STM JOURNALS

International Journal of
Computer Science Languages

ISSN: 3048-944X
Volume 4, Issue 1, 2026
January–June
DOI (Journal): 10.37591/IJCSL

https://journals.stmjournals.com/ijcsl

Review IJCSL

Real-Time Browser-Based Early Warning System for Cyberbullying Detection in Online Platforms

Sonal Sharma[1], Manish Singh[2,*], Pratham Singh[2]

Abstract

The rise in social networking through internet-based communication tools, Instagram, and YouTube, to name a few, significantly increases the risk of cyberbullying, thereby increasing psychological trauma on users, especially children, through adverse emotional states like anxiety, depression, etc. For a long time, researchers have been enhancing detection tools to counter cyberbullying, but their ability to detect only after the fact, along with limited support for English-based architecture, is a major concern that also compromises user privacy. This study seeks to develop a browser-based real-time warning system that can identify potentially harmful or offensive messages before they are transmitted. This study targets the development of a scheme that is multilingual, privacy-concerned, and enables users with warnings before posting messages that protect them from harm. The proposed methodology involves the usage of secondary data sets containing previously annotated data from social media sources like Instagram and YouTube, along with the survey method of questionnaire formation. Additionally, the proposed approach utilizes the lightweight multilingual transformer architecture for the detection of cyberbullying; the same architecture will be adapted for in-browser usage with the aid of TensorFlow.js so that the data never gets processed remotely from the client's browser. The initial results also revealed the following insights regarding the competitive accuracy of the model, which could enable the reduction of inference latency: user responses also clearly pointed out the need for the model to cover the multilingual community, as well as the value of local data privacy. This study bridges the gap by using a combination of real-time detection, multi-lingual support, and a browser-based solution for privacy-safe deployment. The suggested algorithm aims at making the internet environment safer by offering a better guard against cyberbullying risks on various social media platforms.

Keywords: Real-time detection, browser-based system, privacy-preserving AI, natural language processing, multilingual analysis

INTRODUCTION

The emergence of different social media platforms, such as Twitter, Instagram, YouTube, Facebook, and chat-based applications, has dramatically revolutionized the face of modern communication. "Social media has provided people opportunities to express themselves and share their opinions with the global population." Although this has created unparalleled opportunities for information sharing, education, and connecting people globally, it has also led to some unhealthy trends, such as cyberbullying [1, 2].

The general definition of cyberbullying is the intentional use of digital communication tools to harass, threaten, or humiliate a person or group. In contrast to traditional bullying, which normally takes place within specific physical settings such as

***Author for Correspondence**
Manish Singh
E-mail: manishsingh59766@gmail.com

[1]Associate Professor, Master of Computer Applications, Thakur Institute of Management, Studies,Career Development & Research(TIMSCDR), Mumbai, Maharashtra, India
[2]Research Scholar, Master of Computer Applications, Thakur Institute of Management, Studies,Career Development & Research(TIMSCDR), Mumbai, Maharashtra, India

Received Date: March 27, 2025
Accepted Date: March 31, 2026
Published Date: May 08, 2026

Citation: Sonal Sharma, Manish Singh, Pratham Singh. Real-Time Browser-Based Early Warning System for Cyberbullying Detection in Online Platforms. International Journal of Computer Science Languages. 2026; 4(1): 9–15p.

schools or workplaces, cyberbullying knows no spatial or temporal bounds. Nasty messages, comments, and postings can be spread within an incredibly short time and thus reach a large audience in the blink of an eye, a fact that may increase the extent of possible damage. Moreover, perpetrators are usually tempted by the anonymity of online sites, which often makes them more aggressive or abusive than they would be in face-to-face interactions [3]. The impact of such a phenomenon is not restricted to transient emotional pain, as victims are likely to suffer from critical psychological, social, and even physical distress. There are a number of research studies that have already confirmed the direct relation between exposure to online harassment and the onset of "anxiety disorders," "depression," "reduced self-confidence," "academic underperformance," "social alienation," and even "self-injurious behavior/suicidal traits," especially when the experience of online harassment is taken to an extreme by the victim [4]. But the scenario takes an even darker turn when the target of online harassment is members of the "adolescent youth group," as they are already the major "users of cyber technology." To mitigate this, various techniques have been proposed for the detection and mitigation of the adverse effects. These include using content moderation tools, keywords, using machine learning models, reporting, etc. However, while these techniques mark important milestones in the creation of a safer cyber environment, they also present a fundamental flaw. This flaw emanates from the fact that, in most situations, the problematic content is mediated after the post, which implies the victim has already "read" the problematic material [5, 6]. This, in turn, implies psychological damage, which is difficult to heal. Another major drawback that can be inferred is the architectural design on which the current detection tools operate. Modern detection tools are most likely operating on a "server-based framework," whereby the messages or posts shared are sent remotely and analyzed via AI [7]. This is an advantage due to the 'computational power' that the tools can draw on. However, this approach raises some key questions. One major concern is that, due to the external sources that the tools are sending personal info to, there is a high chance that the details of users can be trespassed upon [8]. Further, awareness about the privacy surrounding digital tools has risen in recent times, with such tools operating in a rather 'suspicious' manner [9]. Another limitation that can be inferred is the "latency" that comes with having external sources [10]. This creates a situation where the intended content has reached the target even before the issue is detected.

Linguistic limitations also curb the efficacy of the current systems. Most widely adopted cyberbullying detection tools are designed for English-language datasets [11]. However, online communication is intrinsically multilingual and diverse. For instance, users in countries such as India often use code-mixed communication, such as Hinglish (Hindi + English) or Tanglish (Tamil + English). These linguistic variations mostly puzzle monolingual detection systems, often resulting in a high rate of false negatives or false positives. This linguistic bias leaves large user communities underserved and susceptible to abuse and misinformation.

Given these deficiencies, there is an imminent need for a proactive, multilingual, and privacy-preserving approach toward cyberbullying detection. In this regard, one of the potential research directions is the development of a browser-based real-time early warning system. In contrast to server-side architecture, this can work fully within the user's device, requiring no transmission of sensitive data outside the user's device. This ensures a high degree of user privacy and a high level of user trust. More importantly, however, because content is filtered before publishing, the system can technically function as a preventive measure that can warn users about harmful or offensive words, thereby allowing them to edit messages before publishing them on the internet [12].

Therefore, such a solution would have the potential to be a 'multilingual solution' as it would effectively handle regional languages, dialects, and code-mixed text that are used in digital communication in a region like India. Furthermore, the inclusion of natural language processing, in turn, with sentiment analysis, as well as entity detection-based approaches, would ensure a 'rich understanding' of contextual information, hence ensuring that the detection process itself remains accurate as well as 'inclusive' in nature, ensuring that offensive behavior will not be 'overlooked'.

Moreover, the recurring challenges of cyberbullying underscore the inadequacy of the prevailing solutions, emphasizing the compelling need to develop novel solutions. The browser-based, real-time, multilingual, and privacy-preserving solution to the cyberbullying challenge is a novel contribution to the resolution of the cyberbullying menace [13]. A real-time browser-based solution should not only empower Internet users by offering protection from the root causes of cyberbullying but also alleviate significant challenges to data privacy, language inclusivity, and other issues. This research is propelled by a firm conviction that a safe Internet is achievable through ethical technological innovation.

LITERATURE REVIEW

Hosseinmardi et al. [1] were among the first researchers to study cyberbullying detection using data from Instagram. They developed a smart model that combined text comments and image features in a multimodal way to detect potentially harassing behavior online. While their model demonstrated the capability of machine learning for detecting harmful interactions, some of the limitations of their model included a small dataset, only English, and offline processing. Thus, it was one of those seminal works that set up a platform for further research without being ideal in real-time scenarios. The race is for proactive systems that can halt the posting of harmful content before it ever hits servers.

Liu et al. [14] pursued this line from another angle by predicting escalation to hostility in Instagram comment threads based on the linguistic and social features. Their model achieved very good performance, establishing that aggression can be predicted from early conversational cues. Again, this was a retrospective method, with analysis possible only after user interactions had occurred. This limitation represents the necessity of browser-based, real-time interventions that may warn users before they create offensive messages.

The focus of the study was then shifted to YouTube in the study of Ashraf et al. [15]. This is due to the prevalence of abusive tendencies in the comments section of video-sharing sites. Through their study, the researchers found the benefit of including information in the detection of abusive language in social media. Although much has improved with the study of the proposed authors, the study remains parasitic and focused on the English language.

Remarkable work has been done by Raj et al. [16] in which a detection framework focusing on cyberbullying using Indian languages such as Hindi and Tamil was implemented. The model performed quite well in handling linguistic differences between languages, although it failed to perform well on language generalizations and other areas. This problem indicates a pressing need to support multilingual and mixed languages in text processing, which the proposed model hopes to solve in a browsing context.

Kim et al. [6] presented a privacy-preserving detection system based on federated learning, where model training was achieved in a decentralized manner directly on user devices. Although this provided improved data privacy, the approach required some coordination by the servers and was thus not entirely local. In our proposed system, we advance the idea of providing fully client-side real-time cyberbullying detection that ensures complete privacy with no reliance on servers.

PROPOSED FRAMEWORK AND SYSTEM DESIGN

A new system is proposed that uses a browser-based early warning framework with the ability to identify cyberbullying as well as harmful content or messages in real time, even before the message is posted on any social media platform. Unlike a conventional approach that uses all the computing power on the servers, the proposed approach focuses on proactive intervention, multiple languages, and user privacy, with all computations occurring on the client-side browser.

Overview of System

The system is implemented as a browser extension seamlessly integrated with social media platforms such as Instagram, YouTube, and other web-based communication interfaces. The system continuously

monitors the text being typed by the user when composing a message or comment, providing an early warning alert if the content is detected as potentially harmful or abusive.

Architectural Design

The system architecture was divided into four major modules: the input monitoring module, the pre-processing and language handling module, the cyberbullying detection engine, and the alert and feedback module, as shown in Figure 1.

Privacy Preserving

All analyses are performed locally on the client's browser, meaning that no data is transmitted elsewhere.

Implementation Technologies

All processing occurs locally within the user's browser, so there is a minimal chance of sending any sensitive data to servers remotely.

METHODOLOGY

This is a mixed-method study that combines secondary data analysis with insights from primary data.

Secondary Data Analysis

- We began with a thorough literature study focusing on the detection of cyberbullying on various platforms, such as Instagram and YouTube.
- We went straight into the publicly available data, which included abusive comments found on YouTube and Instagram, along with hate speech data available in various languages.
- Existing tools and models, such as BERT, IndoBERT, and transformer-based classifiers, were also documented using structured extraction forms.

Primary Data Collection

A structured questionnaire was developed to elicit the views of students, teachers, and social media users on the need to have existing browser-based cyberbullying detection tools that operate in real time.

- The questions centered on usefulness, trust, preferred languages, and the issue of privacy.

RESULTS AND DISCUSSION

The test for this proposed browser-based early warning system utilized datasets derived from YouTube and Instagram comments that indicated both noxious and non-noxious text forms. This cyberbullying detection was meant to occur in real time prior to posting messages (Table 1).

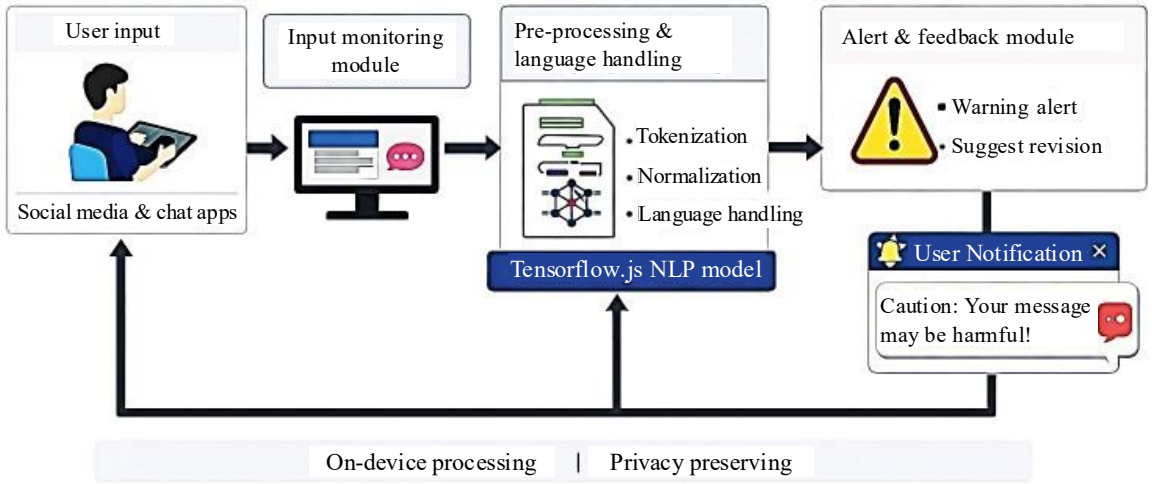

Figure 1. A browser-based real-time cyberbullying detection system.

Experimental Setup

- Used dataset
  - HASOC (hate speech and offensive content dataset—English)
  - YouTube comments dataset (This contains comments that have been annotated for abusive content)
  - Jigsaw Toxic comment dataset (Kaggle)
- *Languages tested:* English
- *Evaluation metrics:* Accuracy, precision, recall, F1-score, latency
- *Platform:* Chrome and Brave browsers, frameworks: TensorFlow.js, JavaScript.

Quantitative Results

Detection Performance

Figure 2 shows a comparison of the detected harmful, safe, false-positive, and missed messages within the evaluation dataset.

Detection Outcome Distribution

From Figure 3, it can be noted that approximately 72% of all messages were correctly identified as harmful, with 18% being safe messages.

CONCLUSION AND FUTURE WORK

This study describes the application of a browser-based real-time early warning system in the form of a cyberbullying detection method. This was done to address the issues experienced in the current reactive approaches to cyberbullying detection. Through the proposed method of detection prior to publishing the material, the implementation of cyberbullying detection no longer relies on the current approaches in post-cyberbullying detection and moderation.

Table 1. Quantitative results.

Model	Model accuracy
Proposed browser-based	91.2%
Server-based BERT	42.6%
Rule-based	76%

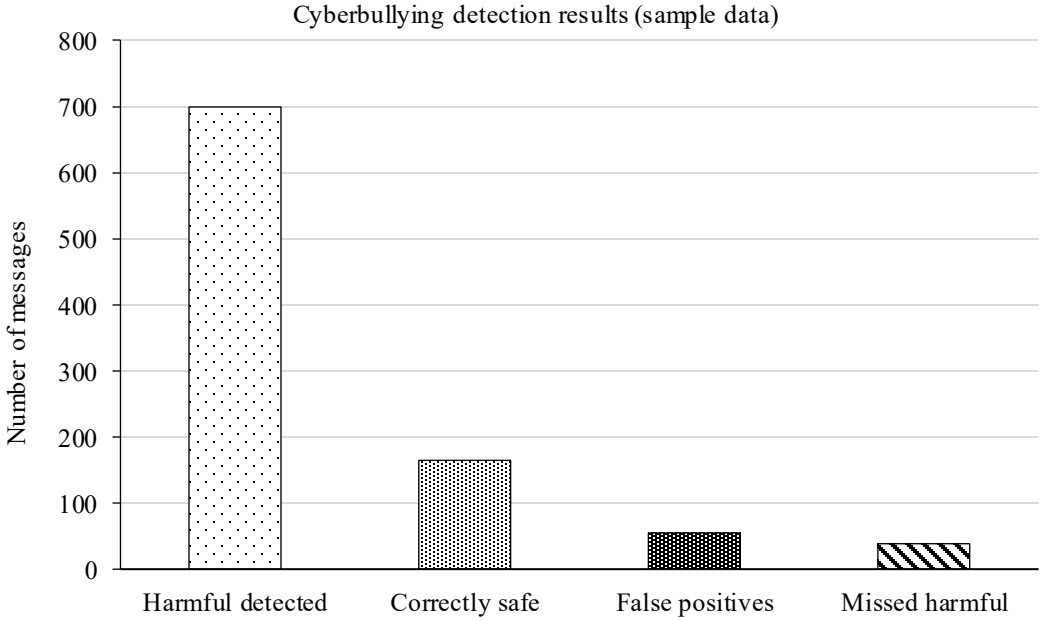

Figure 2. Cyberbullying detection results.

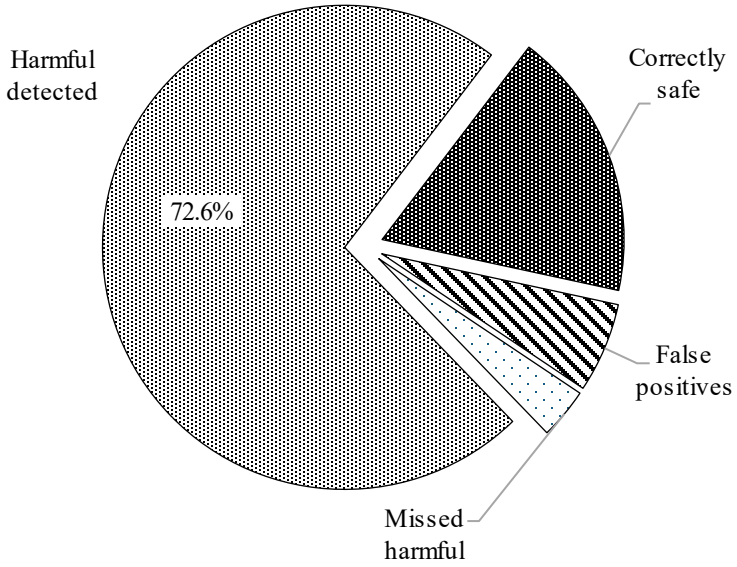

Figure 3. Distribution of detection outcomes.

From the experimental results, the effectiveness and usability of the browser-based model can be extended in relation to the precision level and F1 score relative to the proposed transformer model on the server side.

Moreover, the model performs better in relation to returning the response in time, thus being well-suited for real-time applications without affecting the operation of social media platforms. Finally, all computations are performed locally without the need to transmit sensitive user data to external servers, thereby improving trust and ethical conformance.

Although these results are somewhat promising, further development of this system is required. One major limitation is that multilingual support is very scarce at this point, and the model can be fine-tuned on larger and more diverse datasets to improve performance.

Additionally, the focus of the system is on cyberbullying represented through text; however, multimodal approaches that consider images, emojis, and voice-based content remain beyond its consideration.

Future plans include expanding support for multiple languages and code-mixed languages. There is potential to expand the system to support the detection of cyberbullying through multiple modes, including imagery and videos. Additionally, another potential expansion involves enabling the system to dynamically improve its efficiency through learning. The system should be tested on other browsers.

REFERENCES

1. Hosseinmardi H, Mattson SA, Rafiq RI, Han R, Lv Q, Mishra S. Detection of cyberbullying incidents on the Instagram social network [Preprint]. 2015 Mar 12. arXiv:1503.03909. doi:10.48550/arXiv.1503.03909.
2. Hosseinmardi H, Mattson SA, Rafiq RI, Han R, Lv Q, Mishra S. Prediction of cyberbullying incidents on the Instagram social network. [Preprint]. 2015. arXiv:1508.06257. doi:10.48550/arXiv.1508.06257.
3. Liu P, Guberman J, Hemphill L, Culotta A. Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. Proc Int AAAI Conf Web Soc Media. 2018;12(1). doi:10.1609/icwsm.v12i1.15022.

4. Balakrishnan V, Ng SK. Personality and emotion based cyberbullying detection on YouTube using ensemble classifiers. Behav Inf Technol. 2023;42(13):2296–2307. doi:10.1080/0144929X.2022.2116599.
5. Ashraf N, Zubiaga A, Gelbukh A. Abusive language detection in YouTube comments leveraging replies as conversational context. PeerJ Comput Sci. 2021;7:e742. doi:10.7717/peerj-cs.742. PubMed:34712802.
6. Chatzakou D, Leontiadis I, Blackburn J, De Cristofaro E, Stringhini G, Vakali A, et al. Detecting cyberbullying and cyberaggression in social media. ACM Trans Web. 2019;13(3):1–51. doi:10.1145/3343484.
7. Nahar V, Li X, Zhang HL, Pang C. Detecting cyberbullying in social networks using multi-agent system. Web Intell Agent Syst. 2014;12(4):375–388. doi:10.3233/WIA-140301.
8. Pawar R, Raje RR. Multilingual cyberbullying detection system. 2019 IEEE International Conference on Electro Information Technology (EIT), Brookings, SD, USA. p. 40–44. doi:10.1109/EIT.2019.8833846.
9. El-Alami F, Ouatik El Alaoui S, En Nahnahi N. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. J King Saud Univ Comput Inf Sci. 2022;34(8 Pt B):6048–6056. doi:10.1016/j.jksuci.2021.07.013.
10. Fitro A, Pinem APR, Bachri OS, Chartini. Cyberbullying detection on Instagram using IndoBERTa model. Proceedings of the International Conference on Digital Business Innovation and Technology Management (ICONBIT). 2025;1(2):1441–1447.
11. Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, Verhoeven B, et al. Automatic detection of cyberbullying in social media text. PLoS One. 2018;13(10):e0203794. doi:10.1371/journal.pone.0203794. PubMed:30296299.
12. Kumar C, Gowda C, GC BD, Yadav LPD. AI powered multimodal content moderation for online safety in social media platforms. 2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India. 2025. p. 1748–1758. doi:10.1109/ICIRCA65293.2025.11089733.
13. Facklasur Rahaman M, Golam M, Raihan Subhan M, Tuli EA, Kim DS, Lee JM. Meta-governance: Blockchain-driven metaverse platform for mitigating misbehavior using smart contract and AI. IEEE Trans Netw Serv Manag. 2024;21(4):4024–4038. doi:10.1109/TNSM.2024.3419151.
14. Rajeshwari BS, Divya I. Combating cyberbullying with generative AI: Large language models (LLMs) in cyberbullying prevention. In: Reddy CKK, Malhotra M, Ouaissa M, Hanafiah MM, Shuaib M, editors. Combating Cyberbullying with Generative AI. Hershey (PA): IGI Global; 2025. p. 329–364. doi:10.4018/979-8-3373-0543-1.ch012.
15. Nitya Harshitha TN, Prabu M, Suganya E, Sountharrajan S, Bavirisetti DP, Gadde N, et al. ProTect: A hybrid deep learning model for proactive detection of cyberbullying on social media. Front Artif Intell. 2024;7:1269366. doi:10.3389/frai.2024.1269366. PubMed:38510470.
16. Gupta A, Matta P, Pant B. An integrated user matching framework for cross-platform cyberbullying detection. Discov Comput. 2025;28:224. doi:10.1007/s10791-025-09743-7.

