Abstract

Cyberbullying Detection in Social Media Comments

Problem Statement

Cyberbullying has slowly turned into one of the biggest problems of our generation. Every day, people are getting hurt by words shared online, and the emotional damage keeps growing as social media becomes a bigger part of our lives. What makes it worse is that the systems made to detect this kind of abuse still don't really understand how humans communicate. They look for harsh words or harsh language, but most of the time, the real pain comes from messages that sound normal but carry a hidden meaning. Because of this, two frustrating things happen all the time normal posts get flagged for no reason, while truly hurtful comments pass through without any warning. It shows that these systems are not thinking the way people do. What we really need is a detection method that can feel and interpret context something more human-aware.

Proposed Solution

My idea is to create a smarter machine learning model that can sense the tone, mood, and intent behind a message, not just read the words on the screen. I'll be using a strong pre trained language model like NLP which is Natural Language Processing and BERT which is Bidirectional Encoder Representations from Transformers by applying this method it transfers learning with real social-media data. The goal is to help the model pick up on hidden hostility and subtle emotional cues, like sarcasm or passive aggression, that ordinary systems usually miss.

Expected Outcome

At the end, I'll prepare a full research report that explains the development process, how the model works, its performance. My hope is that this project can contribute to building safer online spaces, where technology can actually recognize and stop harmful conversations before they cause more emotional and personal damage.

Git Link

https://github.com/PanchamiVaradaraju/Cyberbullying-detection-in-Social-Media-Comments/blob/main/README.md