CHAPTER I
INTRODUCTION

In recent years, social media usage in Indonesia has experienced a notable increase. As reported by We Are Social in January 2022, the country had more than 191 million active users (Kemp, 2022). This rapid growth of social media has transformed the way information is disseminated, enabling internet users to become both contributors and active disseminators of information (Alamsyah & Adityawarman, 2017). These platforms have fostered seamless communication among individuals. Unfortunately, this unbridled exchange of information has led to the use of offensive and disrespectful language (Plaza-del-Arco et al., 2021), resulting in growing concerns about online toxicity.

Online toxicity refers to harmful and hostile communication in digital interactions (Hauser et al., 2017). It can manifest in several forms, such as cyberbullying, trolling, and the formation of virtual mobs or "online firestorms" (Salminen et al., 2020). Toxic behavior can cause users to disengage from online platforms (Hosseini et al., 2017) and negatively impact the mental health and well-being of individuals exposed to or subjected to such behavior, potentially leading to self-harm, depression, and anxiety (Aleksandric et al., 2022). Addressing and mitigating online toxicity is crucial to create a safer and healthier online environment.

Effectively identifying online toxicity requires understanding the contextual intricacies of the Indonesian language, as toxic comments depend not only on offensive words but also on the overall message conveyed. Natural Language Processing (NLP) can address this challenge by employing advanced computational techniques to analyze, understand, and interpret human language. A significant advancement in NLP

is the development of transformer-based language models, which excel at capturing context and semantic relationships within text data. This enables the effective identification and assessment of potential instances of online toxicity.

Several studies have explored online toxicity detection in Bahasa Indonesia using language models. Hana et al. (2020) investigated tweet classification for hate speech using SVM, CNN, and DistilBERT, finding SVM without stemming and stopword removal to be the most effective method, achieving 74.88% accuracy. Nabiilah et al. (2023) classified toxic comments into four categories: pornography, hate speech, radicalism, and defamation, using pre-trained models like IndoBERT, Multilingual BERT, and IndoRoBERTa Small, and by employing the same hyperparameters for all models (5 epochs and 8 batch sizes), the optimal F1 score of 88.97% was achieved with the IndoBERT model. Rivaldo et al. (2021) conducted a similar study, using Multilingual BERT and IndoBERT to classify toxic comments into six distinct categories. Employing consistent hyperparameters (3 epochs, 2e-5 learning rate, and a batch size of 32), the study achieved high test accuracy scores across all categories using an unbalanced dataset of 13,000 instances.

These studies highlight the advantages of using pre-trained language models for online toxicity classification in Bahasa Indonesia. However, there is a need to further investigate various hyperparameters to enhance classification performance, particularly with more balanced datasets. Dong et al. (Dong et al., 2020) suggested that utilizing balanced datasets could lead to improved classification outcomes. Therefore, future studies should consider incorporating balanced datasets to enrich the diversity of training data, resulting in more efficient tools for detecting and classifying online toxicity in Bahasa Indonesia. This, in turn, can contribute to creating a safer and healthier online environment for Indonesian users.

This research aims to detect online toxicity, a task made challenging by the complexity and nuances of language. To tackle this issue, we employ the BERT and RoBERTa language models, renowned for their advanced NLP capabilities. We evaluate the effectiveness of these two models in identifying toxic content and assess their performance to determine which model is better suited for this specific application. By leveraging the power of transformer-based language models, we can create more robust solutions for detecting and mitigating toxic behavior in digital interactions, ultimately fostering a healthier and more secure online environment.