Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 2, 2026

CROSS-LINGUAL CYBER-BULLYING: A SYSTEMATIC REVIEW OF DETECTION METHODS

Faheem Abbas[1*], Ali Sufyan[2], Aurangzaeb Khan[3], Muhammad Zain-ul-Abdeen[4], Sundas Amin[5]

*[*1]Department of Information and Communication Engineering, Islamia University of Bahawalpur, Bahawalpur, Pakistan*

[2]Department of Information and Communication Engineering, Islamia University of Bahawalpur, Bahawalpur, Pakistan

[3]Department of Information and Communication Engineering, Islamia University of Bahawalpur, Bahawalpur, Pakistan

[4]Department of Information and Communication Engineering, Islamia University of Bahawalpur, Bahawalpur, Pakistan

[5]Department of Information Security, Islamia University of Bahawalpur, Bahawalpur, Pakistan

*[*1]fahymabbas@gmail.com*

DOI:

Keywords
Cyber-bullying, Roman Urdu, Sentiment Analysis, Emoji Interpretation, NLP, Machine Learning, Deep Learning, Multilingual Transformers.

Article History
Received on 10 Jan, 2026
Accepted on 11 Feb, 2026
Published on 12 Feb, 2026

Copyright @Author

Corresponding Author:
Faheem Abbas

Abstract

The growing use of social media has led to a significant rise in cyber-bullying and hate speech, creating serious social and mental health challenges worldwide.Consequently, numerous automated detection methods have been developed, but their performance varies widely across languages, datasets, and modeling strategies. This paper reviews existing literature on state-of-the-art approaches to cyber-bullying and hate speech detection, with particular emphasis on multilingual and low-resource language settings such as Roman Urdu and English. The reviewed studies are analyzed across several dimensions, including dataset characteristics, preprocessing methods, and feature engineering techniques, followed by an evaluation of machine learning, deep learning, and transformer-based models. The findings indicate that traditional machine learning models provide a strong baseline but struggle with contextual and intent-aware detection. Deep learning approaches achieve improved performance, yet these approaches are still limited by data scarcity and dependence on binary classification. While transformer-based models demonstrate state-of-the-art performance, they struggle with emoji-aware processing, slang interpretation, and differentiating playful teasing from harmful cyber-bullying. By identifying key research gaps, this review underscores the importance of multilingual, emoji-aware, and intent-sensitive cyber-bullying detection frameworks, supporting further research and practical moderation systems.

I. Introduction

The high growth rate in the use of the social media platforms has transformed the way people interact, communicate, and air their views and express feelings in the new digital world [1]. As noted in recent systematic reviews, [2] while these platforms have enabled global connectivity and freedom of expression, they have also facilitated the growth of several harmful online behaviours, such as cyber-bullying, hate speech and offensive statements [3], [4].

Cyber-bullying entails the deliberate, repeated use of digital platforms, social media, messaging applications, or online forums to harass, threaten, or humiliate individuals [5], [6]. Current literature emphasizes that this phenomenon often involves a power imbalance and can result in serious psycho- logical, emotional, and social harm to the victim, distinct from traditional bullying due to the anonymity and infinite reach of the internet [7], [8]. The hate speech can be described as a means of communication that insults, intimidates, or incites violence against other individuals or groups based on their particular characteristics: religion, race, ethnicity, gender, or sexual orientation [9], [10]. Such materials propagate discrimi- nation and have the potential to further aggravate social tension to the point of violence. The offensive refers to the expres- sions that use insulting, profane, or otherwise inappropriate language, potentially causing discomfort or emotional distress without targeting protected groups [11]. While not necessarily fitting into the category of hate speech, offensive speech can also contribute to online toxicity. [12], [13]

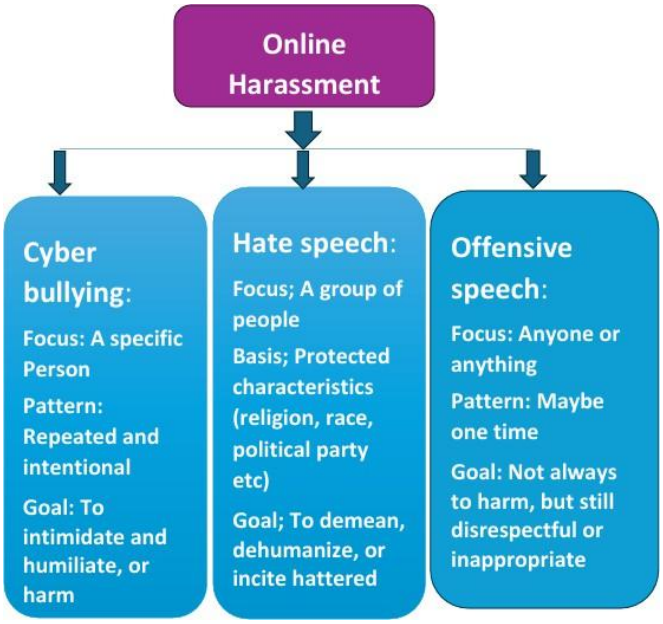

Figure 1: Taxonomies of online harm

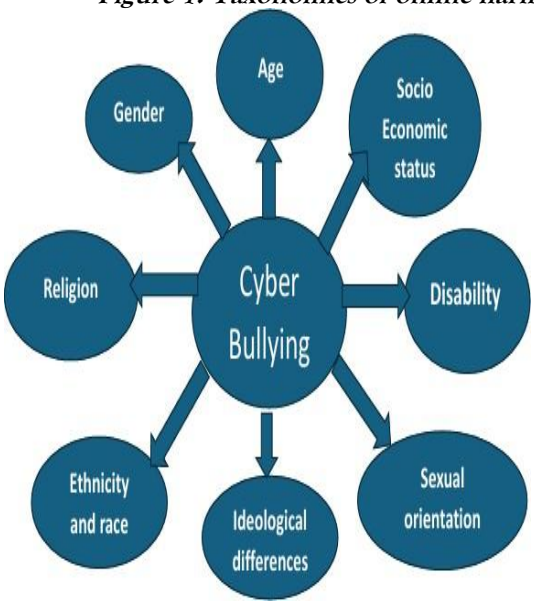

Figure 2: Common behavioral factors used to identify cyber- bullying on digital platforms

Cyber-bullying is informed by a set of demographic, social, and ideological factors that incorporate online interaction and user behaviour [11]. The factors in Figure 2 represent the common characteristics that increase an individual's vulnera- bility to online harassment and abusive behaviour [14]. The age factor in cyber victimization tends to be high among adolescents and young adults. Heightened social media use, peer pressure, and emotional sensitivity during these periods of development further increase exposure to online harassment [15]. Another critical factor is gender, since individuals may be cyber bullied based on gender roles and societal stereotypes. While females are more often victims of emotional abuse, body-shaming, and harassment [16], men are more likely to be victims of aggressive or threatening cyber-bullying [17]. Religion is often used as a pretext for cyber-bullying ex- pressed as hate speech and discriminatory content. People from minority or visually distinct religious groups are especially targeted for mass online attacks because of the intolerance and discrimination rooted in religious belief [18]. In fact, ethnicity and race are commonly used in incidents of cyber-bullying to attack an individual or community using racial slurs, stereotypes, and exclusionary language. Such attacks reinforce social inequality and contribute to hostile online environments [19]. Socioeconomic status may impact instances of cyber- bullying where there is targeting based on perceived wealth, social class, or economic background. Harassment in this context may most often come through mocking, exclusion, or derogatory remarks concerning financial standing. Disability can be considered a critical vulnerability factor, since the majority of the time people with physical, cognitive, or mental disabilities are targeted by bullying due to perceived difference or dependency. Cyber-bullying associated with disability may have serious consequences for self-esteem and psychological well-being [20]. Sexual orientation is one of the most com- mon reasons for cyber-bullying [21].LGBTQ+ individuals face hate speech, harassment, and threats. Online platforms are very often turned into places where discriminatory attitudes regarding sexual identity are loudly expressed [22]. Cyber- bullying may also be based on ideological differences, such as political beliefs, cultural values, or personal opinions. The disagreement over ideology escalates into hostile interactions that use abusive language and targeted harassment in an online discussion [23].

Cyber-bullying and harassment have emerged as significant public health problems, greatly impacting the psychological and emotional state of people, and the youth, in particular [24]. Unlike the occurrence of bullying, which takes place offline, cyber-bullying takes place through digital platforms, social media networks, and messaging platforms, with the ability to share harmful and offensive material instantly and anonymously [25], [26]. The different manifestations of this kind of aggression consist of harassment, hate speech, threats, humiliation, and exclusion, among others, which cause severe effects to the victim, including feelings of anxiety and depres- sion, social isolation, and extreme instances of self-infliction of injury and suicidal tendencies among the affected young individuals [27].

Institute for Excellence in Education & Research

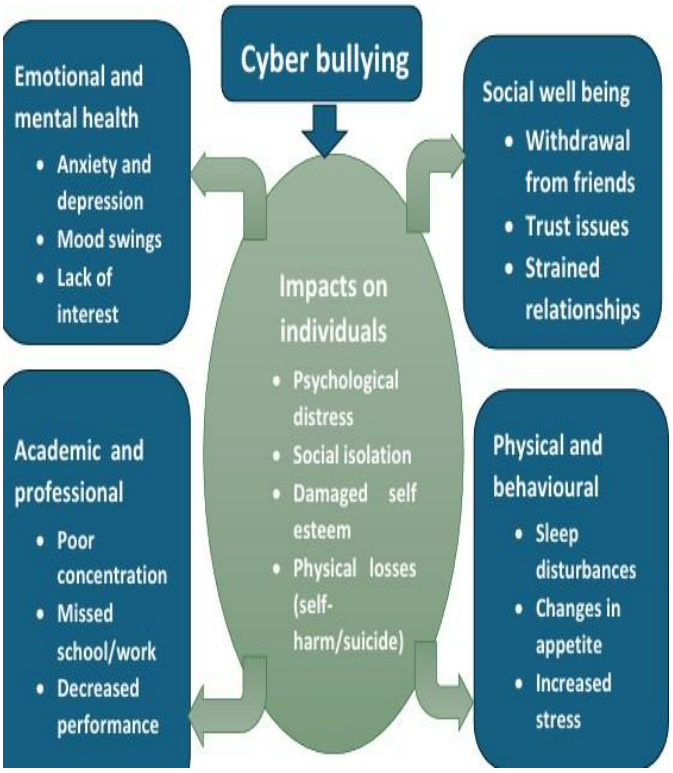

Figure 3: The impacts of cyber-bullying on individuals The impacts of cyber-bullying on individuals include emo-tional and mental health problems, academic and professional life's distortions, social well being issues, physical and be-havioral losses [28]. These phenomena have become major social challenges, particularly affecting adolescents and young adults, and are associated with severe psychological conse- quences, including anxiety, depression, social withdrawal, and in extreme cases, self-harm and suicide [29]. The study in [30], demonstrates that cyber-bullying significantly increases the risk of depression and suicidal ideation among adolescents in India. Victims were found to be more than twice as likely to experience adverse mental health outcomes, highlighting the need for early automated detection systems.

With the exponential growth of user-generated content on social media, manual moderation for abusive content is no longer practical [31]. In this context, automated detection systems for cyber-bullying and hate speech based on NLP and machine learning have recently gained significant attention from the research community [32]. Most early approaches relied on traditional machine learning techniques using hand- crafted features such as Bag-of-Words and TF-IDF. While these methods achieved reasonable results on structured En- glish datasets, they poorly capture contextual meaning, implicit intent, and emotional nuances that are characteristic of real- world social media text [33], [34].

This problem is highly exacerbated in multilingual and low-resource language settings [11], [25]. Social media conversations by the online community of South Asia, especially in Pakistan, are mostly in Roman Urdu or English or both in a single sentence. Roman Urdu is an informal, non-standardized language that presents high variation of spelling, slang usage, and phonetic ways of writing, which poses difficulty for tradi- tional NLP pipelines [35]. A few existing cyber-bullying de- tection systems are designed for English and do not generalize well to Roman Urdu or code-mixed text, reducing the accuracy of detection and making predictions unreliable [26]. Roman Urdu poses some challenges from the point of view of compu- tation. For instance, it can be defined as the Urdu script written with the English alphabets with non-standardized spellings that lack overall annotated corpora or language support [36]. It can further be noted that Roman Urdu is resource-scarce with a lack of support from the perspective of languages such that the approach itself may be inefficient because it may rely upon English language standards that can face challenges from the viewpoint of social media [**?**]. Another critical challenge in cyber-bullying detection is the ever-increasing usage of emojis during online communication. Emojis are crucial to express emotions, sarcasm, and intent that may not have been explicitly expressed by using text only [37]. The same textual content can have drastically different meanings depending on the accompanying emoji. For example, an insulting phrase combined with a laughing emoji could mean friendly teasing, while the same phase with an angry emoji can represent aggressive cyber-bullying [38], [39]. However, most of the existing detection systems either ignore emojis or consider them as noise, hence miss

classifying and performing poor intent recognition. This systematic review of the literature critically discusses key research trends in the detection of cyber-bullying and hate speech on social networks, placing special emphasis on multilingual and low-resource language settings such as Roman Urdu with English. Further, this review discusses prior studies with regard to dataset characteristics, preprocessing and feature engineering techniques, and model performance in terms of traditional machine learning, deep learning, and transformer-based approaches. This study further identifies strengths, limitations, and performance trends across the different models in an attempt to highlight some major research gaps related to;

- Emoji-aware sentiment modeling.
- Intent-sensitive classification.
- Distinguish between friendly teasing and cyber-bullying. This review also lays down a structured framework for the proposed methodology and justifies the need for an emoji-

aware, multilingual cyber-bullying detection framework.

The rest of the article is organized into several sections. Section 2 discusses the datasets used in this study, while Section 3 summarizes data preprocessing and feature engineering techniques. Machine learning, deep learning, and transformer-based approaches are presented in Sections 4, 5, and 6, respectively. Moreover, the limitations and directions for future work are outlined in Section 7. Finally, the article is concluded in Section 8.

II. Dataset Collection

The efficiency of cyber-bullying detection systems is greatly dependent on the quality, size, and linguistic diversity of the datasets on which the systems have already been trained and tested. Most previous works rely on social media platforms such as Twitter, Instagram, and Facebook since they are pub- licly available and contain a high volume of abusive content [5], [40], [41]. The early work focused on English datasets, recent trends from 2020 onwards have emphasized the creation of larger, more diverse benchmarks [33], [42], [43]. These datasets usually range from thousands to millions of instances and often come labelled under binary schemes, such as cyber- bullying versus non cyber-bullying, or hate versus non-hate. By contrast, Roman Urdu and other low-resource South Asian language datasets are scarce, mostly custom-developed by researchers themselves [11], [25], [36]. Roman Urdu datasets are typically of a limited size and display more linguistic variability because of non-standardization of spelling and grammatical constructs [26]. In [44], contemporaries typically created the Roman Urdu dataset by collecting social media posts and manually annotating them into categories related to bullying. These datasets allow for language-specific modeling but, due to their limited size and domain, model generalization is hindered.

Recent works have attempted to further alleviate the issue of linguistic diversity by introducing multilingual and code-mixed datasets that combine Urdu, Roman Urdu, and English content [35]. These studies showed clearly that the multi- lingual datasets reflect more accurately the real patterns of social media communication in which users often use multiple languages within a single post. These datasets also remain heavily dependent on binary or coarse-grained labels that reduce their capacity for capturing subtle variations in intent, such as friendly teasing.

Table I: *Comparison of Datasets Used in Reviewed Studies*

Ref	Languages			Platform				Label Type		Emoji	Source		Dataset Size
	Urdu	Roman	English	Twitter	Instagram	Facebook	Wikipedia	Binary	Multi		Custom	Public	
[35]	✓	✓	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗	- 18,000
[45]	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗	- 8,000
[46]	✗	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗	✓	- 47k (DS1) - - 100k (DS2)
[47]	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	- 214k (Wiki), - 25k (Twitter)
[48]	✗	✓	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗	- 9,000
[49]	✗	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	- 15,000
[50]	✗	✗	✓	✓	✗	✗	✗	✓	✗	✓	✓	✗	- 12,000
[51]	✗	✗	✓	✓	✗	✗	✗	✗	✓	✗	✓	✗	- 20,000
[44]	✗	✓	✗	✓	✗	✓	✗	✓	✗	✗	✓	✗	30,000
[18]	✗	✓	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗	- 10,000
[52]	✗	✓	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗	- 12,000
[53]	✗	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	- 12,000
[54]	✗	✗	✓	✓	✗	✓	✗	✓	✗	✗	✗	✓	- 7,000
[55]	✗	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗	10,000
[56]	✗	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	- 20,000

IIER
Institute for Excellence in Education & Research

The important limitation that is observed in nearly all datasets is the absence of emoji-aware annotation [57].Even though the use of emojis to communicate feeling and sarcasm is becoming more popular, a limited number of people take explicit measures to consider emojis when construction of datasets took place, which is rather odd considering the impor- tance of body language in communication [58]. Others, during preprocessing, either delete all emojis or ignore them and therefore miss out on much contextual and emotional meaning altogether.Overall, the data analysis of the datasets indicates that there is a considerable unequal balance between resource- abundant datasets of English and low resource Roman Urdu datasets. Hence, the necessity to have multi-language, emoji- based, and fine-tuned annotated corpora is experienced. [38], [39], [59].

III. Data Preprocessing And Feature Engineering

Data preprocessing is a crucial task that enhances performance in cyber-bullying detection models, this becomes critical when dealing with noisy and informal social media text [60]. Preprocessing techniques such as lowercasing, tokenization, URL and user mention removals, punctuation removal, and stop words elimination were shared across most of the studies [51]. These techniques aim to reduce noise and standardize textual input before feature extraction.

It gets a lot trickier to preprocess Roman Urdu text since there are spelling inconsistencies and phonetic variations. Many works in Roman Urdu use normalization, in which a set of spelling variants is mapped into a canonical form using handcrafted dictionaries or rule-based mappings. As pointed in [52], normalization can drastically reduce vocabulary spar- sity and improve the performance of deep learning models. However, these techniques are language-specific and are not uniformly applied in different multilingual studies.

The feature engineering approaches, in the literature, broadly include traditional feature-based methods and repre- sentation learning approaches. Most of the earlier machine learning-based studies primarily rely on BoW, TF-IDF, and n-gram features [61]. These features capture word frequency information and fail to capture contextual semantics. These features perform reasonably well with simple classifiers like logistic regression and SVM, however, they cannot capture intent, sarcasm, and implicit aggression.

Later research, therefore, incorporated distributed word embeddings to represent words in dense vectors using semantic similarity [62]. Compared to BoW and TF-IDF, these repre- sentations improve the understanding of context but are still at the word level and full context at sentence level is missing. More recent works utilize contextual embeddings from transformer-based models such as BERT, mBERT and RoBERTa and avoid explicit feature engineering [63]. Recent studies like [47] and [53] have demonstrated the superiority of these architectures.BERT representations were further augmented with emotion and sentiment features extracted using lexicon-based methods, resulting in improved recall and F1- score [47]. However, in this work, emojis were removed as a part of preprocessing, thus rendering the model incapable to capture sentiment from emojis. Throughout the reviewed liter- ature, considerations of slang and emoji-aware preprocessing have been largely neglected. Most research removes emojis and informal abbreviations rather than explicitly modeling them [45]. Such a practice leads to the loss of information, especially in deciding between friendly teasing and actual cyber-bullying. Thus, the currently available preprocessing and feature engineering pipelines are not good enough to provide intent-aware and emotion-sensitive cyber-bullying detection in real-world social media settings.

IV. Machine Learning Approaches

Early cyber-bullying detection research was predominantly done using traditional machine learning algorithms because they are simpler, more interpretable, and computationally inexpensive compared to deep learning algorithms [34], [64]. The most standard classifiers for cyber-bullying detection are logistic regression, naive bayes, SVM, and random forests, while standard features extracted are Bag-of-words and TF- IDF representations [54].

Due to their efficiency and ease of implementation, the baseline models based on logistic regression and naive bayes have been widely adopted [65]. Indeed, various studies such as [66] show that combined with TF-IDF features, these models obtain a reasonable accuracy for datasets in English.

Table II: *Comparison of Preprocessing and Feature Engineering Techniques*

Ref	Lower	Token	StopWord	RU Norm	Slang	Emoji	Feature Representation
[35]	✓	✓	✓	×	×	×	m-BERT, MuRIL
[45]	✓	✓	✓	×	×	×	TF-IDF
[46]	✓	✓	✓	×	×	×	TF-IDF, BERT Embeddings
[47]	✓	✓	✓	×	×	×	EDM, Emotion, Sentiment, Lexicons
[48]	✓	✓	✓	✓	✓	×	TF-IDF
[49]	✓	✓	✓	×	×	×	TF-IDF + LIWC

[50]	✓	✓	✓	×	×	✓	TF-IDF + Deep Learning
[51]	✓	✓	✓	×	×	×	CountVectorizer
[44]	✓	✓	✓	✓	×	×	Word2Vec
[18]	✓	✓	✓	✓	×	×	FastText
[52]	✓	✓	✓	✓	✓	×	TF-IDF
[53]	✓	✓	✓	×	×	×	GloVe + RoBERTa
[54]	✓	✓	✓	×	×	×	Bag-of-Words, TF-IDF
[55]	✓	✓	✓	×	×	×	Word Embeddings
[56]	✓	✓	✓	×	×	×	Embeddings

However, their performance degrades in the presence of contextual ambiguity, sarcasm, and informal language that is common in cyber-bullying scenarios. These models completely rely on surface-level word frequencies and are unable to capture sequential dependencies or semantic intent [13], [67], [68].

In contrast, support vector machines have performed considerably better, especially for high-dimensional feature space [69]. Researchers continue to use SVM as a strong baseline for comparison against newer deep learning methods [49]. How- ever, approaches based on SVM also need to do extensive fea- ture engineering and do not scale well or adapt to multilingual or code-mixed text. Similarly, the Random Forest model is more robust to noise, but also suffers from similar limitations because of reliance on handcrafted features [70]. In fact, all Roman Urdu-specific studies rely on aggressive preprocessing and normalization in order to achieve acceptable results with classical ML models [71]. Hyperparameter optimization was shown to improve the performance of machine learning meth- ods for Roman Urdu hate-speech detection [71].However, even these improved models remain behind on implicit aggression, friendly teasing, and generally emotionally subtle text. Overall, machine learning models are useful baselines, however, their inability to model context and intent limits their effectiveness for fine-grained cyber-bullying detection [72].

V. Deep Learning Approaches

To overcome these limitations of typical machine learning methods, there was an increasing tendency for researchers to turn their attention to the use of deep learning models able to extract contextual and sequential information [73], [74]. Popular architectures include Convolutional Neural Networks, Long Short-Term Memory, Bidirectional LSTM, and various hybrid configurations [73]. The models developed using CNN are doing well in explicit abusive keyword detection and also local textual patterns [75].In [76] CNN architectures are applied on Roman Urdu microtext and outperform the state-of-the-art traditional ML classifiers. On the downside, the CNNs capture mainly local n-gram features, failing to model long-range dependencies, which are essential in the case of capturing implicit bullying or sarcasm.

Models based on LSTMs extend the capability of CNNs by modeling sequential dependencies in the text [77]–[80]. Indeed, several works established that LSTMs outperform other traditional ML and CNN methods, especially for longer sentences and conversational data [52]. BiLSTM models fur- ther increase performance by processing text in both for- ward and backward directions to capture better contextual understanding. In [18],a CNN–BiLSTM hybrid architecture is utilized for the detection of hate speech in Roman Urdu and achieved superior F1-scores compared to the performance of standalone models.

Despite these advances, there are still a number of challenges with deep learning models, the relatively large modeling datasets needed, the tendency of overfitting in low-resource scenarios, and usually, the interpretability of such systems [46]. Most deep learning–based studies formulate cyber-bullying detection as a binary classification problem, failing to distinguish friendly teasing from harmful cyber-bullying [48]. These limitations then motivated shifting to transformer-based architectures [43], [81].

VI. Transformer Based Approaches

Transformer-based models represent the current state of the art in the detection of cyber-bullying and hate speech because they capture global contextual relationships through self-attention mechanisms [33], [82]. Performances by BERT, mBERT, XLM-RoBERTa, and RoBERTaNET are at the top among various other datasets and languages [53], [83].

In contrast, most of the BERT-based models outperform the traditional approaches in ML and deep learning significantly by extracting fine-grained semantic and syntactic information [84]. In [44] a context-aware deep learning model involving BERT was developed to detect hate speech for Roman Urdu and demonstrated significant improvements in performancecompared to CNN and LSTM baselines. The multilingual versions, including mBERT and XLM-RoBERTa, make use of cross-lingual knowledge to perform better and thereby become suitable for code-mixed Urdu-English text as well [35].

Table III: *Evaluation Metrics of Best-Performing Models*

Ref	Models Investigated	Best Model	Accuracy (%)	Precision	Recall	F1-score
[35]	m-BERT, MuRIL	MuRIL	92.0	0.93	0.92	0.92
[45]	Random Forest	RF	88.0	0.89	0.94	0.91
[46]	MLP, SVM, RF + Neutrosophic	MLP + Neutrosophic	95.0	Not discussed	Not discussed	Not discussed
[47]	BERT, XLNet + EDM	BERT + EDM	96–97	0.97	0.98	0.97
[48]	SVM, RF	SVM	86.4	0.85	0.84	0.84
[49]	SVM + NLP	SVM	93.2	0.93	0.94	0.94
[50]	LSTM, Bi-LSTM, GRU	Bi-LSTM	90.0	0.89	0.88	0.89
[51]	Ensemble (RF + LR + DT)	Ensemble	94.7	Not discussed	Not discussed	Not discussed
[44]	CNN, LSTM, Bi-LSTM +Attention	Bi-LSTM + Attention	87.5	0.88	0.89	0.89
[18]	CNN-BiLSTM	CNN-BiLSTM	80.7	0.81	0.82	0.81
[52]	CNN, RNN-LSTM	RNN-LSTM	85.5	0.72	0.68	0.70
[53]	RF, CNN, RoBERTaNet	RoBERTaNet	95.0	0.95	0.97	0.96
[54]	NB, SVM, RF	SVM	78.2	0.77	0.75	0.76
[55]	RNN, Bi-LSTM, DEA-RNN	DEA-RNN	90.5	0.90	0.89	0.89
[56]	LSTM	LSTM	87.6	0.89	0.88	0.88

In [53], RoBERTaNET was proposed as an enhanced RoBERTa-based model augmented with GloVe features, achieving state-of-the-art performance on cyber-bullying datasets. Similarly, [47] combined BERT with an Emotion Detection Model (EDM) and sentiment features, resulting in high recall and F1-scores through explicit modeling of emotional cues. However, transformer-based models have lim- itations, too. Most of the current research removes emojis during preprocessing or treats them as noise, thus limiting emotional expressiveness. Besides, transformer models are usually evaluated using binary classification schemes, not accounting for friendly teasing or the ambiguity of intent [50]. Thus, high scores reported in the literature may not reflect the real effectiveness of moderation in real life.

A comparative analysis of previous studies shows a clear performance hierarchy among different modeling paradigms. Traditional machine learning models perform adequately on clean, English-only datasets but fail to capture contextual and emotional nuances. Deep learning models, particularly BiLSTM-based and hybrid architectures, offer improved con- textual understanding but remain bound by data requirements and binary modeling strategies. Transformer models consis- tently yield the best overall performances across languages and data because of the contextual modeling. Indeed, such models correspond to the most successful attempts to date towards detecting cyber-bullying [44], [53]. Still, they lack emoji-aware processing and intent-sensitive classification, and therefore, they also represent an important gap.

In general, the current state of research shows that, while transformer models are clearly the top-performing architecture for cyber-bullying detection, their true powers are not harnessed as a result of weaknesses in dataset design, pre-processing strategy, and granularity of classification. These observations naturally follow from the need for a multilingual, emoji-aware, intent-sensitive cyber-bullying detection frame- work, proposed herein.

VII. Limitations And Future Work

Despite the significant progress made in cyber-bullying and hate speech detection, several limitations persist in existing research that hinder the effectiveness and real-world applicability of current systems. These limitations mainly arise from linguistic bias, insufficient modeling of emotional cues, simplified classification schemes, and evaluation practices that do not fully reflect deployment challenges.

The key limitations are summarized as follows;

- The major portion of available cyber-bullying detection systems focuses on English language datasets, with lim- ited focus on Roman Urdu and code-mixed text.
- Emoji-aware sentiment and slang handling are largely ignored, despite their importance for the conveyance of emotion and intent in social media communication.
- Most of the studies are based on binary classification, which does not discriminate between friendly teasing and actual cyber-bullying.
- Current models lack intent-aware mechanisms and often misclassify emotionally expressive but non-harmful con- tent.
- Usually, datasets are small, imbalanced, and non-standardized, which restricts model generalization.
- Evaluation methods emphasize accuracy with less em- phasis on practical applicability and moderation require- ments.

Future research directions should aim to address these challenges through following efforts;

- Development of multilingual and emoji-aware datasets.

- Intent-aware and fine-grained classification.
- Enhanced contextual and emotional modeling.
- Improved evaluation frameworks.

Addressing these limitations will be essential for developing more accurate, robust, and socially responsible cyber-bullying detection systems that can operate effectively in multilingual and real-world social media environments.

VIII. Conclusion

This review explains the current literature that has been done on the identification of cyber-bullying and hate speech on social media platforms, their datasets, preprocessing, feature engineering, and performance of diverse models.Traditional models of machine learning like the logistic regression, naive bayes and support vector machines did well on English data sets but were very reliant on the manual features engineering and just did not understand the context.More sophisticated deep learning architectures like CNNs, LSTMs, BiLSTMs and hybrid models showed superior contextualization but require a large quantity of labeled data and are usually limited to binary classification problems.Further enhanced transformer models include BERT, mBERT, XLM-RoBERTa and RoBER- TaNET to perform extremely well in different languages with both semantic and contextual knowledge.Nevertheless, several challenges remain, including the scarcity of datasets for low-resource languages such as Roman Urdu, limited representation of emojis and slang in existing datasets, and the difficulty of distinguishing harmful communication from playful or teasing interactions.

References

[1] J. Chun, J. Lee, J. Kim, and S. Lee, "An international systematic review of cyberbullying measurements," *Computers in human behavior*, vol. 113, p. 106485, 2020.

[2] F. Alkomah and X. Ma, "A literature review of textual hate speech detection methods and datasets," *Information*, vol. 13, no. 6, p. 273, 2022.

[3] S. Chakraborty, A. Bhattacherjee, and A. Onuchowska, "Cyberbullying: A review of the literature," *Available at SSRN 3799920*, 2021.

[4] J. Jeswani, V. Bhardwaj, B. Jain, and B. S. Kohli, "Negative sentiment analysis: hate speech detection and cyber bullying," *Int. J. Res. Appl. Sci. Eng. Technol*, vol. 10, no. 5, pp. 911–917, 2022.

[5] A. C. Roy, T. Mahmud, and T. Abrar, "A multi-class cyberbullying classification on image and text in code-mixed bangla-english social media content," *Natural Language Processing Journal*, p. 100191, 2025.

[6] G. W. Giumetti and R. M. Kowalski, "Cyberbullying via social media and well-being," *Current opinion in psychology*, vol. 45, p. 101314, 2022.

[7] M. Mladenovic´, V. Os˘mjanski, and S. V. Stankovic´, "Cyber-aggression, cyber bullying, and cyber-grooming: A survey and research challenges," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–42, 2021.

[8] Y. Wang, J. Cai, C. Wang, Y.-F. Mu, Z.-Y. Deng, A.-P. Deng, H.-J. Song,
Y. Huang, L. Yin, W. Zhang *et al.*, "The prevalence and association of traditional bullying and cyber bullying with mental health among adolescent and youth students in china: a study after the lifting of covid- 19 restrictions," *BMC public health*, vol. 25, no. 1, p. 618, 2025.

[9] D. Sharma *et al.*, "Hate speech detection research in south asian languages: A survey of tasks, datasets and methods," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 3, pp. 1–44, 2025.

[10] A. Gandhi, P. Ahir, K. Adhvaryu, P. Shah, R. Lohiya, E. Cambria,
S. Poria, and A. Hussain, "Hate speech detection: A comprehensive review of recent works," *Expert Systems*, vol. 41, no. 8, p. e13562, 2024.

[11] A. A. Khan, M. H. Iqbal, S. Nisar, A. Ahmad, and W. Iqbal, "Offensive language detection for low resource language using deep sequence model," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 5210–5218, 2023.

[12] L. Cheng, Y. N. Silva, D. Hall, and H. Liu, "Session-based cyberbullying detection: Problems and challenges," *IEEE Internet Computing*, vol. 25, no. 2, pp. 66–72, 2020.

[13] A. Haider, A. B. Siddique, R. H. Ali, M. Imad, A. Z. Ijaz, U. Arshad,
N. Ali, M. Saleem, and N. Shahzadi, "Detecting cyberbullying using machine learning approaches," in *2023 International Conference on IT and Industrial Technologies (ICIT)*. IEEE, 2023, pp. 1–6.

[14] K. Rudnicki *et al.*, "Systematic review of determinants and consequences of bystander interventions in online hate and cyber bullying among adults," *Behaviour & Information Technology*, vol. 42, no. 5, pp. 527– 544, 2023.

[15] C. Zhu, S. Huang, R. Evans, and W. Zhang, "Cyberbullying among adolescents and children: a comprehensive review of the global situation, risk factors, and preventive measures," *Frontiers in Public Health*, vol. 9, p. 634909, 2021.

[16] A. S. G. Tabares, J. E. Restrepo, and G. Zapata-Lesmes, "The effect of bullying and cyberbullying on predicting suicide risk in adolescent females: The mediating role of depression," *Psychiatry research*, vol. 337, p. 115968, 2024.

[17] S.-S. Lee, S. Jun, and S. Jung, "Gender differences in the moderating effect of cyberbullying peers between cyberbullying offending and cyberbullying victimization in south korea," *Gender Issues*, vol. 42, no. 3, p. 22, 2025.

[18] M. Zohaib *et al.*, "Detecting hate speech in roman urdu using a convolutional-BiLSTM-based deep hybrid neural network," *PeerJ Com- puter Science*, vol. 11, p. e3342, 2025.

[19] N. Agustiningsih, A. Yusuf, A. Ahsan, and Q. Fanani, "The impact of bullying and cyberbullying on mental health: a systematic review," *International Journal of Public Health Science*, vol. 13, no. 2, pp. 513–520, 2024.

[20] D. T. Qamar, D. H. H. Ali, D. M. J. Aftab, D. M. N. Iqbal, M. S. Zaman, and S. Anser, "Effects of cyberbullying on the mental health of students with disabilities in punjab," *Journal of positive school psychology*, vol. 7, no. 6, pp. 809–823, 2023.

[21] F. Angela, R.-d. Maria-Luisa, N. Annalaura, and M. Ersilia, "Online sexual harassment in adolescence: a scoping review," *Sexuality Research and Social Policy*, vol. 21, no. 4, pp. 1480–1499, 2024.

[22] M. Garaigordobil and E. Larrain, "Bullying and cyberbullying in lgbt adolescents: Prevalence and effects on mental health." *Comunicar: Media Education Research Journal*, vol. 28, no. 62, pp. 77–87, 2020.

[23] O. F. Malik and S. Pichler, "Linking perceived organizational politics to workplace cyberbullying perpetration: the role of anger and fear," *Journal of Business Ethics*, p. 1, 2022.

[24] S. I. Ali and N. B. Shahbuddin, "The relationship between cyberbullying and mental health among university students," *Sustainability*, vol. 14, no. 11, p. 6881, 2022.

[25] S. Hussain, M. S. I. Malik, and N. Masood, "Identification of offen- sive language in urdu using semantic and embedding models," *PeerJ Computer Science*, vol. 8, p. e1169, 2022.

[26] A. Mehmood, M. S. Farooq, A. Naseem, F. Rustam, M. G. Villar, C. L. Rodr´ıguez, and I. Ashraf, "Threatening urdu language detection from tweets using machine learning," *Applied Sciences*, vol. 12, no. 20, p. 10342, 2022.

[27] C. Li, P. Wang, M. Martin-Moratinos, M. Bella-Ferna´ndez, and H. Blasco-Fontecilla, "Traditional bullying and cyberbullying in the digital age and its associated mental health problems in children and adolescents: a meta-analysis," *European child & adolescent psychiatry*, vol. 33, no. 9, pp. 2895–2909, 2024.

[28] N. Agustiningsih, A. Yusuf, A. Ahsan, and Q. Fanani, "The impact of bullying and cyberbullying on mental health: a systematic review," *International Journal of Public Health Science*, vol. 13, no. 2, pp. 513–520, 2024.

[29] P. Peprah, M. S. Oduro, R. Okwei, C. Adu, B. Y. Asiamah-Asare, and W. Agyemang-Duah, "Cyberbullying victimization and suicidal ideation among in-school adolescents in three countries: implications for prevention and intervention," *BMC psychiatry*, vol. 23, no. 1, p. 944, 2023.

[30] C. Maurya, T. Muhammad, P. Dhillon, and P. Maurya, "The effects of cyber bullying victimization on depression and suicidal ideation among adolescents and young adults: a three year cohort study from india," *BMC Psychiatry*, vol. 22, no. 1, p. 599, 2022.

[31] M. H. Obaid, S. K. Guirguis, and S. M. Elkaffas, "Cyberbullying detection and severity determination model," *IEEE Access*, vol. 11, pp. 97 391–97 399, 2023.

[32] Z. Mansur, N. Omar, and S. Tiun, "Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities," *IEEE Access*, vol. 11, pp. 16 226–16 249, 2023.

[33] M. Abusaqer and J. Saquer, "A comparative analysis of transformer and traditional ml models for cyberbullying detection on twitter (now x)," in *2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2025, pp. 1607–1612.

[34] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: a review," *IEEE access*, vol. 9, pp. 88 364–88 376, 2021.

[35] F. Razi and N. Ejaz, "Multilingual detection of cyber bullying in mixed urdu, roman urdu, and english social media conversations," *IEEE Access*, 2024, early Access.

[36] I. U. Khan, A. Khan, W. Khan, M. M. Su'ud, M. M. Alam, F. Subhan, and M. Z. Asghar, "A review of urdu sentiment analysis with multilin- gual perspective: A case of urdu and roman urdu language," *Computers*, vol. 11, no. 1, p. 3, 2021.

[37] H. Setyawan, "Contemporary issues in linguistics: A systematic litera- ture review on emoji and emoticon," *Elsya: Journal of English Language Studies*, 2024.

[38] K. Maity, S. Saha, and P. Bhattacharyya, "Emoji, sentiment and emo- tion aided cyberbullying detection in hinglish," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 5, pp. 2411–2420, 2022.

[39] D. Jadhav, S. Vijayalakshmi, T. S. Palathara *et al.*, "Rating-based cyberbullying detection with text, emojis on social media," in *2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT)*, vol. 1. IEEE, 2024, pp. 1–6.

[40] A. G. Philipo, D. S. Sarwatt, J. Ding, M. Daneshmand, and H. Ning, "Cyberbullying detection: Exploring datasets, technologies, and ap- proaches on social media platforms," *ACM Computing Surveys*, 2024.

[41] T. Ahmed, S. Ivan, M. Kabir, H. Mahmud, and K. Hasan, "Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 99, 2022.

[42] H. Mehta and K. Passi, "Social media hate speech detection using explainable artificial intelligence (xai)," *Algorithms*, vol. 15, no. 8, p. 291, 2022.

[43] K. Saifullah, M. I. Khan, S. Jamal, and I. H. Sarker, "Cyberbullying text identification: A deep learning and transformer-based language modeling approach," 2024.

[44] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Context-aware deep learning model for detection of roman urdu hate speech on social media platform," *IEEE Access*, vol. 10, pp. 121 133–121 151, 2022.

[45] C. Valarmathi *et al.*, "NLP-driven detection of cyber-bullying comments in instagram social network," in *2025 4th International Conference on Computing and Information Technology (ICCIT)*. IEEE, 2025.

[46] Y. M. Ibrahim, R. Essameldin, and S. M. Saad, "Social media forensics: An adaptive cyber bullying-related hate speech detection approach based on neural networks with uncertainty," *IEEE Access*, vol. 12, pp. 59 474– 59 484, 2024.

[47] A. Al-Hashedi *et al.*, "Cyber bullying detection based on emotion," *IEEE Access*, vol. 11, pp. 53 907–53 918, 2023.

[48] S. W. Azumah *et al.*, "Deep learning approaches for detecting adversarial cyber bullying and hate speech in social networks," in *2024 2nd International Conference on AIBThings*. IEEE, 2024.

[49] J. Sathya and F. M. H. Fernandez, "Effective automatic cyber bullying detection using a hybrid approach SVM and NLP," in *2024 International Conference on ADICS*. IEEE, 2024.

[50] D. Jadhav, S. Vijayalakshmi, and T. S. Palathara, "Rating-based cyber bullying detection with text, emojis on social media," in *2024 Interna- tional Conference on Emerging Electronics and Computer Technologies (ICEECT)*. IEEE, 2024.

[51] K. Agarwal *et al.*, "NLP-powered identification of online harassment," in *2024 International Conference on Communication, Computer and Power (IC2PCT)*. IEEE, 2024.

[52] A. Dewani, M. A. Memon, and S. Bhatti, "Cyber bullying detection: advanced preprocessing techniques & deep learning architecture for roman urdu data," *Journal of Big Data*, vol. 8, no. 1, p. 160, 2021.

[53] A. A. Jamjoom *et al.*, "RoBERTaNET: Enhanced RoBERTa transformer based model for cyber bullying detection with GloVe features," *IEEE Access*, vol. 12, pp. 58 950–58 959, 2024.

[54] M. M. Islam *et al.*, "Cyber bullying detection on social networks using machine learning approaches," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 2020.

[55] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, and H. D. E. Al-Ariki, "Dea-rnn: A hybrid deep learning approach for cyberbullying detection in twitter social media platform," *IEEE Access*, vol. 10, pp. 25 857–25 871, 2022.

[56] M. Atif *et al.*, "Cyberbullying detection and abuser profile identification on social media for roman urdu," *IEEE Access*, vol. 12, pp. 76 543– 76 560, 2024.

[57] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on deep-learning-based cyberbullying detec- tion," *Future Internet*, vol. 15, no. 5, p. 179, 2023.

[58] M. F. Almufareh *et al.*, "Integrating sentiment analysis with machine learning for cyber bullying detection on social media," *IEEE Access*, 2025, early Access.

[59] A. Philipo, J. Ding, D. Sarwatt, J. Mohamed, A. Yusufu, M. Danesh- mand, and H. Ning, "Sentiment-enhanced cyberbullying detection mod- els on social media platforms," *ACM Transactions on the Web*, 2025.

[60] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: Garbage in, garbage out," *PLoS One*, vol. 15, no. 12, p. e0243300, 2020.

[61] R. Bayari and A. Bensefia, "Text mining techniques for cyber bullying detection: state of the art," *ASTES Journal*, vol. 6, no. 1, pp. 783–790, 2021.

[62] M. S. Jahan and M. Oussalah, "A systematic review of hate speech au- tomatic detection using natural language processing," *Neurocomputing*, vol. 546, p. 126232, 2023.

[63] M. Umer, E. A. Alabdulqader, A. A. Alarfaj, L. Cascone, and M. Nappi, "Cyberbullying detection using pca extracted glove features and rober- tanet

IEER Institute for Excellence in Education & Research

transformer learning model," *IEEE Transactions on Computational Social Systems*, 2024.

[64] M. Raj, S. Singh, K. Solanki, and R. Selvanambi, "An application to detect cyberbullying using machine learning and deep learning techniques," *SN computer science*, vol. 3, no. 5, p. 401, 2022.

[65] S. Gupta, I. B. Jain, M. Saxena, P. K. Sarangi, A. K. Sahoo, and A. K. Agrawal, "Cyber bullying detection and classification using machine learning algorithms," in *2024 International Conference on Cybernation and Computation (CYBERCOM)*. IEEE, 2024, pp. 167–171.

[66] M. H. Shohan *et al.*, "Use of natural language processing for the detection of hate speech on social media," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 51, no. 2, pp. 86– 96, 2025.

[67] V. Jain, A. K. Saxena, A. Senthil, A. Jain, and A. Jain, "Cyber-bullying detection in social media platform using machine learning," in *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*. IEEE, 2021, pp. 401–405.

[68] G. B. Anwar and M. W. Anwar, "Textual cyberbullying detection using ensemble of machine learning models," in *2022 International Conference on IT and Industrial Technologies (ICIT)*. IEEE, 2022, pp. 1–7.

[69] J. Bhagya and P. Deepthi, "Cyberbullying detection on social media using svm," in *Inventive Systems and Control: Proceedings of ICISC 2021*. Springer, 2021, pp. 17–27.

[70] M. P. Rao, N. Kota, D. Nidumukkala, M. Madoori, and D. Ali, "Enhancing online safety: Cyberbullying detection with random forest classification," in *2024 10th International Conference on Communica- tion and Signal Processing (ICCSP)*. IEEE, 2024, pp. 389–393.

[71] W. Ashiq *et al.*, "Roman urdu hate speech detection using hybrid machine learning models and hyperparameter optimization," *Scientific Reports*, vol. 14, no. 1, p. 28590, 2024.

[72] L. Ketsbaia *et al.*, "A multi-stage machine learning and fuzzy approach to cyber-hate detection," *IEEE Access*, vol. 11, pp. 56 046–56 065, 2023.

[73] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cy- berbullying detection solutions based on deep learning architectures," *Multimedia Systems*, vol. 29, no. 3, pp. 1839–1852, 2023.

[74] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on deep-learning-based cyberbullying detec- tion," *Future Internet*, vol. 15, no. 5, p. 179, 2023.

[75] T. Mahmud, M. Ptaszynski, and F. Masui, "Deep learning hybrid models for multilingual cyberbullying detection: Insights from bangla and chittagonian languages," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2023, pp. 1–6.

[76] A. Dewani, M. A. Memon, S. Bhatti *et al.*, "Detection of cyber bullying patterns in low resource colloquial roman urdu microtext using natu- ral language processing, machine learning, and ensemble techniques," *Applied Sciences*, vol. 13, no. 4, p. 2062, 2023.

[77] D. Sultan, M. Mendes, A. Kassenkhan, and O. Akylbekov, "Hybrid cnn-lstm network for cyberbullying detection on social networks using textual contents," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, 2023.

[78] M. Gada, K. Damania, and S. Sankhe, "Cyberbullying detection using lstm-cnn architecture and its applications," in *2021 International Con- ference on Computer Communication and Informatics (ICCCI)*. IEEE, 2021, pp. 1–6.

[79] E.-Y. Daraghmi, S. Qadan, Y.-A. Daraghmi, R. Yousuf,
O. Cheikhrouhou, and M. Baz, "From text to insight: An integrated cnn-bilstm-gru model for arabic cyberbullying detection," *IEEE Access*, vol. 12, pp. 103 504–103 519, 2024.

[80] S. Ghosh, A. Chaki, and A. Kudeshia, "Cyberbully detection using 1d- cnn and lstm," in *Proceedings of International Conference on Communi- cation, Circuits, and Systems: IC3S 2020*. Springer, 2021, pp. 295–301.

[81] H. Lin, P. Siarry, H. Gururaj, J. Rodrigues, and D. K. Jain, "Special issue on deep learning methods for cyberbullying detection in multimodal social data," *Multimedia Systems*, vol. 28, no. 6, pp. 1873–1875, 2022.

[82] W. Tapaopong, A. Charoenphon, J. Raksasri, and T. Samanchuen, "Enhancing cyberbullying detection on social media using transformer models," in *2024 5th Technology Innovation Management and Engineer- ing Science International Conference (TIMES-iCON)*. IEEE, 2024, pp. 1–5.

[83] L. Yuan, T. Wang, G. Ferraro, H. Suominen, and M.-A. Rizoiu, "Transfer learning for hate speech detection in social media," *Journal of Compu- tational Social Science*, vol. 6, no. 2, pp. 1081–1101, 2023.

[84] M. Behzadi, I. G. Harris, and A. Derakhshan, "Rapid cyber-bullying detection method using compact bert models," in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*. IEEE, 2021,