Journal of Information Systems Engineering and Management
2025, 10(24s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Cyber Shield: Protecting the Digital Space from Bullies

Dr. B. Vijaya Kumar[1*] , Dr. S. Balaji[2] , Mr.R.Suresh[3]
([1,2,3] Professors),
Harini B[4], Shiamamadanki Ramalingame[5], Nandana K P[6]
Sri Manakula Vinayagar Engineering College

ARTICLE INFO

Received: 28 Dec 2024

Revised: 14 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Introduction: The cyberbullying detection system combines Artificial Intelligence (AI), Natural Language Processing (NLP), and deep learning to offer a sophisticated solution for detecting toxic online interactions. In contrast to conventional keyboard-based filtering, which tends to misclassify harmless content or miss implicit bullying, this system uses BERT (Bidirectional Encoder Representations from Transformers) to better comprehend the context and meaning of text. Cyberbullying may have catastrophic impacts on the victim, including, anxiety, depression and social exclusion. The requirement for an automated and smart detection system has become imperative as social media keeps evolving into a central mode of communication

Objectives: Cyber shield creates a highly precise and content sensitive cyberbullying detection system based on deep learning techniques. The system will improve social media surveillance by minimizing false positives and negatives in detecting bullying. It also targets real-time content categorization, which ensure that offending interactions are reported in real-time for intervention. The second major objective is to enhance detection of sarcasm, implicit bullying, and developing slang words that tend to outsmart conventional filtering methods. Finally, the system is scalable and flexible to ensure it can be used on other social media websites and languages.

Methods: The cyberbullying identification system operates within an organized method, beginning from data preprocessing to preprocess and normalize text data. Preprocessing involves tokenization, removing stop words, and lemmatization to normalize text input. After this, the system uses BERT to make its feature extraction and context understanding. BERT is going to understand the context of a word based on its relationship with the surrounding words in some sentence. Then we further train it on labelled data that has bullying and non-bullying text samples. The model was trained on comments and some performance indicators such as accuracy, precision, recall and F1-score were used to evaluate the model.

Results: This approach enhances the detection of cyberbullying beyond what traditional models can do. The context-aware processing improves the detection of some discrete and implicit form of bullying and offensive content that are incorporated in the neutral language text. Also, the real time processing abilities allows harmful content to be identified immediately and allowing timely intervention.

Conclusions: Cyber shield is extremely useful for identifying inappropriate content on social media through real-time monitoring with contextual understanding. Further enhancement could include multilingual processing, it will improve detection accuracy and increase the availability of this system. We can also add the detection of multimedia content such as videos, audios…

Keywords: BERT, Machine Leaning, Cyberbullying Detection, OCR, social media.

INTRODUCTION

With the evolution of the internet nowadays social media has become a highly used socializing and connecting platform. And it comes under many names and with many unique features. Furthermore, it allows to communicate via a numerous multimedia format such as text, video, audio… It is a great way to create and share content and

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

awareness with other social media users especially because it has become an important part of their daily lifestyle. Even though social media has a lot of pros it also has its cons. And one of them is cyberbullying.

Cyberbullying is the use of internet to harass, embarrass, hurt or make fun of someone online [1]. And having a conversation online allows the users to maintain anonymity which, in some ways, gives the bullies the freedom of writing hurtful comments and posting offensive posts. And the above-mentioned anonymity makes it hard to identify those bullies. But those offensive comments can affect the bullied person's mental health. They can cause depression, anxiety and in the worst case they can lead to suicidal thoughts [2]. The lack of filtration in the feed of any social media platform results in the above-mentioned consequences. So having a filter in every social media would help to maintain a healthy network.

For this, to analyse the sentences and words Natural language processing is the optimal option. NLP is a subfield of artificial intelligence which is predominantly used in Large Language Models (LLM). So, NLP allows you to identify the harmful language used in the comments and captions. So, NLP will be helpful in breaking down the textual content and analysing it to find offensive words or group of words to conclude that the comment or caption is cruel. NLP uses many techniques like cleaning, tokenization, stemming, lemmatization, stop word removal…

Therefore, NLP can be used for textual content. However, cyberbullying does not only occur through comments or captions. Cyberbullying can be also reflected through the posts which would come under the category of visual content. And for that deep learning is used for visual data analysis that is images [4]. It is a subset of machine learning which tries to imitate the human brain that is the neurons. So, deep learning uses multiple convolutional layers to analyse an image to whether classify the whole image as bully or non-bully. Otherwise, it can be used to extract the text that image may contain one of such images can be memes. Those types of posts contain pictures as well as texts. So, once the text is extracted, we can use NLP to process it and identify the bully post and remove them from the feed. So, integrating textual and visual content analysis in order to detect the bullying comments, captions and posts would be favourable for a healthy and bully free environment on social media.

This project consists of analysing the textual content using NLP and using BERT which is a transformer-based model that google created for applications including question answering, sentimental analysis and language comprehension [3]. So, it will help to establish the meaning of words in relation to other words in a sentence. And using deep learning we will extract the textual content on images and use NLP to process it. By extracting text from various media, OCR technology enables the analysis and flagging of dangerous content that could otherwise go unnoticed in systems that are solely text-based. Let us assume that someone has posted a meme with offensive language, so in this case OCR can extract that text from the image and process it using the trained BERT model [8].

OBJECTIVES

Cyber shield is expected to provide an insightful and automated feature for prevention and the detection for identifying harmful social media interactions. And using NLP techniques, with BERT [5] at the centre of the system ensures context-aware classification of social media posts.

Unlike keyword-based filtering which often misclassify content that is not offending anyone in anyway or fails to recognize implicit bullying. It is to overcome this, that we are moving to BERT to understand the overall context of the sentence before detecting it.

ARCHITECTURE

This system aims to detect abusive social media interaction using NLP, specifically we have BERT. Unlike traditional models that tend to not consider context or the sarcasm and indirect bullying, this model would provide more context sensitivity and precise detection. By incorporating real-time processing, it facilitates rapid intervention and moderation through a dashboard, minimizing extended exposure to abusive content.

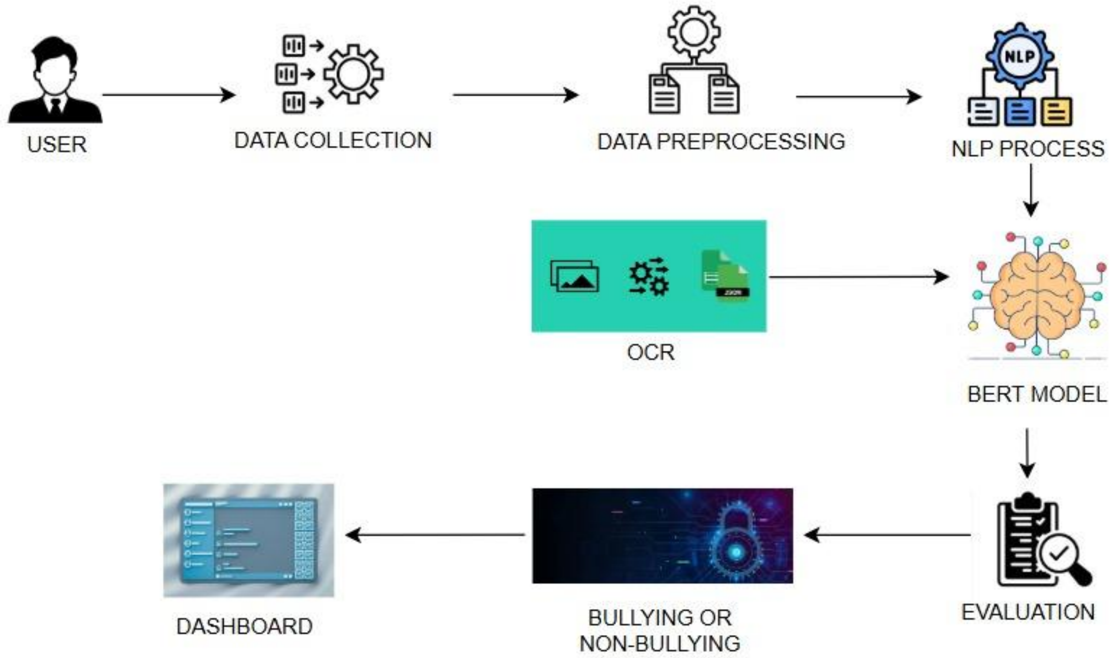

Fig 1.1: Architecture diagram

It uses OCR technology to get text from images and screenshots, so there is complete coverage of online interactions. The data is then pre-processed where noise removal, tokenization, removal of stop words and lemmatization are used to clean the text and make it ready for analysis.

After being pre-processed, the text is then fed through BERT, which learns deep contextual features. BERT, unlike traditional model, knows relationships between words in both directions and can recognize concealed bullying trends, sarcasm and new slang. The processed text is then assessed and classified as bullying or non-bullying.

METHODS

The cyberbullying detection system employs a step-by-step method that includes data preprocessing, feature extraction, model training, classification, and evaluation to ensure proper and accurate identification of toxic content. The detection begins with data preprocessing, where raw social media text, that is text with the hashtags and mentions, is purified by eliminating noise, stop words, lemmatization, and tokenization. All these procedures assist in normalizing and formatting the input data for analysis and enhance the accuracy.

Following preprocessing, feature extraction is accomplished using BERT. Instead, it reads the sentences in both forward and backward directions, so the semantic relationships and meaning based on the context are retained. It has better classification accuracy by introducing lexical challenges such as identifying implicit bullying, sarcasm and slang keywords.

The BERT model is trained on a wide range of labelled data which are example of bullying and non-bullying comments that are then fine-tuned. To achieve good generalization, the dataset is split into train and test sets. During training, hyperparameters tuning along with loss function optimization is performed enabling the model to enhance in the prediction of classes. The model is built on this data and helps in getting category labels for the social media messages, i.e., to find out if the message is bullying or non-bullying. Real-time processing of this content allows the system to recognise which content is offensive so that we can intervene before the content further spreads on the social media.

In addition, the system is integrated with Optical Character Recognition (OCR) and can extract text from images. As cyberbullying usually takes place via memes, screenshots or any other form of multimedia, but we are going to focus only on images, OCR detects the text embedded into them. This makes the system more effective at tracking more kinds of online interactions beyond typical text-based post.

This system is evaluated using accuracy, precision, recall, F-score. Precision measures how many comments flagged as cyberbullying are genuinely cyberbullying, and recall measures the rate at which the system can find all real instances of cyberbullying. And the F-score is the harmonic mean of the precision and recall.

RESULTS

The cyberbullying detection system, based on BERT along with OCR to counter implicit bullying, sarcasm and evolving slang, achieves much better results than common keyword-based systems. By analysing whole words and their contexts, it improves classification accuracy rather than just relying on keywords.

The addition of OCR technology extends the system's application beyond text conversation, where the system can decipher text in an image, a common medium for sharing and for cyberbullying. The system has been tested with precision, recall, and F-score algorithm which assures its high credibility and accuracy as a scalable flexible solution that will provide for a safer online social interaction in major social media channels.

DISCUSSION

The utilization of BERT and OCR for identifying and detecting toxic interactions on social media provides contextual classification and it is not keyword driven. Therefore, it recognizes implicit bullying, sarcasm, and evolving slang. Real-time detection helps to react faster to the harmful connection and avoid showing unwanted images for a long time, which is possibly more suitable for the style of current internet communications.

It demonstrated a balance between flexibility as well as performance when deployed on multiple social media sites and the precision, recall and F1-score metrics demonstrated that it is efficient and reliable. Further enhancement can be made by including multilingual capabilities and including other types of multimedia (audios, videos…).

REFRENCES

[1] Varun Jain, Vishant Kumar, Vivek Pal, Dinesh Kumar Vishwakarma. "Detection of Cyberbullying on Social Media Using Machine Learning."

[2] Pradeep Kumar Roy, Fenish Umeshbhai Mali. "Cyberbullying Detection Using Deep Transfer Learning."

[3] Kazi Saeed Alam, Shovan Bhowmik, Priyo Ranjan Kundu Prosun. "Cyberbullying Detection: An Ensemble Based Machine Learning Approach"

[4] Mohammed Hussein Obaida, Saleh Mesbah Elkaffas and Shawkat Kamal Guirguis. "Deep Learning Algorithms for Cyber-Bulling Detection in Social Media Platforms"

[5] Belal Abdullah Hezam Murshed, Jemal Abawajy (Senior Member, IEEE), Suresha Mallappa, Mufeed Ahmed Naji Saif, Hasib Daowd Esmail Alariki. "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platforms"

[6] Mohammed Al-Hashedi, Layki Soon, Hui-Ngo Goh, Amy Hui Lan Lim, Eu-Gene Siew. "Cyberbullying Detection Based on Emotion."

[7] Ayesha Atif, Amna Zafar, Muhammad Wasim, Talha Waheed, Amjad Ali, Hazrat Ali (Senior Member, IEEE), Zubair Shah. "Cyberbullying Detection and Abuser Profile Identification on Social Media for Roman Urdu."

[8] Fakhra Razi, Naveed Ejaz. "Multilingual Detection of Cyberbullying in Mixed Urdu, Roman Urdu, and English Social Media Conversations"