CHAPTER 1

INDTRODUCTION

This chapter includes the following subtopics, namely: (1) Rationale; (2) Theoretical Framework; (3) Conceptual Framework/Paradigm; (4) Statement of the problem; (5) Hypothesis (Optional); (6) Assumption (Optional); (7) Scope and Delimitation; and (8) Importance of the study.

1.1 Rationale

The development of technology including social media in Indonesia gives rise to many benefits. Such as making work easier and social interaction in society. With ease of access and use, social media is a place for information dissemination and exchange. In Indonesia, social media is the most widely used platform for exchanging information and make some opinions to express their thoughts through social media such as Instagram and Twitter. Based on research conducted by Wearesocial Hootsuite released in January 2020, social media users in Indonesia grow up to 160 million with the use of the Instagram platform at 79% of the population [1].

Despite of all benefits from social media such as free, fast to spread information, easy to use, social media give some risk and vulnerable. The social media has a huge influence for spreading some information, whether good information or bad, social media also can be platform to spread hate opinion or behavior towards another person or group causes abuse on social media which is a crime. It also proved that hate crime is getting increased from one day to another.

This phenomenon has caused unrest in Indonesia, there are even regulations governing the use of social media, namely the Law and circulars on hate speech through Article 27 paragraph (3) of the UU ITE, Article 45 paragraph (1) of the UU ITE and Circular (SE) Kapolri number SE / 6 / X / 2015 [2], with the aim of the community wisely using technology, especially social media.

Several cases of the spreading of hate speech, among others, during the presidential election in 2019. Groups or individuals consistently spread hate speech on social media. In her research entitled " Social Media and Democracy in the Information Age in 2014," Devie said that almost all political actors are competing for influence through the use of social media for political purposes. They scramble to influence society through social media channels [3].

Hate speech classification also is a challenging issue. There’s no exact definition of hate speech, and that is one of many issue for classifying hate speech is challenging, because of no exact definition of hate speech and it depend on the context.

Studies that related to identifying hate speech has been done by Saurabh et.al present a sarcasm detection using Recurrent Neural Network. The model implemented with Tensorflow

and python programming language. The author trained 2000 tweets, and used 2 recurrent neural network layers, with each layer consist of 256 LSTM cells. Increasing the neural network layers increases the computations and decreasing it, affect the accuracy. So deciding the appropriate number of layers of neural networks plays an important roles. On other research, research by I Ketut Gede Darma Putra, et.al, using Convolutional Neural Network for classifying hate speech on Twitter using Indonesian language. Author using TF-IDF as term weighting method for feature extraction.

Classifying hate speech requires a special approach and care. Hate speech is a form of discrimination and hatred that can have a negative impact on the individual or group it is aimed at, so it is important to identify and deal with it appropriately. In the process of classifying hate speech, it is important to consider the context and implications of the action. This requires a careful and responsible approach, as it can have legal and social implications. Some things to consider include the Dataset. The dataset that used for classifying hate speech must be a fair representation of the different types of hate speech and must have non-hate speech data as a reference and also model selection. The model chosen must have the ability to understand the context and implications of hate speech. Preprocessing must take into account that hate speech can take various forms, such as sarcasm, irony, or euphemisms. Based on previous research [6], they suspect that there is any offensive term that could be feature.

Therefore, the detection of hate speech on social media is carried out to developing technology who can detect and filtering hate speech that can take various forms. Identifying hate speech on social media is one way to measure the potential and level of hate speech on social media content. The purpose of this thesis is to identifying hate speech on social media users based on specific aspect because hate speech will have negative impact on individuals or other social media users, so that social media users are wiser in using social media.

This report is containing chapter 1 for introduction, chapter 2 is related studies that related on this research, chapter 3 is for methodology, and result also conclusion is occur on last chapter which is chapter 4.

1.2 Theoretical Framework

In the Build model referred to in Figure 2, the training process will build one of Recurrent Neural Network (RNN) architectures, there is Bidirectional LSTM (long short term memory) model. BiLSTM has a memory cell consisting of 4 main components: input, word embedding, recurring connection, forget gate, activation function and output.

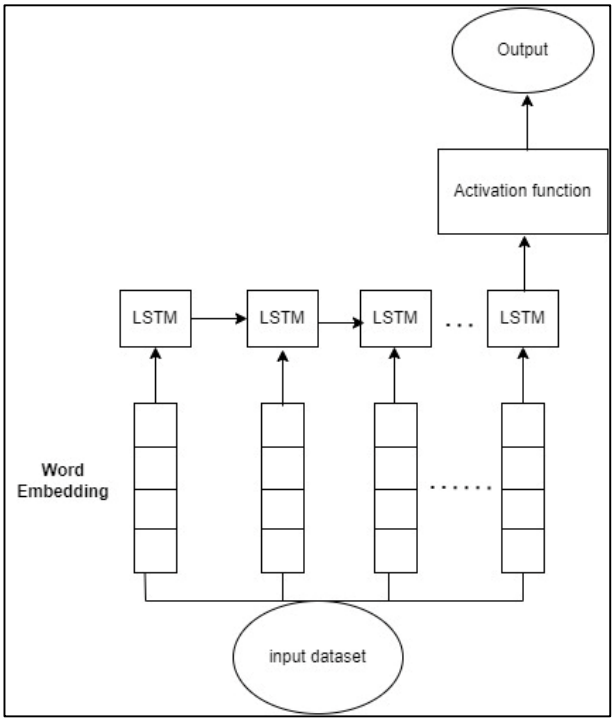

Figure 1 LSTM process

Figure 1 show building an LSTM network is to identify the unnecessary information that will be removed from the cell at that step using the sigmoid function. For the activation function, in this research using Sigmoid. Sigmoid is activation function that have S shape and have dropout between 0 and 1.

However, the LSTM method like figure 1 show, has a limitation that is only using one input in the form of past information to learn information from one direction, that is the forward direction [14] so, another architecture of LSTM was introduced, namely Bidirectional Long Short-Term Memory (Bi-LSTM) which has 2 inputs in the form of forward and backward.

In the hate speech detection task, BiLSTM can explore the features needed to identify readings that have hate speech. In this case, BiLSTM can help identify changes in sentences that can change the meaning of sentences in totality. That way, the use of BiLSTM in hate speech detection can provide better results compared to other methods.

1.3 Conceptual Framework/Paradigm

Based on previous studies [6], result of identifying hate speech it could be consist of offensive terms, that sometimes cannot have detected. So that is multi label classification problem. In this research we try to use multi-label dataset in order to detect hate speech for aspect itself.

Table 1 Example of sentence and output

Sentence	Output
bangsat bajingan ini sudah diciduk mampus	Hate Speech Individual
sekarang kita berada di rejim partai komunis indonesia lengserkan jokowi pemimpin tukang mengkriminalisasi ulama	Hate Speech Group

On table 1, show the example of sentence that become input of purposed model, which on the output it could be classified into aspect hate speech. For example on second sentence, because the sentence consist of group term, it classified into “hate speech group”.

1.4 Statement of the Problem

Hate speech classification sometimes hard to handle and detect the hate speech because of hate speech are many variations. So the problem of this research is how to build model that classifying multi-label hate speech and solve the problem from previous research that one of their embedding cannot achieve good result because of the dataset is small.

1.5 Objective and Hypotheses

The purpose of this study is to propose method to detect hate speech on Social media on several aspect in comments and measure its accuracy and compare it with previous studies [6],[15]. On Previous research [15], regarding sarcasm detection was successful using the RNN-BiLSTM method. So the assumption of this research is that this method can be used to detect hate speech based on each aspect.

Also using different method for feature extraction to avoid miss classifying. On previous research [15] that using TF – IDF for measure word frequency text can be classified as hate speech from the frequency of use of insulting words so more tweets are classified as True Negative than True Positive. This high True Negative result indicates that it is easier to classify hate speech tweets than non-hate speech because sometimes there are misclassification

1.6 Assumption

Using purposed model which is Recurrent Neural Network with architecture BiLSTM and fasttext word embedding can be more better than previous study and more better to detect hate speech in social media. And also grouping the dataset assumed that reduce miss classifying on hate speech detection.

1.7 Scope and Delimitation

1. Data is taken on two social media, namely Instagram and twitter.
2. Even on dataset, label of data is many, in this experiment the label used in this dataset is only Abusive, Other, and group.

1.8 Significance of the Study

The technology developed in this research can assist law enforcement officials based on [2] in identifying and processing acts of discrimination and hate speech, because the model can learn the features needed to better recognize hate speech in text. This better accuracy can help keep people safe online and protect vulnerable individuals from discrimination. Another significance of this study is to develop effective methods to deal with multi-label hate speech: Hate speech often has several labels, such as Individual hate speech, Abusive hate speech, Group hate speech, and so on. By using BiLSTM, the model can learn patterns and relationships between these labels, so that it can recognize and categorize hate speech more effectively.