CHAPTER 1

INDTRODUCTION

This chapter includes the following subtopics, namely: (1) Rationale; (2) Statement of the problem; (3) Hypothesis; (4) Assumption; (5) Scope and Delimitation; and (6) Importance of the study.

1.1 Rationale

Current technological developments can be seen from the existence of the internet and the emergence of various types of social networks [1]. Social media offers many conveniences for its users, but it can also have a negative influence. An example in recent years is uncontrolled communication and the widespread use of abusive language and hate speech on social media. Abusive language with hate speech can occur anytime and anywhere, so it will be very easy to find on the internet. This is because there are still many who use social media irresponsibly under the pretext of 'this is my social media, so it's up to me what to say'. In addition, netizens always take refuge in Indonesian human rights law regarding freedom of speech and expression, which actually has limitations on these freedoms [2].

Abusive language is an expression that contains harsh words and dirty phrases both in spoken and written language so that it can involve sarcasm, harassment, hate speech, etc [3]. Abusive language is an expression that contains harsh words and dirty phrases both in spoken and written language so that it can involve sarcasm, harassment, hate speech, etc. Whereas hate speech is speech directly or indirectly that refers to content or communications containing threats, insults, or discrimination against individuals or groups based on race, religion, ethnicity, gender, sexual orientation, or other attributes [4]. Hate speech that contains abusive language will often accelerate social conflict due to the use of words or phrases that trigger emotions. Even though abusive language is sometimes only used for jokes (not to offend people), the use of abusive language on social media can still cause conflict due to misunderstandings among netizens [5]. In addition, children will also be exposed to language that is not appropriate for their age so that it can cause severe psychological harm to their development as a result of abusive language and hate speech spread on social media [6], [7]. So that this can have negative social impacts and cause many problems in society.

Making hate speech on social media that contains abusive & offensive language is actually strictly prohibited by law in Indonesia. This is written in Indonesian law Number 19 of 2016 Article 45A Paragraph 2 concerning changes to Law Number 11 of 2008 and Electronic Transaction Information. It is clearly explained in the law that anyone who intentionally and without rights disseminates information aimed at creating feelings of

hatred or hostility towards certain individuals and or groups of people based on ethnicity, religion, race, and intergroups will be subject to imprisonment and or fines [8], [9]. However, we still find a lot of hate speech containing abusive sentences on Indonesian social media.

Based on the 2021-2022 Indonesia Internet Survey (Q1) report made by the Association of Indonesian Internet Service Providers (APJII), writes that Indonesia's internet penetration has reached 77.02% with 210 million internet users out of the total population of the Republic of Indonesia of 272.2 million people [10]. And judging from the statistics on social media use in Indonesia in 2022, the number of active social media users is 191.4 million with an average usage of three hours and seventeen minutes. Twitter is one of the most used websites and social media with users of 58.3% of the total population in 2022 [11].

According to a recent survey by Amnesty International USA [12] and research by Joetta Di Bella & Fred C. Sauter III at Montclair State University [13], it shows that the Twitter platform has experienced an increase in abusive language and hate speech since Elon Musk acquired Twitter. And based on the CSIS National Hate Speech Dashboard Indonesia [14] there are nearly 11,000 tweets on Twitter containing abusive language and hate speech in Indonesia in 2021-2022. The reason that there is still a lot of use of abusive language and hate speech on the internet or social networks is the lack of effective tools to filter the use of these sentences, the lack of empathy among internet users, and the lack of parental supervision.

Abusive language and hate speech detection often suffers from bias and ambiguity [15], [16]. One approach that used to be done was based on keywords which had significant shortcomings, such as the inability to capture the full context, resulting in false positive and false negative detections, and requiring intensive keyword list maintenance. The approach with the subjectivity of human annotation also has shortcomings because it can have different interpretations for each person [17]. Therefore, the detection of abusive language and hate speech becomes very important because in addition to experiencing bias and ambiguity, individuals or groups who spread hate speech can be prosecuted due to legal implications [18]. Currently, text classification approaches can be used for detection, which is a field of Natural Language Processing that helps categorize or classify types of sentences into certain categories [19]. One of the text classification methods that is currently popular and has been proven to have high performance is the Bidirectional Encoder From Transformers (BERT) [20] and the Robustly Optimized Bidirectional Encoder From Transformers Approach (RoBERTa) [21].

1.2 Statement of the Problem

In previous research, the detection of abusive language and hate speech was often carried out using traditional machine learning methods. These methods include Naive Bayes, Support Vector Machines (SVM) [22], and Random Forests [23]. These models

rely on features manually extracted from text, such as word frequency, n-grams, and other linguistic features. Although this method is quite effective, it has limitations in handling complex context and language nuances and dealing with data imbalances, especially in social media texts that are often unstructured. Apart from that, other studies also use word embedding to produce vector representations with fixed dimensions, where words with similar meanings are placed close together in vector space. This allows machine learning models to understand the semantic relationships between words. However, traditional word embeddings [24] have limitations in handling context, as each word has the same fixed representation regardless of the context in which it is used.

So, different from previous research, this research uses a special model for Indonesia, namely IndoBERTweet [25] and IndoRoBERTa [26],[27]. IndoBERTweet is a model adapted from BERT and optimized for Indonesian, especially in the context of text from social media such as Twitter. IndoRoBERTa is an optimized version of IndoBERTweet, following the same principles as RoBERTa which is a development of BERT. IndoRoBERTa uses a larger data set, dynamic masking, next sentence prediction removal, and various hyperparameter adjustments to improve accuracy. Apart from that, this research also tries to use data balancing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) [22], random oversampling, and random undersampling. This is to see the effect of data balancing techniques on the performance of the IndoBERTweet and IndoRoBERTa models in detecting abusive language and hate speech on Twitter. There are several main questions in this research.

1. Will the use of the pre-trained IndoBERTweet model on the BERT algorithm and the pre-trained IndoRoBERTa model on the RoBERTa algorithm provide good accuracy in detecting multi-label abusive language and hate speech Indonesian on Twitter compared to previous studies [23] [28]?

2. How does accuracy affect the use of data balancing methods to overcome data imbalance in the IndoBERTweet and IndoRoBERTa models?

1.3 Objective and Hypotheses

The objectives of this research are to:

1. Compare the performance of the pre-trained IndoBERTweet model based on the BERT algorithm and the pre-trained IndoRoBERTa model based on the RoBERTa algorithm with previous studies [23][28] in detecting multi-label abusive language and hate speech in Indonesian on Twitter.

2. Looking at the impact of different data balancing methods, such as Synthetic Minority Over-sampling Technique (SMOTE), random oversampling, and random

undersampling, on the accuracy of the IndoBERTweet and IndoRoBERTa models in detecting abusive language and hate speech on Twitter.

The hypotheses of this research are:

1. The pre-trained IndoBERTweet based on the BERT algorithm will outperform the accuracy of the IndoRoBERTa model and traditional machine learning methods in detecting multi-label abusive language and hate speech in Indonesian on Twitter. This is because IndoBERTweet is specifically trained on Indonesian Twitter data, capturing unique linguistic nuances and utilizing contextual embeddings to better understand and classify nuanced language.

2. The application of data balancing methods, such as SMOTE, random oversampling, and random undersampling, will significantly improve the accuracy of the IndoBERTweet and IndoRoBERTa models in detecting abusive language and hate speech on Twitter.

1.4 Assumption

Several assumptions were made in this study. First, it is assumed that the pre-trained IndoBERTweet and IndoRoBERTa models are capable of capturing and understanding the contextual embeddings and semantic relationships in the text. These models are expected to leverage their deep learning architecture to handle complex language structures, providing a significant improvement over traditional machine learning methods and static word embeddings. Second, the application of data balancing techniques such as SMOTE, random oversampling, and random undersampling will effectively address the issue of data imbalance in the training datasets. This balancing is presumed to lead to more accurate and generalizable models by preventing the models from being biased towards the majority class and improving their performance in detecting minority class instances. Lastly, the evaluation metrics and methodologies used to assess the performance of the models are robust and provide a comprehensive measure of their effectiveness in detecting abusive language and hate speech. This includes metrics such as accuracy, precision, recall, and F1-score, which collectively give a holistic view of the models' capabilities.

1.5 Scope and Delimitation

The study utilizes datasets derived from Twitter, containing a variety of text forms, including slang, abbreviations, and mixed languages that are prevalent on social media. This study only focuses on abusive language and hate speech, without distinguishing between categories or sarcastic words. And This study also utilizes IndoBERTweet and IndoRoBERTa, models designed for Indonesian which are expected to increase detection accuracy. To address the issue of data imbalance, techniques such as Synthetic Minority

Over-sampling Technique (SMOTE), random oversampling, and random undersampling are employed. The performance of the models is evaluated using metrics such as accuracy, precision, recall, and F1-score.

1.6 Significance of the Study

The significance of this study lies in its potential contributions to the field of Natural Language Processing (NLP) and its practical implications for social media moderation and online safety. Specifically, the study holds the following importance:

1. By utilizing IndoBERTweet and IndoRoBERTa, this research pushes the boundaries of NLP applications tailored to the Indonesian language. These models, fine-tuned for the specific nuances of Indonesian social media text, can serve as benchmarks for future research and development in the area of NLP for low-resource languages.

2. This research also aims to improve accuracy a in multi-label classification of abusive language and hate speech on social media platforms. These improvements are critical to developing automated moderation tools that can effectively identify and mitigate harmful content, thereby contributing to a safer online environment.

3. By exploring the impact of data balancing techniques such as SMOTE, random oversampling, and random undersampling, this research contributes to a broader understanding of how to handle imbalanced data sets in NLP tasks specifically in the IndoBERTweet and IndoRoBERTa models. These insights can be useful for researchers and practitioners facing similar challenges in other fields.