stemming and using Freeze Parameter

Testing with the second scheme is carried out with *Dataset* which does not go through the *Stemming* process and uses *Freeze Parameter*. In this Test some *Layer* of the *BERT* Model is disabled in order to produce a simpler model. From the three models used, the model performance is as follows:

Experiments using the *Non-Stemming* Dataset and using the *Freeze Parameter* obtained some results from the three *Pretrained Model* used. First, by using the *akahana/indonesia-emotion-roberta* model, the accuracy results are 90% where the values *Precision*, *Recall*, and *F1-Score* are not much different from the results obtained. obtained with Dataset *Non-Stemming* which does not use *Freeze Parameter* with a range of 85% to 95%. for the amount of difference *loss* between *Evaluation* and *Training*, the distance is quite large with *Training loss* always decreasing as *Epoch* increases, but *Evaluation Loss* tends to increases as *epoch* increases.

Experiments using *indolem/indobertweet-base-uncased* showed more or less the same as the previous experiment, using the *Non-Stemming* Dataset and without using *Freeze Parameter* with an accuracy of 78%. The values between *Precision*, *Recall*, and *F1-Score* show results in the range of 64% to 85% and the *Loss* difference in this model is *Training Loss* always decreases with increasing *epoch* but for *Evaluation Loss* the value from one *epoch* to another *epoch* tends to stagnate.

Experiments using the *StevenLimcorn/indonesian-roberta-base-emotion-classifier* model more or less produce values that are not much different from those used with the *akahana/indonesia-emotion-roberta* model where the accuracy is 88%. The value between *Precision*, *Recall*, and *F1-Score* also shows good results where the lowest value is at 81%, namely the *Precision* value of the *label Fear* and the highest value is in *Precision* of the label *Happy* with a value of 94%. Meanwhile, the difference in loss is not much different from the results produced by the *indolem/indobertweet-base-uncased* model where the difference in *Loss* in this model is that *Training Loss* always decreases with increasing *epoch* but for *Evaluation Loss* the value from one *epoch* to another *epoch* tends to stagnate.

E. Testing Dataset Stemming and without using Freeze Parameter

Testing with the third scheme is carried out with *Dataset* which goes through the *Stemming* process and without using *Freeze Parameter*. In this test *Dataset* goes through the *Stemming* process and as a result the affixed words are removed.

From the three models used, the model performance is as follows:

		Model 1	Model 2	Model 3
Precision	Sadness	51%	53%	61%
	Anger	80%	84%	81%
	Love	81%	82%	77%
	Fear	64%	68%	71%
	Happy	74%	78%	82%
Recall	Sadness	57%	58%	61%
	Anger	69%	73%	78%
	Love	78%	79%	83%
	Fear	79%	85%	80%
	Happy	73%	76%	76%
F1-Score	Sadness	54%	55%	61%
	Anger	74%	78%	79%
	Love	79%	81%	80%
	Fear	70%	76%	75%
	Happy	73%	77%	79%
Accuracy		70%	73%	75%

Gambar 11. Experimental results of *Stemming* method and without using *Freeze Parameter* using 3 *Pretrained Model*

In the experiment using the *Stemming* dataset and without using the *Freeze Parameter*, several results were obtained from the three *Pretrained Model* used. The first using the *akahana/indonesia-emotion-roberta* model, the results are much different from the 2 datasets that have been tested previously with an accuracy of 70% where this result is very far from the first two experiments using the same model with an accuracy of 90 %. Then for the values of *Precision*, *Recall*, and *F1-Score* are also classified as getting *Range* which is quite small with a range of 51% to 81%. Then for the difference in *Loss* itself, *Training Loss* always decreases with increasing *epoch* and *Eval Loss* also shows the same result, but for *Eval Loss* the decrease is not too drastic like *Training Loss*.

The second experiment using *Pretrained Model* from *indolem/indobertweet-base-uncased* and the results obtained with an accuracy of 73%. There is also a decrease from the experiment with the *Non-Stemming* dataset, but it is not significant considering that it only decreases by 4% to 5%. Then for the range of values *Metrics Precision*, *Recall*, and *F1-Score*, the results are almost the same as experiments with the previous 2 dataset types with a value range of 53% to with 85%. For the difference in *Loss* from this experiment, *Training Loss* always decreases with increasing *Epoch* while *Eval Loss* as *Epoch* increases, there is a moment where the value increases but after back down.

The third experiment using *Pretrained Model* from *StevenLimcorn/indonesian-roberta-base-emotion-classifier* and the results obtained with an accuracy of 75%. There is also a significant decrease from experiments with the *Non-Stemming* dataset considering that experiments with the previous 2 datasets yielded an accuracy of 88%. Then for the range of values *Metrics Precision*, *Recall*, and *F1-Score* shows the results with a value range of 61% to 83%. For the difference in *Loss* from this experiment, *Training Loss* always decreases with increasing *Epoch* while *Eval Loss* as *Epoch* increases, there is a moment where the value increases but after back down.

F. Testing Dataset Stemming and using Freeze Parameter

Testing with the third scheme is carried out with *Dataset* which goes through the *Stemming* process and without using *Freeze Parameter*. In this test *Dataset* goes through the *Stemming* process and as a result the affixes are removed and *Freeze Parameter* is used to disable some *Layer* so that the model used is simpler. From the three models used, the model performance is as follows:

		Model 1	Model 2	Model 3
Precision	Sadness	0%	55%	57%
	Anger	97%	83%	78%
	Love	0%	84%	78%
	Fear	48%	64%	71%
	Happy	0%	78%	78%
Recall	Sadness	0%	57%	58%
	Anger	27%	74%	75%
	Love	0%	79%	82%
	Fear	82%	85%	75%
	Happy	0%	76%	76%
F1-Score	Sadness	0%	56%	57%
	Anger	42%	78%	77%
	Love	0%	82%	80%
	Fear	60%	73%	73%
	Happy	0%	77%	77%
Accuracy		32%	73%	72%

Gambar 12. Experiment results using *Stemming* method and using *Freeze Parameter* using 3 *Pretrained Model*

The next experiment is using the *Stemming* dataset and using the *Freeze Parameter*. The first experiment using *Pretraine Model* from *akahana/indonesia-emotion-roberta* with poor results with the previous 3 types of datasets with an accuracy of 32%. Then the values *Precision*, *Recall*, and *F1-Score* show values that are smaller than the three previous dataset types with a range of 0% to 97%. For the difference *Loss*, there is no allusion between *Evaluation* and *training loss* where *Training Loss* is always greater than *Eval Loss*.

The second experiment using *Pretrained Model* from *indolem/indobertweet-base-uncased* and the results obtained with an accuracy of 73%. There is also a decrease from the experiment with the *Non-Stemming* dataset, but it is not significant considering that it only decreases by 4% to 5%. Then for the range of values *Metrics Precision*, *Recall*, and *F1-Score*, the results are almost the same as experiments with the previous 2 dataset types with a value range of 55% to with 85%. For the difference in *Loss* from this experiment, *Training Loss* always decreases with increasing *Epoch* while *Eval Loss* always stagnates as *Epoch* increases.

The third experiment using *Pretrained Model* from *StevenLimcorn/indonesian-roberta-base-emotion-classifier* and the results obtained with an accuracy of 72%. There is also a significant decrease from experiments with the *Non-Stemming* dataset considering that experiments with the previous 3 datasets yielded an accuracy of 88% for *Non Stemming* and and 75% for *Stemming* . Then for the range

of values *Metrics Precision*, *Recall*, and *F1-Score* shows the results with a value range of 57% to 82%. For the difference in *Loss* from this experiment, *Training Loss* always decreases with increasing *Epoch* while *Eval Loss* as *Epoch* increases, there is a moment where the value increases but after back down.

V. CONCLUSION

From all the research that has been done, some conclusions can be drawn as follows:

nolistsep

1) the use of the *Stemming* feature always results in poorer model performance when applied to a dataset containing non-standard words. This is because the *Stemming* feature does not work optimally if it cuts affixed words in non-standard sentences which results in a lack of performance from the model.
2) Using *Freeze Parameter* has very little effect on model performance. When compared with models that do not use *Freeze Parameter*, the difference in accuracy is only 1%.
3) BERT is a very easy *overfit* model, so setting parameters such as the *freeze* parameter will make the model more *general* but have a few percent lower accuracy.
4) BERT is a pretty good and effective model for classifying text which has been proven to produce an accuracy performance level of 90% and only takes a total of 5 minutes for 1 *epoch* only.
5) the performance of the model using a dataset that contains many non-standard sentences has the best performance at 90% by using *Pretrained Model* named *akahana/indonesia-emotion-roberta* and without going through the *Stemming* and using *Freeze Parameters*

PUSTAKA

[1] (2019). Pemerintah lakukan dua pendekatan tangani konten ujaran kebencian. diakses April 2022.

[2] Ash-Shidiq, M. A. dan Pratama, A. R. (2021). Ujaran kebencian di kalangan pengguna media sosial di indonesia: Agama dan pandangan politik. pages 1–2. Universitas Islam Indonesia.

[3] M, E., Nilizadeh, S., Nguyen, D., G, V., dan E, B. (2018). Peer to peer hate : Hate speech instigators and their targets. pages 52–61.

[4] Microsoft (2020). Microsoft study reveals improvement in digital civility across asia-pacific during pandemic. diakses April 2022.

[5] Rustian, R. S. (2012). Apa itu sosial media. diakses April 2022.

[6] Saputri, M., Mahendra, R., dan Adriani, M. (2018). Emotion classification on indonesian twitter dataset.