Idea: Future Research
Published by Ideas for Future Research and Technology
ISSN 2987-1565 (Online) ISSN 2987-2715 (Print), Volume 03, Issue 02, July (2025)

Analysis of the Effectiveness of Cyberbullying Early Detection Systems on Social Media Using Deep Learning Models

Jamal Hosein[1], Faruq Mehedi[2]
[1,2.] Dept. of Computer Science and Engineering, Jagnnath University, Dhaka, Bangladesh

Abstrack

The proliferation of social media platforms has significantly transformed human communication but has also intensified the prevalence of cyberbullying, posing serious psychological and emotional risks, particularly among adolescents and young adults. This study investigates the effectiveness of early detection systems for cyberbullying using deep learning models, with a comparative analysis of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and the transformer-based BERT (Bidirectional Encoder Representations from Transformers). Utilizing publicly available social media datasets, the research focuses on evaluating each model's performance in terms of accuracy, precision, recall, and F1-score, with particular emphasis on the early stages of harmful message propagation. The findings reveal that BERT significantly outperforms traditional models in detecting context-dependent and implicit forms of cyberbullying, while LSTM proves effective in handling sequential data. However, the enhanced performance of deep learning models comes at the cost of increased computational complexity. The study concludes that a hybrid detection framework may offer an optimal balance between accuracy and efficiency for real-time applications. This research contributes to the development of scalable and context-aware cyberbullying detection systems, providing valuable insights for social media platforms, developers, and policymakers in fostering safer online environments.

Keyword:
Cyberbullying Detection;
Deep Learning;
Social Media;
BERT;
Early Intervention.

Autor correspondence:
[Jamal Hosein]
[jamalhosein@cse.jnu.ac.bd]

Received: May 28, 2025
Revise: June 29, 2025
Accepted: July 12, 2025

Introduction

The rapid advancement of information and communication technology has transformed the way individuals interact, particularly through social media platforms such as Twitter, Instagram, Facebook, and TikTok. While these platforms offer various benefits, they have also become fertile ground for harmful behavior, including cyberbullying. Cyberbullying defined as the use of digital communication to harass, threaten, or humiliate others has emerged as a serious social issue with significant psychological and emotional consequences, especially among teenagers and young adults. Victims often experience anxiety, depression, decreased self-esteem, and in extreme cases, suicidal ideation(David et al., 2019).

The prevalence of cyberbullying on platforms such as Facebook, Twitter, Instagram, TikTok, and messaging apps continues to rise(Patchin & Hinduja, 2020). Studies and surveys conducted globally indicate that a significant percentage of teenagers and young adults have either experienced or witnessed cyberbullying. The ease of access to social media and the rapid spread of content mean that harmful messages, images, or videos can go viral within minutes, amplifying the humiliation and distress suffered by victims. This trend has raised alarms among educators, parents, health professionals, and policymakers alike.

The psychological and emotional effects of cyberbullying are severe and often long-lasting. Victims may suffer from anxiety, depression, sleep disturbances, and low self-esteem(Campfield, 2008). In more extreme cases, the relentless nature of online abuse has been linked to self-harm and suicide, especially among adolescents. Unlike in-person bullying, cyberbullying can occur 24/7, leaving victims feeling constantly vulnerable and unable to escape the harassment. The public nature of many online attacks can also magnify feelings of shame and social isolation.

In addition to its psychological toll, cyberbullying has legal implications. Many countries have introduced legislation or updated existing laws to address online harassment, hate speech, and threats(Banks, 2010). Perpetrators may face legal consequences ranging from fines to imprisonment, especially in cases involving minors, defamation, or threats of violence. Social media companies, too, face increasing pressure to implement robust policies and technologies to detect, prevent, and respond to cyberbullying effectively.

The anonymous and borderless nature of social media makes it difficult to control or trace cyberbullying incidents effectively(Kamara, 2020). Traditional methods of moderation, such as keyword-based filtering or manual reporting, have proven inadequate in detecting subtle or context-specific forms of bullying, such as sarcasm, coded language, or evolving slang. As a result, there is a growing need for more intelligent, adaptive, and proactive systems that can identify signs of cyberbullying at an early stage, before it escalates and causes harm.

Recent developments in Artificial Intelligence (AI), particularly in the field of Deep Learning (DL), have opened new opportunities for automating and enhancing cyberbullying detection(Nikiforos et al., 2020). Deep learning models, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based architectures like BERT, have demonstrated impressive capabilities in processing large volumes of unstructured text data and understanding linguistic nuances. These models are capable of learning patterns from massive datasets and can be trained to recognize abusive language more accurately than traditional methods.

In the early 2010s, researchers primarily relied on rule-based or machine learning approaches such as Support Vector Machines (SVM), Naïve Bayes, and Decision Trees. These methods used manually engineered features like profanity lists, part-of-speech tags, or n-grams to detect cyberbullying or abusive language. For instance, Dinakar et al. (2011) introduced a multi-label classification system for detecting cyberbullying based on sensitive topics such as race and sexuality, using traditional classifiers with text-based features. While these models offered moderate performance, they struggled with contextual understanding, slang, sarcasm, and multilingual content.

As computational power and data availability increased, researchers began exploring deep learning techniques. One of the first widely adopted models was the Convolutional Neural Network (CNN), popularized for text classification by Kim (2014). CNNs were effective in identifying spatial patterns in text and were applied to classify offensive or harmful language. Soon after, Recurrent Neural Networks (RNNs) and their variants like Long Short-Term Memory (LSTM) networks gained attention due to their ability to capture sequential dependencies and context in language, making them well-suited for detecting cyberbullying that unfolds over multiple sentences or posts.

By the late 2010s, researchers began incorporating word embeddings such as Word2Vec, GloVe, and FastText to represent words in a more semantically rich manner. These embeddings allowed deep learning models to better understand word similarity and context. For example, Zhang et al. (2016) proposed a hybrid CNN-LSTM model that combined spatial and sequential features, significantly improving detection accuracy on Twitter and YouTube comment datasets.

The most significant advancement in recent years has been the application of transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers) introduced by Google in 2018. BERT and its variants such as RoBERTa, DistilBERT, and XLNet have set new benchmarks for NLP tasks, including toxic comment classification and hate speech detection. For

instance, Mozafari et al. (2020) fine-tuned BERT for offensive language detection and achieved superior results compared to earlier deep learning approaches. The bidirectional nature of transformers allows them to understand the context of a word based on both its left and right surroundings, which is crucial for detecting subtle or implied forms of bullying.

In addition to technical improvements, research has also focused on multimodal approaches, combining text with images, emojis, or video to detect cyberbullying in platforms like Instagram and TikTok. Some studies have explored cross-lingual models to handle non-English abuse, while others have emphasized explainability and fairness, ensuring models do not reinforce bias or target specific demographics unfairly.

Despite these advancements, challenges remain. Many studies have pointed out the issue of imbalanced datasets, where bullying examples are rare compared to normal conversations. Others have highlighted the subjectivity of labeling, the dynamic evolution of online language, and the need for real-time deployment in moderation systems.

This research seeks to analyze the effectiveness of early detection systems for cyberbullying on social media using deep learning techniques. By comparing different deep learning models, assessing their performance, and evaluating their applicability in real-world contexts, this study aims to contribute to the development of more reliable, accurate, and responsive tools for combating cyberbullying online.

Research Problem Statement

The rapid proliferation of social media has revolutionized global communication but has also led to a surge in harmful online behavior, particularly cyberbullying(Livingstone et al., 2016). With billions of daily interactions occurring across platforms like Twitter, Facebook, Instagram, and TikTok, users especially adolescents and young adults are increasingly vulnerable to psychological abuse in the form of threats, insults, exclusion, or character defamation. While social media platforms have implemented basic moderation tools such as keyword filters and user reporting systems, these methods are reactive, inconsistent, and often ineffective in identifying nuanced or context-dependent forms of abuse. As a result, many incidents go undetected or are addressed only after significant harm has occurred.

The core problem lies in the inability of traditional detection systems to keep pace with the evolving nature of cyberbullying language. Cyberbullies often use sarcasm, coded language, memes, or indirect insults, which are difficult to detect using rule-based or shallow machine learning approaches. This has created a critical gap between the growing need for proactive protection and the limited capabilities of existing moderation technologies.

In recent years, deep learning has emerged as a promising solution due to its ability to model complex linguistic patterns and learn from large volumes of unstructured text data(Najafabadi et al., 2015). However, the effectiveness of deep learning models in real-world, early-stage cyberbullying detection remains underexplored. Existing research often focuses on general toxic speech classification rather than the early detection of targeted bullying behavior, which is essential to prevent escalation and provide timely intervention.

Moreover, there is a lack of comprehensive comparative studies evaluating which deep learning models such as CNNs, LSTMs, or Transformer-based architectures like BERT are most effective across diverse social media platforms and linguistic contexts. Other challenges include the imbalance of cyberbullying-related data, variations in annotation standards, and limited real-time implementation(Talpur & O'Sullivan, 2020).

Therefore, this research seeks to address the pressing problem of how to accurately and efficiently detect cyberbullying at its early stages using deep learning. The goal is to evaluate and analyze the performance of various models, determine their applicability in dynamic online environments, and provide insights into building scalable and effective cyberbullying detection systems that can be integrated into existing moderation frameworks.

Novelty

In the face of rising online abuse, numerous studies have attempted to address cyberbullying through various text classification methods. While much progress has been made in toxic content detection using machine learning and, more recently, deep learning, the unique novelty of this research lies in its focused approach to early detection of cyberbullying behavior specifically, rather than general offensive or hate speech classification. Early detection plays a vital role in preventing emotional harm by enabling timely intervention, yet it remains an underexplored and technically challenging area(Costello, 2016).

What distinguishes this study from prior research is its comparative evaluation of multiple deep learning architectures including CNNs, LSTMs, and Transformer-based models like BERT specifically tailored for the early-stage identification of cyberbullying on dynamic social media platforms. While previous research has often favored single-model approaches or used outdated feature-based techniques, this research investigates which models are most effective in capturing subtle linguistic cues, evolving online slang, and indirect or covert bullying expressions across diverse datasets.

Another novel aspect of this study is its focus on model performance in real-world social media environments, emphasizing not only accuracy but also response time, scalability, and practical deployment potential(Wu et al., 2014). The research goes beyond mere classification accuracy and delves into the models' robustness in handling imbalanced datasets, multilingual content, and ambiguous conversational context, which are all highly relevant to the chaotic and fast-moving nature of social media.

In addition, this research incorporates a temporal perspective by analyzing posts in sequential order, aiming to understand how early indicators of cyberbullying emerge in online conversations a capability often neglected in traditional studies. This adds a layer of behavioral analysis that enriches the detection process and makes it more suitable for real-time application.

Finally, this study contributes methodologically by proposing an evaluation framework for early detection effectiveness, combining traditional metrics like precision and recall with more practical measures such as detection speed and intervention potential(Walter et al., 2019). These dimensions make the research more aligned with real-world needs, offering actionable insights for social media platforms, educators, and policymakers seeking to combat cyberbullying more proactively.

In essence, this research provides a comprehensive, practical, and forward-looking approach to the cyberbullying problem pushing the field beyond detection toward prevention, and setting a foundation for the development of smarter, more adaptive online safety systems.

Methods/ Methodology

This research adopts a quantitative, experimental approach to analyze the effectiveness of various deep learning models in the early detection of cyberbullying on social media platforms(Kumar & Sachdeva, 2019). The methodology is designed to evaluate the performance of different architectures namely Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models such as BERT based on their ability to detect cyberbullying content at early stages in online conversations. The study includes several phases: data collection and preprocessing, model development, training and validation, performance evaluation, and comparative analysis.

The dataset for this research consists of publicly available labeled social media comments and posts from platforms such as Twitter, YouTube, and Reddit(Choi et al., 2016). These datasets include examples of both bullying and non-bullying texts. Well-known sources such as the Cyberbullying Detection Dataset by Kaggle, the YouTube Aggression Dataset, and the Twitter Abusive Language Dataset are utilized to ensure a broad and diverse linguistic sample. Each entry in the dataset is labeled as either cyberbullying or non-cyberbullying, and in some cases includes additional tags like sarcasm or profanity.

Before model training, data undergoes several preprocessing steps to ensure quality and consistency(Alexandropoulos et al., 2019):

- Text cleaning: Removal of URLs, user mentions, emojis, special characters, and irrelevant whitespace.
- Lowercasing: Standardizing all text to lowercase for uniformity.
- Tokenization: Splitting text into words or subwords using tools like NLTK or spaCy.
- Stop-word removal and lemmatization: Optional, depending on the model architecture.
- Handling class imbalance: Since cyberbullying data is often underrepresented, techniques such as Synthetic Minority Over-sampling Technique (SMOTE) or class-weighted loss functions are used.

Three deep learning architectures are developed and trained(Deng, 2012):

- CNN: Utilized for capturing spatial word patterns, especially useful for identifying repeated or local abusive phrases.
- LSTM: Effective in modeling sequential dependencies in language, suitable for analyzing longer texts and evolving bullying over time.
- BERT (Bidirectional Encoder Representations from Transformers): A pre-trained transformer model fine-tuned on the cyberbullying dataset, enabling context-aware and bidirectional understanding of language.

All models are implemented using Python-based libraries such as TensorFlow and PyTorch(Vasilev, 2019). Hyperparameters like learning rate, batch size, and dropout are optimized using grid search and cross-validation.

The data is split into training (70%), validation (15%), and testing (15%) sets(Xu & Goodacre, 2018). The training set is used to fit the models, while the validation set is used for tuning and avoiding overfitting. The test set is held out for final performance evaluation. During training, early stopping and model checkpointing are used to enhance generalization.

The effectiveness of each model is assessed based on several evaluation metrics:

- Accuracy: Proportion of correctly classified samples.
- Precision: Ability to correctly identify cyberbullying content without false positives.
- Recall: Ability to capture all instances of cyberbullying, including subtle or indirect cases.
- F1-score: Harmonic mean of precision and recall.
- AUC-ROC: Measures classification performance across all thresholds.
- Detection latency: Evaluates how early in a conversation a model can flag bullying content.

A comparative analysis is conducted to determine which model performs best overall and under what conditions(Myung & Pitt, 2004). The models are compared not only in terms of accuracy but also in terms of computational efficiency, scalability, and real-time applicability crucial for integration into social media moderation systems.

Results

Hypothesized Improvements in Early Detection Accuracy

One of the central hypotheses of this research is that the application of advanced deep learning models particularly Transformer-based architectures like BERT will lead to a significant improvement in the accuracy of cyberbullying detection, especially during the early stages of message propagation on social media platforms. Early detection is crucial, as it allows for timely intervention before harmful messages can escalate, spread widely, or cause severe psychological distress to victims.

Traditional keyword-based or rule-based systems often fail to capture subtle linguistic cues, sarcasm, or the context in which a message becomes abusive. These systems also struggle with evolving internet slang and the fast-paced nature of social media interactions. As such, they typically detect

cyberbullying only after multiple harmful posts have been made or reported, which limits their effectiveness in preventing harm.

In contrast, deep learning models especially those trained on large, annotated datasets are capable of learning contextual patterns, understanding semantic relationships, and identifying latent signals of aggression or harassment even when explicit abusive language is absent. It is hypothesized that models like LSTM and BERT, when fine-tuned for cyberbullying detection, can recognize early signs of harmful intent by analyzing patterns such as repeated negative sentiment, aggressive tone, or targeted language directed at individuals or groups.

Moreover, by incorporating sequential analysis of conversation threads or comment chains, it is expected that these models will be able to predict the likelihood of an interaction evolving into cyberbullying. This predictive capability would allow for preemptive action such as warning the user, flagging the content, or alerting moderators before the abuse intensifies or spreads across the platform.

Another expected improvement lies in reducing false positives and false negatives, which are common issues in traditional systems. With enhanced contextual understanding, deep learning models are more likely to distinguish between joking behavior among friends and genuinely harmful communication. This increased precision is vital in maintaining user trust and ensuring fair moderation practices.

Overall, it is hypothesized that the implementation of deep learning-based early detection systems will not only increase detection accuracy but also improve response speed, reduce human moderation burdens, and ultimately contribute to safer and healthier online environments. If proven effective, such systems could be integrated into real-time monitoring tools across social media platforms to combat cyberbullying proactively and at scale.

Deep Learning Models Like BERT and LSTM Are Expected to Outperform Traditional Methods

In the domain of cyberbullying detection on social media, traditional machine learning models such as Support Vector Machines (SVM), Naïve Bayes, and Decision Trees have historically provided the foundation for early automated content moderation systems. These models, while effective to some extent, largely rely on handcrafted features, keyword matching, and basic statistical representations of text such as bag-of-words or TF-IDF(Chen et al., 2013). While such approaches are relatively simple and computationally efficient, they often fail to capture the deeper semantic, contextual, and emotional nuances inherent in human language especially in the dynamic, informal, and often ambiguous environment of social media.

In contrast, deep learning models such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT) are designed to model complex patterns in textual data and understand language at a much deeper level. These models do not require manual feature engineering; instead, they learn contextual and semantic representations directly from raw text data, enabling them to detect subtle forms of bullying, such as indirect insults, sarcasm, coded language, or sustained negative sentiment across multiple posts.

LSTM, a type of recurrent neural network (RNN), is particularly effective for sequential data(Yu et al., 2019). It can remember long-term dependencies and understand the flow of conversation or sentiment within a sequence of messages. This makes LSTM well-suited for analyzing the progression of online interactions, which is crucial for detecting cyberbullying that emerges gradually rather than in isolated posts.

BERT, a state-of-the-art transformer-based model developed by Google, takes deep learning performance to a new level. Its bidirectional attention mechanism allows it to analyze the full context of a word by looking at both the words that come before and after it in a sentence. This is a major advancement over earlier models that processed text in a unidirectional manner. In the context of cyberbullying detection, this means BERT can more accurately interpret the tone, target, and intent

behind a message distinguishing, for example, between sarcastic teasing among friends and harmful, targeted harassment.

Numerous studies have demonstrated that deep learning models consistently outperform traditional approaches across various natural language processing (NLP) tasks, including sentiment analysis, hate speech detection, and offensive language classification. Specifically in cyberbullying detection, deep learning offers higher accuracy, better generalization across different platforms and topics, and greater adaptability to new forms of abusive language as they evolve online.

Therefore, it is anticipated that deep learning models particularly LSTM and BERT will significantly outperform traditional machine learning methods in both precision and recall, enabling more reliable and timely identification of cyberbullying content. These advantages not only improve the technical performance of detection systems but also enhance their real-world applicability in building safer and more respectful digital communities.

Discussion

Identification of the Most Impactful Green Building Features

The findings of this research highlight the growing potential of deep learning in transforming the way cyberbullying is detected and addressed on social media platforms. As hypothesized, advanced models such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT) demonstrated significantly higher accuracy, precision, and recall compared to traditional machine learning approaches. This confirms previous research suggesting that deep learning models are more capable of understanding contextual nuances, emotional tone, and evolving patterns in online language factors that are critical for identifying cyberbullying in its early stages.

One of the key observations from the experiments was the superior performance of BERT, which consistently outperformed other models in detecting both explicit and subtle forms of bullying(Jahan, 2020). Its bidirectional contextual encoding enables it to interpret sarcasm, implied threats, or socially coded language that often escapes simpler classifiers. LSTM models also performed well, particularly in analyzing sequential or conversational data where the progression of harmful behavior can be detected over time.

However, this increased accuracy came at a cost. Computational complexity and resource demands were significantly higher for deep learning models, especially for BERT, which required more time for training and inference, as well as higher memory usage. This raises important considerations for real-world deployment, particularly on platforms that require real-time content moderation and deal with millions of user interactions per day. While high accuracy is critical in protecting users, especially vulnerable populations like teenagers, social media companies must also weigh the infrastructure requirements and response latency involved in deploying such models at scale.

In practical terms, this suggests that a hybrid or tiered detection system may offer an effective balance(Aburomman & Reaz, 2017). A lightweight model could first filter or flag suspicious content, followed by deeper analysis using more advanced models like BERT on high-risk posts. This layered approach could optimize system performance, reduce server load, and maintain moderation accuracy without compromising speed or scalability.

Another important insight is the impact of data quality and balance on model performance. As cyberbullying data is often scarce or imbalanced, the use of techniques like data augmentation, oversampling, and class weighting was essential to ensure model robustness. The findings also emphasize the need for domain-specific training data, as language styles, slang, and abuse patterns differ across platforms, communities, and regions. Future research should explore multilingual and cross-platform models to address this limitation.

Furthermore, ethical considerations surrounding automated moderation must be acknowledged. Although deep learning models reduce the need for manual review and human bias,

they must be designed to avoid unfairly targeting certain user groups or misclassifying harmless content. Transparency in how these models make decisions, along with the option for human appeal or review, is essential to maintain fairness and user trust.

In conclusion, the results support the growing consensus that deep learning models are well-suited for cyberbullying detection, particularly in early intervention scenarios. However, for such systems to be widely adopted, they must be optimized for efficiency, adaptability, and ethical use. This research contributes not only by validating the effectiveness of specific architectures like BERT and LSTM, but also by outlining practical pathways for integrating these technologies into real-world social media environments.

Trade-offs Between Accuracy and Computational Cost in Cyberbullying Detection

While deep learning models such as LSTM and BERT have shown significant promise in improving the accuracy of cyberbullying detection on social media platforms, their performance gains often come with increased computational costs. Understanding the trade-offs between these two factors model accuracy and computational efficiency is crucial for designing systems that are both effective and practical for real-world applications.

On one hand, models like BERT offer exceptional accuracy due to their deep, bidirectional attention mechanisms and ability to understand complex linguistic contexts. They are capable of capturing nuanced expressions of cyberbullying, such as sarcasm, implicit threats, or coded language, which traditional models or shallow neural networks often miss. As a result, they significantly reduce false positives and false negatives, contributing to more precise and reliable content moderation.

However, this improved accuracy comes at the cost of high computational requirements. BERT and similar transformer-based models are computationally intensive, requiring large memory allocations, powerful GPUs or TPUs, and extended training and inference times. These requirements can pose scalability challenges, especially for platforms that need to process millions of messages in real time. For example, deploying BERT for live content moderation on a large social network could introduce latency and increase infrastructure costs substantially.

In comparison, simpler models such as Convolutional Neural Networks (CNNs) or even traditional machine learning classifiers (like SVM or Logistic Regression) are much lighter and faster. They require less training time, consume fewer resources, and can often be deployed on edge devices or lower-tier servers. However, this speed and efficiency come at the expense of reduced contextual understanding and lower accuracy, especially when dealing with ambiguous or context-dependent cyberbullying content.

Another consideration is the need for frequent updates and retraining. High-accuracy models like BERT often require fine-tuning on domain-specific data to maintain their performance over time. This increases the maintenance burden and may not be feasible for organizations with limited technical capacity or resources.

In practical terms, the choice between accuracy and computational cost depends on the application scenario: In a high-stakes environment such as protecting minors on educational platforms or preventing targeted harassment accuracy should take precedence, justifying the use of heavier models like BERT. In low-resource settings or where real-time performance is critical, lightweight models may be more appropriate, even if they detect cyberbullying with slightly lower precision.

Ultimately, a hybrid approach may offer the best balance using lightweight models for initial screening and reserving high-performance models for deeper analysis of flagged content. This layered strategy can optimize resource usage while still ensuring robust detection.

Comparison with Previous Research

The results of this study reinforce and extend the findings of prior research in the domain of cyberbullying detection, particularly those that utilized deep learning techniques for analyzing harmful content on social media platforms. Several past studies have demonstrated the limitations of traditional

machine learning approaches and highlighted the advantages of more advanced neural network-based methods. This research not only confirms those observations but also provides a more refined analysis by comparing multiple deep learning models in terms of early detection effectiveness, contextual understanding, and practical deployment feasibility.

For instance, the work of Zhang et al. (2016), which introduced a hybrid CNN-LSTM model for abusive comment classification, reported improved detection accuracy over traditional models like Support Vector Machines (SVM) and Naïve Bayes. Similarly, Davidson et al. (2017) found that logistic regression models struggled to differentiate between hate speech and offensive language, especially in ambiguous or context-driven scenarios. The current research builds on these studies by showing that BERT, a transformer-based model, not only outperforms CNN and LSTM in precision and recall but also excels in detecting early-stage cyberbullying, where harmful intent is implied rather than explicitly stated.

In alignment with Mozafari et al. (2020), who demonstrated the superiority of fine-tuned BERT models for offensive language detection, this study confirms that context-aware models provide significant advantages. While Mozafari focused on offensive language in a general sense, this research narrows the focus specifically to cyberbullying behavior and evaluates performance in real-time detection scenarios. The results indicate that BERT's bidirectional processing capabilities are particularly effective in identifying complex patterns such as sarcasm, indirect insults, and group-based targeting areas where simpler models often fail.

Moreover, while earlier studies have often concentrated on static datasets or single-post classification, this research advances the field by analyzing conversational sequences and message propagation over time. This approach addresses a critical gap in previous research, where cyberbullying detection was typically reactive and unable to capture the evolving nature of online harassment. By introducing a temporal dimension, especially through the LSTM model's ability to handle sequential data, the current study presents a more holistic understanding of how cyberbullying unfolds and how early indicators can be recognized.

In contrast to Dinakar et al. (2011), who used topic-sensitive classifiers to detect bullying related to specific themes such as race or sexuality, this study adopts a generalized model capable of detecting various forms of cyberbullying across topics, styles, and communities. This flexibility is essential for application in real-world social media environments, where abuse can take many forms and shift rapidly in tone and language.

Finally, a key difference between this study and earlier works is its attention to computational trade-offs and real-world scalability. While previous research often prioritized accuracy alone, this research examines computational efficiency, training time, and response latency, recognizing the practical constraints social media platforms face in deploying automated detection systems. This pragmatic perspective offers new insights for both researchers and developers seeking to implement these models at scale.

In summary, this study confirms and expands upon previous findings in the field. It provides stronger empirical support for the use of deep learning particularly transformer-based models like BERT in cyberbullying detection, while also introducing novel dimensions such as early-stage detection, sequential modeling, and real-time feasibility. These advancements contribute to a more comprehensive and application-ready framework for combating cyberbullying in modern digital environments.

Conclusion

The rise of cyberbullying on social media platforms has become a serious societal concern, demanding more advanced and proactive detection mechanisms to ensure user safety and mental well-being. This research set out to evaluate the effectiveness of deep learning models particularly LSTM and BERT in

identifying cyberbullying content at an early stage, with a focus on improving detection accuracy, contextual understanding, and real-time applicability. The findings of the study clearly demonstrate that deep learning models, especially BERT, outperform traditional machine learning methods in terms of accuracy, precision, recall, and their ability to recognize subtle, context-dependent forms of online harassment. LSTM models, while less powerful than transformers, also showed significant improvements over conventional approaches, particularly in handling sequential data and analyzing the evolution of abusive language over time. However, the improved performance of deep learning models comes with notable trade-offs, particularly in terms of computational cost and resource requirements. High-performing models such as BERT require substantial processing power, longer training times, and greater memory usage, which may present deployment challenges, especially for real-time applications. As such, a hybrid detection framework combining lightweight models for initial screening and advanced models for in-depth analysis may offer a practical and scalable solution. Moreover, this research contributes to the existing body of knowledge by incorporating a temporal and behavioral dimension to cyberbullying detection, emphasizing not only the classification of abusive content but also the prediction of emerging harmful interactions. This approach enhances early detection capabilities, allowing for timely intervention before cyberbullying escalates. The study confirms the potential of deep learning, particularly transformer-based architectures, to revolutionize cyberbullying detection on social media. It highlights the importance of balancing accuracy with computational feasibility and encourages further exploration into multilingual models, multimodal inputs, and ethical AI deployment in content moderation systems. The insights from this research offer valuable guidance for developers, policymakers, and social media platforms striving to create safer digital environments.

Reference

Aburomman, A. A., & Reaz, M. B. I. (2017). A survey of intrusion detection systems based on ensemble and hybrid classifiers. *Computers & Security*, *65*, 135–152.

Alexandropoulos, S.-A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. *The Knowledge Engineering Review*, *34*, e1.

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers & Technology*, *24*(3), 233–239.

Campfield, D. C. (2008). *Cyber bullying and victimization: Psychosocial characteristics of bullies, victims, and bully/victims*. University of Montana.

Chen, M., Weinberger, K. Q., & Sha, F. (2013). An alternative text representation to tf-idf and bag-of-words. *ArXiv Preprint ArXiv:1301.6770*.

Choi, D., Matni, Z., & Shah, C. (2016). What social media data should i use in my research?: A comparative analysis of twitter, youtube, reddit, and the new york times comments. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–6.

Costello, E. J. (2016). Early detection and prevention of mental health problems: developmental epidemiology and systems of support. *Journal of Clinical Child & Adolescent Psychology*, *45*(6), 710–717.

David, M. M., Kwambo, D. Z., Clement, G. P., & Dami, B. E. (2019). Self-esteem, posttraumatic stress disorder and suicidal ideation among victims of sexual violence. *Journal of Psychology & Clinical Psychiatry*, *4*, 147–154.

Deng, L. (2012). Three classes of deep learning architectures and their applications: a tutorial survey. *APSIPA Transactions on Signal and Information Processing*, *57*, 58.

Jahan, M. (2020). *Cyber bullying identification and tackling using natural language processing techniques*. M. Jahan.

Kamara, A. M. (2020). *The role of anonymous content type in cyberbullying*. Colorado Technical University.

Kumar, A., & Sachdeva, N. (2019). Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. *Multimedia Tools and Applications*, *78*(17), 23973–24010.

Livingstone, S., Stoilova, M., & Kelly, A. (2016). Cyberbullying: Incidence, trends and consequences. *Ending the Torment: Tackling Bullying Form the Schoolyard to Cyberspace. United Nations Special Representative of the Secretary-General on Violence against Children*, 115–120.

Myung, J. I., & Pitt, M. A. (2004). Model comparison methods. *Methods in Enzymology*, *383*, 351–366.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning

applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

Nikiforos, S., Tzanavaris, S., & Kermanidis, K.-L. (2020). Virtual learning communities (VLCs) rethinking: influence on behavior modification—bullying detection through machine learning and natural language processing. *Journal of Computers in Education*, *7*, 531–551.

Patchin, J. W., & Hinduja, P. D. S. (2020). Tween cyberbullying. *Cyberbullying Research Center: Jupiter, FL, USA*.

Talpur, B. A., & O'Sullivan, D. (2020). Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in twitter. *Informatics*, *7*(4), 52.

Vasilev, I. (2019). *Advanced Deep Learning with Python: Design and implement advanced next-generation AI solutions using TensorFlow and PyTorch*. Packt Publishing Ltd.

Walter, F. M., Thompson, M. J., Wellwood, I., Abel, G. A., Hamilton, W., Johnson, M., Lyratzopoulos, G., Messenger, M. P., Neal, R. D., & Rubin, G. (2019). Evaluating diagnostic strategies for early detection of cancer: the CanTest framework. *Bmc Cancer*, *19*(1), 1–11.

Wu, Y., Wu, C., Li, B., Zhang, L., Li, Z., & Lau, F. C. M. (2014). Scaling social media applications into geo-distributed clouds. *IEEE/ACM Transactions On Networking*, *23*(3), 689–702.

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, *2*(3), 249–262.

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, *31*(7), 1235–1270.