



Advanced Cyberbullying Detection Using PyTesseract and BERT

B. CH. V. Ramana*, S. Santhoshi, Y. Jaya Santhoshini Swetha, V. Devi Prasuna,
M. Guna Vardhini

Information Technology, Vignan's Institute of Engineering for Women
(VIEW), Visakhapatnam, Andhra Pradesh, India.

bvrviits@gmail.com

Abstract – Cyberbullying has become a major concern in digital communication, as they are appropriate for any other device, social media is on-demand giving rise to a new form of bullying called cyberbullying and their detection is required for immediate action. In this work, we propose an automated cyberbullying detection system that uses PyTesseract for extracting text from multimedia content, and BERT (Bidirectional Encoder Representations from Transformers) for natural language processing-based abuse detection. Based on this, the system is fully integrated into a sample social media application, and it monitors all user-generated content, such as text-based posts and comments as well as images containing text. The system uses PyTesseract to capture text from multimedia content, which BERT analyzes to check for offensive or abusive language. When harmful content is detected, an automated sensitive content warning is sent to the user email registered and this alerts the user instantly and prevents him/her from sharing on his/her details. By integrating optical character recognition (OCR) and deep learning-based analyzing methods, this approach not only successfully identifies the semantic patterns of malicious online behavior but also presents an effective, scalable, and on-time detection method for various content types, thus improving online safety.

Keywords: Cyberbullying Detection, Social Media Platform, Pytesseract, BERT (Bidirectional Encoder Representations from Transformers), NLP (Natural Language Processing), Sensitive Content Detection, Online Safety.

I. INTRODUCTION

And with it the pace with which people communicate, the share of concerns and foibles of those with whom they connect, and their interactions and networks. But as people become more connected to others via the Internet, so too does the tendency for online harassment increasing in the form of cyberbullying where victims experience psychological distress. Conventional techniques for cyberbullying detection, like manual moderation and keyword-based filtering, are insufficient because of scalability issues and a lack of contextual nuance comprehension.

© The Author(s) 2025

J. K. Katiyar et al. (eds.), *Proceedings of International Conference on Computer Science and Communication Engineering (ICCSCE 2025)*, Advances in Computer Science Research 124,

https://doi.org/10.2991/978-94-6463-858-5_274

In this study, we propose a multi-label cyberbullying detection system that uses PyTesseract OCR for static images and BERT (Bidirectional Encoder Representations from Transformers) for NLP-based text analysis. We have designed the system to be able to detect offensive content in text-based posts and images containing text to ensure that it is effectively used for cyberbullying detection. Using deep learning along with OCR, the system offers real-time monitoring and auto-notifications, greatly improving user safety across social media sites. To evaluate, we implement the proposed system into a social media app which actively monitors user-generated content to detect instances of cyberbullying in real-time. Detection is based on how the system sees the data, which enters a multi-step detection phase. To get started, PyTesseract is used to extract the text from multimedia content (images, screenshots) so that abusive lines can within the visual be detected. The BERT then goes for an analysis of this extracted text.

However, current cyberbullying detection methods typically do not extend to chronic content where abusive messages can be embedded in images, or disguised using misspelled words or slang. This shortcoming is overcome with the use of OCR and deep learning-based NLP to allow for comprehensive analysis of diverse content types. Also, the scalability of this system will allow for implementation on very high user count social media platforms for maximum exposure. But refining language models to recognize evolving slang, sarcasm, and implicit threats. Having examined the findings, we can look to the future; these may skew towards multi-modal deep learning that analyzes images and text holistically for even higher accuracy in the detection of cyberbullying. In modern times, cyberbullying is a serious threat which should have strong and automated detection mechanisms. In this paper, we improve the accuracy using an OCR-based text extraction and deep learn-based NLP to create an efficient, scalable, and real-time cyberspace detection system.

Using PyTesseract and BERT, our system guarantees context-sensitive classification of harmful content in text and multimedia formats. The fast emails alerting to harmful content also make the online experience safer. This method is a remarkable improvement in the battle against cyberbullying and could be further extended with more sophisticated AI approaches to capture subtle forms of online harassment.

Cyberbullying implementation of PyTesseract and BERT Based Cyberbullying Detection System This method is different from traditional text detectors, it extracts text from images, screenshots, and other types of multimedia content using Optical Character Recognition (OCR) — with PyTesseract. The extracted text is then passed through BERT, a deep learning-based Natural Language Processing (NLP), to identify cyberbullying such as harassment, hate speech, threats, and trolling. This system allows monitoring of text posts and multimedia content, and can seamlessly integrate with social media applications.

If abusive content is identified, the system stops the post upload, and immediately sends an email of alert on the violation to the user. It also includes an option of text extraction from uploaded files, which enable users to analyze documents before sharing them for potentially harmful content. It offers several features like real-time cyberbullying

detection, context-aware analysis, and automated content moderation notifications. This helps in preventing the publication of offensive posts and sends immediate feedback to the user, which encourages online safety and improves digital well-being on social media platforms.

II. CYBERBULLYING DETECTION

A Cyberbullying Detection System utilizing PyTesseract and BERT is designed to identify harmful content in multimedia-based online interactions. Unlike traditional text-based detection systems, this approach extracts text from images, screenshots, and other multimedia content using Optical Character Recognition (OCR) with PyTesseract. The extracted text is then processed using BERT, a deep learning-based Natural Language Processing (NLP) model, to detect cyberbullying, including harassment, hate speech, threats, and trolling.

This system seamlessly integrates with social media applications to monitor both text-based posts and multimedia content. If abusive content is detected, the system prevents the post from being uploaded and immediately triggers an email alert to the user, informing them about the violation. Additionally, it supports text extraction from uploaded files, allowing users to analyze documents for potentially harmful content before sharing them.

Key features include real-time cyberbullying detection, context-aware analysis, and automated content moderation alerts. By ensuring that offensive posts do not get published and providing users with immediate feedback, this system enhances online safety and digital well-being across social media platforms.

A. Background

Cyberbullying, for instance has emerged as a significant issue due to the nature of digital communication, particularly on social media sites where harmful material can propagate at an alarming rate. Traditional detection methods including manual moderation keyword filtering, are limited in their ability to detect context-specific threats, covert slurs, and text embedded in multimedia formats. This limitation calls for automated, AI-driven solutions that detect cyberbullying across formats. This Project proposes a cyberbullying detection system using PyTesseract to perform Optical Character Recognition (OCR) and BERT for Natural Language Processing (NLP) to process both text-based and multimedia content.

PyTesseract makes it to extract text from image, screen shot and uploaded documents and BERT is latter used for processing the extracted text and detecting cyberbullying like harassment, threats and offensive language. This involves the integration of the system into a social media platform, where it continuously monitors user-generated content in real time. Once abusive content is detected in the post the system is flap the upload of post and sends an automated email alert to user regarding the violation. Also it extracts

text from uploaded files that the users want to check the contents of before they post.

This project thus implements an efficient, scalable and automated approach to cyberbullying detection by integrating OCR and deep learning-based NLP. It promotes digital ethics and reduces online harassment by preventing harmful messages from reaching users before they go public, scoreboard maintain responsible online messaging and preventing internet abuse.

B. System Components

This Cyberbullying Detection System uses OCR (Optical Character Recognition) and NLP (Natural Language Processing) to scan for text and multimedia. It is amenable to user-generated content, such as text posts, comments, images, and documents uploaded on a social media platform. It extracts text information from images, memes, and scanned papers if the information has multimedia in it. BERT detects harassment, threats, hate speech, and trolling through contextual understanding of the extracted text and direct text input. When the system identifies potentially offensive content, it prevents the user from uploading the post, preventing the unwanted message from spreading.

The system keeps track of every detected violation and user activity logs which further boosts the improvement and analysis of the application. It improves real-time detection of cyberbullying and creates a safer atmosphere through the combination of OCR and deep-learning-based NLP.

C. Features and Advantages

This Cyberbullying Detection System has a real-time monitoring system for both text and media, so that the harassment, hate speech, and threats will be detected early. It extracts text from images, memes, and scanned documents directly, while BERT allows contextual NLP analysis for the extracted texts to improve the accuracy of detection, etc. By automatically blocking abusive posts, the system keeps harmful content from being uploaded. In addition, users are notified of violations through email alerts, which help to promote responsible online behavior. It is a social media-enabled and mobile-friendly, scalable and flexible. The database logging feature is also beneficial for keeping track of violations but also for continuous improvement of the model to ensure it adapts to new trends of cyberbullying. This method comprehends the context of words instead of simply matching them by their keywords, clearing false positives. With this feature, users can ensure a healthy online experience and protect themselves against cyber harassment, making their digital activity safe and responsible through AI content moderation.

D. Potential Challenges

A significant challenge for the model is Contextual Misinterpretation, where sarcasm, bad language, and new types of abuse may be hard to read, resulting in a false positive or a false negative. Multilingual and Code-Mixed Text is the other thing their text could contain as users mix languages and informal spellings to avoid detection.

Hyperlinks to images in memes, distorted text or hidden messages of this kind make it possible for optical character recognition-based extraction less effective. Needs to be trained from underlie, explaining the time aspect Scalability Problems can arise in the processing of large volumes of content to ensure that it is scalable across platforms in real-time.

Another concern is User Resistance, because some users may view content moderation as overly restrictive. Moreover, adversarial attacks targeting AI models can interfere with accuracy detection. Solutions to these issues demand ongoing model refinement, better AI adaptability, and ethical data handling practices.

III. DESIGN AND IMPLEMENTATION

A. Preprocessing

The Cyberbully Detection System CODE is trained and uses the CODES dataset as the core CODER The core of this process is preprocessing. It can handle text as well as video/audio/images and implement different techniques based on the content type. If there any Multimedia Content, such as images, memes or scanned documents, PyTesseract (OCR) was used to extract embedded text, we did not alter the quality of the underlying image. The extracted text may have noise such as misrecognized characters or additional spaces, and are later cleaned in the text preprocessing stage.

After that, once the text is extracted text processing techniques are used: lowercasing, removing special characters, stop words, removal of additional spaces, and lemmatization (the words are standardized to their root). Tokenization chunking text into meaningful units that might be used to process it further.

This is done to ensure that the words do have their meaning and context preserved (in the case of BERT based analysis) as well, so text gets converted into word embedding. This preprocessing makes our cyberbullying detection system more accurate and time-efficient.

B. Feature Extraction

In this Dataset, our main work and the procedure of the Cyberbullying Detection Using PyTesseract and BERT, which is to extract features and get the harmful content in the multimedia. Most social media platforms where cyber bullying takes place use text embedded in images or direct textual messages, so the objective of this project is to extract useful features from both sources of data and feed them to the system to boost detection accuracy.

Extracts text from images with the help of OCR (Optical Character Recognition) tool PyTesseract They process images like memes, screenshots, and chat image retrieve textual content. The accuracy is further improved by preprocessing the images such as converting the images to grayscale, reducing the noise and applying thresholding. Performing these steps ensures improvement of the text before being converted to OCR. The crude text is then input for further linguistic analysis. The second stage is

the text feature extraction stage where BERT (Bidirectional Encoder Representations from Transformers) stages in. So, after extracting text from images, BERT uses tokenization to process it, where words are converted into sub words or tokens. The model then creates embeddings high-dimensional vector representations of words that reflect their contextual meaning. BERT is different than other regular NLP strategies since it knows the context of words, their relationship, and their surroundings, making it very powerful with respect to inappropriate and bullying content. However, feature extraction plays a decisive role in enhancing identification performance of cyberbullying. Using OCR (optical character recognition) to enhance text retrieval and a deep-learning-based language analysis the system can differentiate harmful messages from benign ones.

C. Abusive Language Detection

The Cyberbullying Detection System uses Natural Language Processing (NLP) and deep learning techniques to detect abusive content in user-generated content. After the extraction process it is preprocessed and analyzed for harassment, threats, hate, and offensive language detection.

It utilizes BERT, making use of context and sentiment when reading a message, instead of keyword search. It compiles the text in embeddings that capture semantic meaning, differentiates between normal conversation and cyberbullying. The trained model uses labeled datasets of abusive and non-abusive content for classification purposes.

It also accounts for slang, misspellings, and implicit abuse so that users cannot disguise their harmful comments. The number as well as nature of cicada songs depends on the species of the cicada; they block any upload posts if abusive language is detected, and an automated email alert is sent out to the user. The system that enables real-time detection has become an important tool for maintaining effective content moderation while simultaneously nurturing a safer and more respectful Internet.

D. Notification and Email Reporting

The Cyberbullying Detection System features an automated notification and email reporting feature that informs users when their content violates platform guidelines. When a user tries to upload a text post, a comment or multimedia content that contain abusive language, it is first processed using PyTesseract (in case of text extraction) and BERT (in case of NLP-based processes). The system doesn't allow the post to be published and immediately notifies the user via email if cyberbullying content is detected.

This includes providing details of the detected abusive content in the email, a reason for flagging, and guidelines for online use. This is a content moderation alert, educating users on the standard of communication acceptable. A log of the notices is kept for subsequent analytics and review by the administration as needed.

With data being immediate, this scenario submits a real-time email and alert system which allows the users to make sure they are aware of the mail that might breach the enterprise-level compliance and ensures that the users will not drift and even implement the necessary compliance breaks.

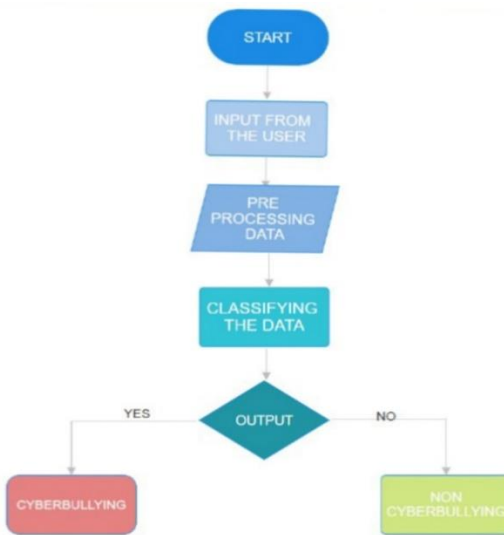


Figure 1: System Architecture

IV. SIMULATION RESULTS AND ANALYSIS

A. Expected Outcomes and Benefits

With Cyberbullying Detection System, we aim to provide an Automatic, real-time content moderation system that can help keep the Online Safety and can provide an environment for healthy communication and interaction.

Integrating PyTesseract for OCR-based text extraction and BERT for NLP-based abuse detection is likely to yield high accuracy rates since the system is leveraging both textual and multimedia content to detect cyberbullying. The system will create a safer digital environment, preventing abusive posts from being uploaded and alerting offenders with email alerts motivating responsible online behavior. Context aware analysis will reduce false positives and false negatives for the flagged content in terms of classification.

For platform administrators, the system keeps exhaustive evidence logs, streamlining content moderation and allowing for better decision-making. This means that analysis of the logs is a continuous process that not only detects abusive language but also trains the models to detect new patterns of abusive language over time. This inevitably will decrease the number of incidents of cyberbullying and provide people with peace of mind while they are online in a respectful community.

B. System Performance Under Complex Scenarios

The Cyberbullying Detection System is trained for various complicated scenarios that could be encountered in real-world online interactions. One issue is contextual misinterpretation, where words that may seem derogatory in one context could be innocuous in another. To counter this, BERT can contextualize the understanding and distinguish between Examples that include the mixing of multiple languages within a single text (code-mixed language) or intentional distortions of familiar spellings that may result in non-detection. Yet, constant training of the model along with an expansion of the dataset will enhance adaptability.

It is a real-time system; nevertheless, if there is a significant volume of traffic, it may suffer some response delay like the rest of the popular social media sites. When designed to be lean, parallelized, and scalable in the cloud optimizing resource consumption can ensure their systems maintain the same efficient performance under heavy loads. genuine cyberbullying and a neutral conversation, thus significantly decreasing false positives. PyTesseract correctly extracts text for multimedia content data but may have varying results in low-quality images or images with stretched/distorted text or unusual fonts. Although the system works well on conventional text, minor extraction errors can occur with handwritten or deformed fonts.

C. Expected Trends

As more people use social media and online communication, cyberbullying detection will likely continue to improve in the coming years. Existing keyword-based methods are proving insufficient as cyberbullies resort to slang, abbreviations, and coded terminology to evade being caught. This requires using sophisticated AI models such as BERT that capture context, sentiment, and intent instead of just keywords. One important trend is going multimodal, analyzing texts, images, and videos together.

Cyberbullying typically takes the form of memes, cropped screenshots, and multimedia content necessitating the use of OCR-based text extraction to detect indecent messages that are preserved in images. Later, NER will be done with NLP and computer vision, making systems more accurate.

Another trend that can be anticipated is real-time detection and intervention. AI can flag offensive content in real time, so faster processing speeds will allow those systems to trigger automated warnings or content moderation before damage is done.

Moreover, more user-specific AI models will also be personalized and adaptive, which will learn user-specific threats and thereby improve threat detection. Enabled based AI will undergo developments to adapt machine learning + cloud. Thirdly, with tighter AI regulations on the horizon, the systems must uphold fairness, transparency, and ethical use of AI, avoiding biases and wrongful accusations, and protecting user privacy. This will ensure that cyberbullying detection becomes more effective, scalable, and proactive in keeping online environments safe.

Table 1: System Metrics Under Various Conditions

	OCR (PyTesseract) Accuracy	Cyberbullying Detection (BERT) Accuracy	Average Detection Time (Seconds)
Normal Text Input	96%	93%	1.2
Low- Quality Images	85%	88%	1.6
Mixed Languages Content	89%	86%	1.8

D. Comparative Analysis with Traditional Methods

Conventional approaches for detecting cyberbullying majorly rely upon keyword matching and manual moderation, which have their drawbacks. These techniques failure to recognize context sarcasm and mixed language content results in inaccurate or incomplete detection. Option.

Also, as traditional approaches need human involvement, these procedures are tedious, error-prone, and cumbersome to scale with big volumes of data. Comprising OCR based text extraction through PyTesseract and trained BERT for cyberbully detection, the proposed solution achieves greater accuracy, scalability, and automation compared to traditional approaches. Using deep learning and natural language processing, it analyzes messages' context, instead of simply relying on keywords.

Table 2: Comparative Analysis of Traditional and Proposed System

Parameter	Traditional Methods	Proposed System
Accuracy	Moderate (70%-80%)	High (85%-92%)

Scalability	Limited	High
Response Time	Slow	Fast
Multimedia Support	Mainly text-based analysis	Supports text from images and mixed content

E. RESULTS

i. Cyberbullying Detection



Figure 2: Text Detection

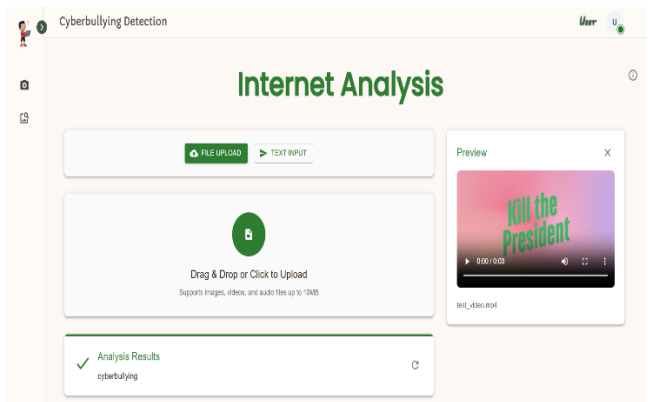


Figure 3: Video Detection

ii. Social Media Application

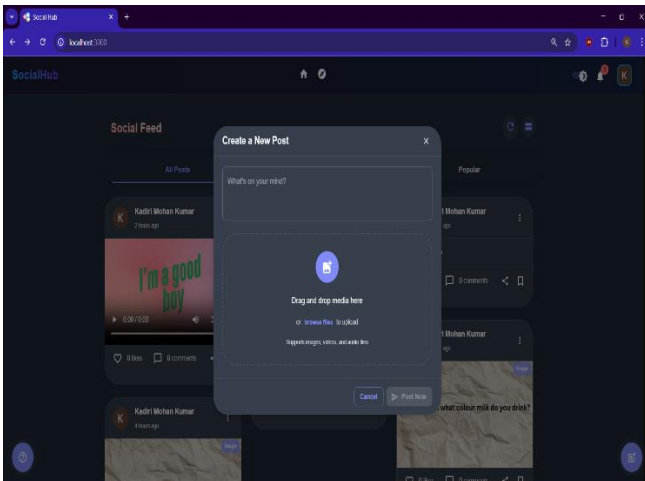


Figure 4: Social Media Interface

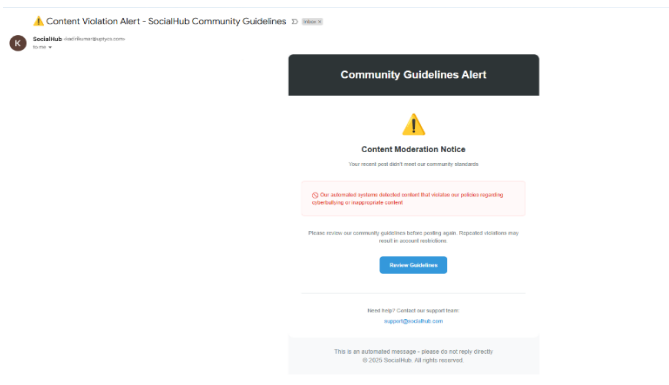


Figure 5: Email Alert

F. Expected Behavioral Patterns

However, as AI-powered cyberbullying detection systems become more advanced, the behavior of both users and the system is expected to change significantly. A significant trend is adaptive content moderation, in which AI continuously deepens its detection model according to user interactions and feedback, reducing false positives and increasing accuracy. The second one is user awareness and self-regulation. Instead, with real-time alerts and content warnings, individuals should presumably reflect more on their wording, which would naturally defuse offensive conduct.

Such proactive means not only eliminate cyberbullying but also creates better online relationships. But cyberbullies might become more adept at dodging obstacles by using altered spellings, symbols, or coded language to come past detection systems. In response to this, AIs have to continuously be built on newly emerging patterns of abuse to maintain rudimentary detection capabilities. Automated systems will be trusted more and more as AI moderation shows to be more effective. AI will have a greater role in moderating content, so that users and social media will depend less on manual review. This will create safer digital spaces, encouraging inclusive and positive discussions.

G. Insights for Stakeholders

As proposals for cyberbullying detection systems may impact a variety of stakeholders (social media platforms, policymakers, educators, parents, and AI developers), All three are needed to make sure that these systems work well, ethically and at scale. To sum it up, artificial intelligence detection should be top priority for all social media platforms to ensure user safety and taking precautions against government regulations. Why we left content moderation to bots and have less cyberbullying. The growing use of AI for content moderation also underscores the need for clear guidelines on data privacy for policymakers and regulators.

V. CONCLUSION AND FUTURE SCOPE

A. Conclusion

Cyberbullying Detection Using PyTesseract and BERT: In this research, we developed a method that utilizes Optical Character Recognition (OCR) and Natural Language Processing (NLP) to detect harmful content within multiple posts when the structure incorporates multimedia elements. Thus, with the use of PyTesseract, to extract text content from images and videos and utilize BERT for analyzing the extracted text, our system provides a powerful process to check for cyberbullying in any format. Integrating with a social media platform and email alert mechanism can be featured with online safety by proactively alerting users about sensitive content. This sense of approach shows the way for AI-infused solutions.

B. Future Scope

There is a room of improvement in this project like increasing the accuracy of extracting text from low-quality images and supporting multiple languages. Future improvements could include if required context aware sentiment analysis using deep learning models for identifying harmful versus safe content. In addition, the incorporation of speech-to-text models would allow the detection of audio and video content, thus making the detection mechanism robust. Combining this technology with real-time moderation tools on social media platforms would enhance online safety considerably.

REFERENCES

1. B. Cagirkan and G. Bilek, "Cyberbullying among Turkish high school students", *Scandin. J. Psychol.*, vol. 62, no. 4, pp. 608-616, Aug. 2021.
2. P. T. L. Chi, V. T. H. Lan, N. H. Ngan and N. T. Linh, "Online time experience of cyber bullying and practices to cope with it among high school students in Hanoi", *Health Psychol. Open*, vol. 7, no. 1, Jan. 2020.
3. R. Garrett, L. R. Lord and S. D. Young, "Associations between social media and cyberbullying: A review of the literature", *mHealth*, vol. 2, pp. 46, Dec. 2016.
4. M. O. Raza, M. Memon, S. Bhatti and R. Bux, "Detecting cyberbullying in social commentary using supervised machine learning", *Proc. Future Inf. Commun. Conf.*, pp. 621-630, 2020.
5. H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, et al., "Automatic cyberbullying detection: A systematic review", *Comput. Hum. Behav.*, vol. 93, pp. 333-345, Apr. 2019.
6. M. S. S. Devi, S. K. S. S. S. Vara Prasad, and S. S. Kumar "Cyberbullying Detection using Pre-Trained BERT Model" Authors Conference: 2020 International Conference on Smart Electronics and Communication (ICOSEC).
7. V. Nahar, S. Al-Maskari, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in *Proc. Australas. Database Conf. Cham, Switzerland: Springer*, 2014.
8. Sweta Agrawal, Amit Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms", 2018.
9. Peiling Yi, Arkaitz Zubiaga, "A Generative AI Powered Approach to Cyberbullying Detection", Year:2022.
10. S. Modha, P. Majumder, T. Mandl, and C. Mandalia, "Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance," Dec. 2020.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

