Cyberbullying Classification Using Three Deep Learning models: GPT, BERT, and RoBERTa

Muhammad Abusaqer and Charles Fofie Jr

Department of Math and Computer Science

Minot State University

Minot, ND, USA

muhammad.abusaqer@minotstateu.edu; charles.fofiejr@minotstateu.edu

Abstract

This research paper presents a study on the classification of cyberbullying on social media feeds using three deep learning algorithms of GPT -3, BERT, and RoBERTa. Cyberbullying is a growing concern in social media, so it is crucial to develop systems for detecting and preventing it. Cyberbullying involves using technology to harass, threaten, embarrass, or target individuals based on age, gender, religion, etc. This paper proposes a system that leverages both machine learning and deep learning algorithms to detect cyberbullying and reduce its impact, particularly on teen suicides. The study trains the deep learning models on a dataset of 46,692 tweets.

Additionally, the study compares the performance of these deep learning models to traditional machine learning algorithms, including Support Vector Machines (SVM), Naive Bayes, and Decision Trees. The study results demonstrate that the deep learning models outperform the traditional machine learning algorithms in detecting cyberbullying. This study makes two contributions to the field. Firstly, it is one of the first studies to use the newly released deep learning models of GPT 3.0 from Open AI, BERT from Google, and RoBERTa from Facebook AI. Secondly, it supplies a performance comparison between these state-of-the-art deep learning models and traditional machine learning algorithms. The results of this study could also help develop tools to assist in monitoring social media for cyberbullying feeds and immediately deleting them.