Chapter 1 INTRODUCTION

This chapter discusses the background or motivation of the research, problem statement and research question, objective, hypothesis, the research methodology, and contribution of the research.

1.1 Background

At this moment, hatred to a person or group can be expressed in the media that can cover a wide range of audiences. This method aims to seek a similar opinion from numerous people and make them usually to hating people or groups who become the target of hatred. A hatred speech towards a person or group is known as hate speech. Hate speech is a form of communication to express hatred by writing things such as inciting, insulting, disparaging or demeaning characteristics of a person or group based on their religions, beliefs, races, nationality, skin colors, ethnicity, genders, sexual orientations, and disability that can trigger violence, discrimination and social conflicts [1], [2], [3].

In the article [4], the author mentioned that the hate speech issue often found in Indonesia is mostly dealing with politics. It is a global phenomenon that does not only occur in Indonesia. In the United States of America in 2016, the presidential election won by Donald Trump was one of the proofs [4]. While in 2018, Indonesia has been colored by hatred cases that happened in the local leader election. In 2019, there were several cases of hatred committed by political party partisans by spreading hoax and negative framing of presidential and vice-presidential candidates besides their candidate. If hate speech actions in 2018 and 2019 happened again, the public would be confused, and worse, several opinions without evidence would be built that made hatred towards a particular group arise again [5].

Hate speech actors in Indonesia often use a social network to spread their hatred words. That social network is Instagram. Instagram is used by the actors because it has been widely used by many people to share photos or videos. Based on Statista[6], Instagram users in Indonesia are ranked fourth in the world with an estimated number of 59 million users. From millions of users on Instagram, there is a possibility of several users become an actor of hate speech by sharing either photo or video using hatred words in the caption of the post. The hatred

post that the user share is an opinion former to lead the opinion of people who agree with the hate speech. They who agree with the hate speech actor or perpetrator will comment on a negative sentence that exceeds the first post negativity. In 2017, Yuliana, M. E. et al. found that 60% of hate speech occurred in the comments field of a post [7]. Moreover, the negative and hatred comment was not just going to stop at argument contest between the pros and cons in the comment field, but it might be a real threat if it was not detected quickly.

Research [1] has detected hate speech in the Indonesian language using the Random Forest Decision Tree and the union of word unigram and bigram in the Twitter dataset. However, their model falsely detects the name of a person as hate speech since the proportion of hate text in their dataset is dominated by the name of a person. The sequence of occurrences and context of several words in sentences is not considered in their research. Their research was then continued by the research [2] but the dataset was different from the previous research [1]. In their research [2], Indonesian language Instagram comments were used for classification using Fasttext as a classifier. They obtained a smaller number of accuracy but the result of the model [2] when using the [1] dataset is even smaller than it.

The researcher in [2] mentioned that their result was small because of the Fasttext was not suitable for a small dataset, occurrences of offensive terms were not considered, and their annotator was subjective when labeling the data. From the problem mentioned earlier, the author proposes an approach combining the rule of hate speech, word2vec with Skip-gram model, methods for imbalanced data, and Convolutional Neural Network (CNN) to detect hate speech in the Indonesian language. Today, only a few studies have discussed the research on hate speech detection in Indonesia language using a combination rule of hate speech, word2vec, and deep learning method.

1.2 Problem Statement and Research Question

Several studies have conducted hate speech text detection in various languages, such as the Indonesian language [1], [2]; and English [8][9]. In research [1], researchers used the combination of Bag of Words (BOW) model, word unigram, and Random Forest Decision Tree (RFDT) in detecting hate

speech from Indonesian language Twitter data with the accuracy of 81.7%. While their best accuracy is obtained by the combination of word unigram and bigram combined with the BOW model and RFDT with 93.5%. Research [2] conducted further research from research [1] by using Fasttext as a classifier to detect hate speech text in Indonesian Instagram comments. Their best score result was 65.7% in F1-score measurement, but compared to previous research [1], the result of RFDT using the dataset of [2] is 50.1%. It is smaller than using Fasttext.

Researchers in research[8] studied a combination of Support Vector Machine(SVM), and character quad-grams to detect hate speech using the English twitter dataset with 78% accuracy as their best result. Research [10] used Support Vector Machine (SVM) and Fasttext skip-gram model to classify IMDB movie reviews in English with an overall accuracy of 86.7%. They also classified Thai language movie reviews from Pantip.com by using the combination of SVM machine learning and the BOW model as the word embedding with an overall accuracy of 86.7%. In contrast, research [9] detected hate speech using Paragraph2vec and Multi-Layer Perceptron (MLP) as a classifier that obtained an accuracy of 99% Area Under the Curve(AUC).

According to research [1] and [10], the BOW model worked well when counting the frequency of words that appeared in the document. As a result, research [1] suffered from the missing context of person names and several words in the sentence, such as "*ahok*", "*babi*", "*penista agama*", etc. Although research [2] used the Fasttext as a classifier, several particular words similar to the research [1] problem were also falsely classified by the method. Therefore, selecting a suitable word embedding is important so that it can detect the context of words. Word embedding methods that popular in text classification are Word2vec [10],[11]; Fasttext[2],[10]; and Glove[11]. Word2vec is the old word embedding where it uses one word to make a vector. Fasttext is an improvement of word2vec where it uses character n-grams when making a vector of words. Glove is also an improvement of word2vec where it considers the occurrence of words when making the vector of words. From three word embeddings, word2vec is chosen because in the previous research [11] word2vec obtained 92% of accuracy which is higher compared to the Glove result. However, Fasttext in research [2] showed that it obtained 65.7% accuracy when detecting

hate speech in the Indonesian Language. It can be seen that word2vec can improve the accuracy of detecting hate speech as [11] obtained 92% and Fasttext in [2] only obtained 65.7%. Another reason to choose word2vec over Fasttext is the time to train the data for word embedding. In [12], Fasttext shows to have a long time to train on a large dataset. On the other hand, word2vec has an advantage, which is faster than Fasttext in processing a large dataset. Moreover, Fasttext is also used as a word embedding to compare the result of using the method with the result of word2vec that is proposed in this research.

Three word embedding methods above can perform two models, namely the Continuous Bag-of-Words (CBOW) model and the Skip-gram model[12]. CBOW model is an improved version of the BOW model where it learns by predicting words based on their context (surrounding words)[12]. It can learn a large amount of data faster than the Skip-gram model. However, CBOW is only good at predicting the words that often appear in the corpus. Skip-gram model is designed to learn by predicting the context of surrounding words to the word itself with a less amount of training data because it can represent words that rarely appear in the document[12]. The skip-gram model is chosen over CBOW because it can detect the word in the context of surrounding words which can solve the problem of the context of person names and several words that appear in the research [1] and [2]. By combining the skip-gram model with word2vec, the word embedding can gather words and rare words that were found in the sentence and represent it into the vector faster than using Fasttext. Certainly, it will improve the prediction from the system.

Imbalanced dataset problem always presents in the machine learning problem including deep learning when the number of data in one class outnumbers the number of data in the other class[13]. The impact of an imbalance dataset to the model is the result of F1-score, precision, and recall will obtain a zero or undefined number because the model only predicts one class that is mostly learned by the model. For example, if the dataset used for the experiment are consists of 90% of not hate speech data and 10% of hate speech data, then it is most likely the dataset are imbalanced. To handle the imbalanced dataset problem, the commonly used methods are oversampling, undersampling[13], etc.

As provided in research [1] and [2], the definition of hate speech is a reference to determine whether the text is hate speech or not hate speech. Although the hate speech has been successfully detected by the research [1] and [2], their model is falsely classified several sentences that contain specific person names and terms.

Another problem suffered in [2] is the occurrence of offensive terms in data and the subjectivity of the annotator when labeling the hate speech data. Thus, a rule is needed to eliminate the subjectivity of the annotator. Besides that, the rule also provides the criteria of text that must be labeled as hate speech.

This thesis focuses on hate speech detection in the Indonesian language using the rule of hate speech and Word2vec with Skip-gram model to understand the context of the word in a sentence. As aforementioned, several balancing methods will be used to solve the imbalance dataset problem in this research. Then, the rule of hate speech, word2vec, and balancing method will be combined with deep learning for text classification. Due to the explanation of the problem, the research question of this thesis is "How is the performance of the combined Word2vec Skip-gram model, balancing dataset method, and Convolutional Neural Network (CNN) in comparison with Fasttext method in detecting hate speech in the Indonesian language Instagram comments?".

1.3 Objective

This research focuses on detecting hate speech in the Indonesian language using word2vec and hate speech rule with Convolutional Neural Network (CNN). As a solution to the problem stated in the background, two objectives are conducted and resolved in this study. The objectives of this thesis are as follows.

1. Studying the combination of rule of hate speech, Word2vec with the skip-gram model, and CNN in detecting hate speech.
2. Building a system that implements a method in objective point one to detect hate speech in the Indonesian language.

1.4 Hypothesis

In this research, five rules to detect hate speech are formed from Indonesia Criminal Code[14], circular letter of the Indonesia police officer[15], the National Commission of Human Rights book[3], also several studies who performed hate speech detection using the Indonesian language [1],[2] and other languages[16],[17]. Studies [16], [17], and rules in [18], [19], [20], [21] have proved that rule of hate speech can detect the hate speech text better than not using any rule.

As mentioned in subsection 1.2, Research [11] has proved that word2vec with the skip-gram model can obtain a high accuracy of 92% when detecting tweets about the Zika virus. On the other hand, Fasttext which was used in research [2] obtained an accuracy of 65.7% to detect hate speech on Instagram. From the accuracy number, it can be seen that the word2vec with the skip-gram model has an advantage which can improve the Fasttext accuracy. Word2vec with the skip-gram model also can represent the word as one word where it is faster than Fasttext that used the n-grams method when representing the word. Using the skip-gram model in word2vec, it has a similar ability to Fasttext when representing rare words that appear in the dataset. Therefore, word2vec with the skip-gram model will be better than Fasttext.

Research [11] found three window sizes with values 5, 10, and 15 which affects the result of classification accuracy when used in their word2vec with the skip-gram model. Therefore, these three window size values are used in the experiment to get the best classification accuracy result from the proposed model. The dataset used in this research is not balanced where the percentage of not hate speech comments in the dataset is 90% and the percentage of hate speech comments is 10%. This will cause an overfit where most of the data were false classified to one majority class which is the not hate speech class. To handle the imbalanced dataset problem, four methods are used, namely 10-fold cross-validation, random undersampling, random oversampling, and class weight method. These four methods are then combined with three window size values and CNN to obtain the best F-Measure accuracy.

The combination of TextCNN and word2vec that suitable as proved in [22] used to predict Movie Reviews dataset, Stanford Sentiment Treebank dataset,

Stanford Sentiment Treebank with binary labels dataset, Subjectivity dataset, TREC Question dataset, Customer Reviews dataset, and Opinion Polarity MPQA dataset in English. The combination of CNN and word2vec was used in the study [22] to detect the subjectivity in the Subjectivity dataset with an accuracy of 93.2%. Other studies, such as [23] and [24] have proved that CNN provides better accuracy performance compared to several traditional methods, i.e, Logistic Regression, Support Vector Machine, Naive Bayes, Linear Discriminate Analysis, and K-Nearest Neighbor. Based on such studies, the hypotheses of this thesis are:

- Hypothesis 1: Window size of word2vec combined with balancing dataset method and the proposed model will affect the accuracy, F1-score, precision, and recall of the model.
- Hypothesis 2: The combination of the proposed method will give better accuracy values rather than only using the Fasttext word embedding as a classifier for hate speech text classification.

1.5 Research Methodology

The steps of research methodology that used to complete this research are as follows:

1. Problem definition
   This section defines the background/motivation, problem statement and research question, the objective, hypothesis, and research methodology related to hate speech detection.
2. Literature review
   This section is a summary and exploration results from previous research related to the basic theory of hate speech, hate speech rule, word2vec, and Convolutional Neural Network. This section also did a review and analyze the performance of previous studies.
3. System design
   This section discussed the process of hate speech detection. Moreover, it also explains the proposed hate speech rule, proposed hate speech detection method, dataset overview, and proposed evaluation method.
4. Experiment and analysis

This section explains the experimental design of this study including the purpose of the experiment, experiment scenario, performance comparison, etc. Furthermore, the analysis of experiments and discussion also explained.

5. Conclusions and future works
   This section discussed the conclusion of all experiments that have been done in the experiment section. Several methods can be used for improving the method are mentioned in the future work section.

1.6 Contribution

This research produces two contributions for the next research under the same problem, especially for those who are interested in hate speech detection. The contributions are as follows:

1. Rule of hate speech to label the dataset. Not many hate speech detection researches in the Indonesian language build and use rules to label and detect hate speech. Most of it only focuses on the method to finish the problem of contexts and improves the classification accuracy by adding more features.
2. Dataset used in the experiment. The data sets were collected from Instagram using Instagram-scraper from GitHub of rarcega with title instagram-scraper. It contains more than 10.000 comments from hundreds of captions and images in the form of text. It is posted by the user under several hashtags which indicated there are hate speeches in it. Hence, hashtags used by the author are under several areas such as politics, economy, etc. All data sets material and raw JSON files can be accessed at https://github.com/sirent/IGHateSpeechDataset.