1. Introduction

Social media platforms like Twitter have grown in popularity as a hub for interaction and information sharing in the current digital era. However, cyberbullying, often known as online bullying, is a serious problem in the internet world. Cyberbullying is the term for aggressive, harassing, or intimidating behavior committed through a digital platform, such as the Twitter platform [1]. Cyberbullying has severe repercussions that demand serious attention.

Research has highlighted the significant harm that cyberbullying causes to its targets and the terrible psychological toll it exacts. People who are the targets of cyberbullying usually struggle with extreme stress, enduring despair, and overwhelming worrying impacts that can snowball into a dangerous deterioration of their mental health [2]. The concerning link between cyberbullying and its negative impact on mental well-being emphasizes the urgent requirement for effective methods of detection and support, especially on complex and widespread platforms like Twitter.

Many users are active on Twitter, and there is much information, making it difficult to find and recognize tweets that contain cyberbullying traits. Finding instances of cyberbullying on Twitter is, therefore, a difficult task. To address this issue, several studies have been conducted to create techniques for identifying cyberbullying on social media.

According to research [3], Twitter cyberbullying detection techniques take user credibility and the classification of texts into account. The objective is to provide an effective method for identifying cyberbullying in Indonesia. The development of Doc2Vec and Convolutional Neural Network (CNN) algorithms for detecting cyberbullying on Indonesian Twitter is covered in research [4]. According to research [5], utilizing Indonesian on Twitter, deep learning is used to classify bullying. The objective is to create a valuable system for classifying bullying information on the Twitter network in Indonesian.

CNN (Convolutional Neural Network) and GRU (Gated Recurrent Unit) were chosen from machine learning approaches. CNN, renowned for image recognition, proves adept at sequential data like text [6], making it fitting for discerning cyberbullying language traits. GRU, a specialized recurrent neural network, excels at modeling sequential dependencies, effectively identifying nuanced contextual language patterns indicative of cyberbullying [7].

This study uses four classification methods: CNN, GRU, CNN – GRU hybrid, and GRU – CNN hybrid. In building the model, TF-IDF is used as the extraction feature and GloVe as the expansion feature. The limitation of the problem in this study is that the tweet data contains Indonesian. The main objective of this study is to thoroughly evaluate and compare the effectiveness of these strategies in detecting cyberbullying, thereby advancing the ongoing effort to create safer online interactions.