Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

EXPLORING TRENDS AND PATTERNS OF CYBERBULLYING THROUGH BIG DATA ANALYSIS

Muhammad Zahid Khan[1], Dr. Muhammad Ashraf[2,3], Engr. Muhammad Akram khan[3], Dr. Akbar Khan[4], Fabiha[5], Muhammad Anas[6]

MS Scholar Department of Computer Engineering BUITEMS Quetta, Pakistan. Email: Zahidkhan827@yahoo.com
Associate Professor Department of Computer Engineering BUITEMS Quetta, Pakistan. Email: muhammad.ashraf@buitms.edu.pk
Lecturer Department of Computer Engineering BUITEMS Quetta, Pakistan. Email: mohammad.akram@buitms.edu.pk
Assistant Professor Department of Computer Engineering BUITEMS Quetta, Pakistan. Email: akbar.khan@buitms.edu.pk
MS Scholar Department of Computer Science BUITEMS Quetta, Pakistan. Email: fabihakhan38@gmail.com
Department of Computer Engineering BUITEMS Quetta, Pakistan. Email: manaswazir@gmail.com

DOI: https://doi.org/

Keywords
Cyberbullying Detection, Machine Learning, Natural Language Processing, BERT Transformer, Social Media Analysis, Big Data Analytics, Digital Safety, Automated Content Moderation.

Article History
Received on 08 Nove 2025
Accepted on 15 Dec 2025
Published on 25 Dec2025

Corresponding Author: *

Muhammad Zahid Khan

Abstract

Cyberbullying has emerged as a pervasive digital threat affecting 20-40% of youths worldwide, transcending traditional bullying boundaries through its persistent, anonymous, and cross-platform nature. This research addresses critical gaps in cyberbullying detection and pattern recognition through comprehensive big data analysis of 60,233 social media communications across multiple platforms. The study employs three distinct machine learning approaches to develop sophisticated automated detection capabilities: a BERT-based transformer model, a Bidirectional LSTM with attention mechanism, and an Advanced Random Forest classifier with comprehensive feature engineering. The research methodology integrates natural language processing, statistical analysis, and qualitative examination to identify cyberbullying trends, demographic patterns, and linguistic predictors. The BERT model achieved superior performance with 89.86% accuracy and 96.23% AUC, demonstrating exceptional capability in identifying cyberbullying content while maintaining high recall (96.00%) to minimize false negatives. The BiLSTM model provided competitive results (87.57% accuracy, 94.97% AUC) with valuable interpretability features, while the Random Forest approach (82.00% accuracy, 93.19% AUC) revealed critical feature importance insights, identifying average word length (7.68%), sentence count (6.80%), and question count (5.83%) as primary predictors. Feature importance analysis revealed that cyberbullying detection requires sophisticated understanding of structural linguistic patterns beyond explicit content markers, with interrogative patterns and text complexity serving as significant discriminators. The study acknowledges the increased challenges created by the digital transition

to COVID-19 and advises strategies at multiple levels based on its evidence-based conclusions. This research enhances theoretical understanding about patterns of digital aggression, and provides practical solutions for a real-time detection system, educational interventions, and policy recommendations. This multidisciplinary approach, which blends perspectives from computer science, psychology, and sociology, improves the capacity to study cyberbullying and offers scalable technological solutions for making online spaces safer.

INTRODUCTION

The Digital Age has dramatically transformed the ways in which people connect, work, and create community through new methods and means of communicating and sharing information globally, but the Digital Age has also provided additional opportunities for people to use technology in negative ways. With more than 97% of American adolescents having the ability to access the Internet [1], they can connect with, communicate with, and learn from others from all over the world. However, this increased ability to communicate and learn about one another has created the issue of cyberbullying, which presents a unique and significant societal problem requiring the efforts of professionals to find creative and thoughtful solutions.

Cyberbullying does not occur during specific timeframes and does not occur at specific geographic locations, as traditional school bullying does; as a result, cyberbullying is available to perpetrators 24 hours a day, 7 days a week, and through many different digital mediums. Consequently, cyberbullying creates a continuum of harassment combined with anonymity, leading to an almost increasing frequency of harassment that occurs at the same time as face-to-face interactions. The psychological consequences of this level of harassment likely differ significantly from the psychological consequences of traditional school bullying, making the problems faced by victims of cyberbullying unique from those of victims of traditional school bullying, especially when considering that the perpetrator may be harassing the victim in a digital environment that is either unfamiliar or confusing to the victim [2]. In order to fully understand the phenomenon of cyberbullying, it is necessary to conduct extensive research to understand the various trends, data, and contexts that underlie the phenomenon in today's society.

Research indicates that between 20-40% of adolescents will experience cyberbullying at some point during their lives, suffering from serious adverse consequences to their mental health, social relations, and education [1]. Furthermore, cyberbullying does not restrict its consequences to individual victims, but impacts the broader community, including families and educational systems (which typically do not have the knowledge or resources to appropriately manage incidents occurring via the Internet).

The rise of cyberbullying has profoundly changed since technological advancements have changed how cyberbullies can bully their victims [3]. The transition from Electronic Mail and SMS messaging to Social Media and other newer forms of Digital Technology has led to new challenges for Researchers, Teachers, and Government Officials to continue to learn and adapt to the ever-changing patterns of digital aggression / bullying and develop new methods for preventing and intervening against it.

The demographics of Cyberbullying have also changed significantly between 2006 and 2012. The rate of adolescent victimization increased from 15 percent to 21 percent; whereas perpetration increased

from 12 percent to 15 percent of Adolescents.[4]. The rate of victimization among female Adolescents increased 10 percent (17% to 27%) whereas Male Adolescents had only increased by 3 percent (12% to 15%).[4]. This shows that GENDER INEQUALITY plays a role in CYBERBULLYING. Another aspect to consider is that Sexual Minorities show much higher instances of both Traditional-Untraditional and CYBERBULLYING overall than heterosexual populations. Thus, cyberbullying may serve as a 'tool' that allows for broader types of identity-based discrimination.[4]. All of these aspects highlight the need for Layered (or multiple layers of) Analysis.

The structure of the Victimisation or Perpetration frameworks for traditional bullying established by Olweus et al., developed in the1990's has to be updated and revised for Digital / Cyberbullying [5]. There are new dynamics in Online environments that affect all characteristics of Harassment, such as User participation as an audience, Content Permanency, Viral Exposure, and User Anonymity.

Big data analytics and neural network machine learning has transformed the field of research on the topic of Cyberbullying. Although traditional qualitative methods (e.g., interviews, surveys, qualitative observations) are important elements of early cyberbullying studies, there are too many limitations when compared to the rapid increase in use of the Internet and mobile digital technologies for Cyberbullying research (number of subjects, geography, methodology). Technology has given researchers new ways (e.g., through Access to large amounts of data) to analyse and test hypotheses about Cyberbullying trends and predictors at a previously impossible scale.

The progressive developments in the field of Machine Learning and Artificial Intelligence (AI) have made it possible to detect Cyberbullying and Harassment; this is evidenced by the continued development of BERT (Bidirectional Encoder Representations from Transformers), Bidirectional LSTM (Long Short-Term Memory) Networks with Attention Mechanisms, and Ensemble Methods achieving greater than 90% accuracy identifying Cyberbullying [6][7][8]. The advancements in Machine Learning have resulted in the ability to detect Cyberbullying, and identify Cyberbullying patterns, immediately on a scale that was previously impossible through manual monitoring, which has led to a large increase in information collected and analyzed.

Despite these advances in technology, there are still many gaps in our knowledge about social, psychological, and contextual factors contributing to an increase in Cyberbullying. While Machine Learning has been very good at pattern recognition and classifying data based on patterns, it does not provide an understanding of the underlying causes and motivations of individuals, nor does it provide us with information about the social dynamics surrounding Cyberbullying. This indicates the importance of continuing to develop and foster interdisciplinary research between technology-based research and traditional social sciences. Additionally, the COVID-19 pandemic and the large increase in digital engagement during the pandemic have created an immediate need for research regarding the global crisis' impact on Digital Aggression. Educational institutions struggle with cyberbullying due to unclear jurisdictional boundaries, limited resources, and insufficient technical expertise [9][10]. Schools increasingly face expectations to address digital harassment without adequate technological infrastructure, policies, or trained personnel, leaving victims with minimal institutional support.

There are significant limitations to current prevention strategies. For instance, only one-third of all cyberbullying victims report an incident to an adult since many victims prefer to tell their parent 29 percent of the time compared to 17 percent of the time when telling a member of the school's staff [4]. The reason victims underreport incidents is that they are concerned about adults' digital aptitude, they

have limited technology access to report their incident, and they doubt that effective action against the perpetrator will be taken.

Gaps in the research related to cyberbullying exist because there are not as many methods to integrate multiple data and analytical instruments. While investigations of individuals produce good findings for their respective methods, synthesizing the findings through a methodological perspective will more accurately demonstrate that cyberbullying is both a technological and social issue.

It is critical to address cyberbullying as part of United Nations Sustainable Development Goals 4 (Quality Education) and 16 (Peace, Justice and Strong Institutions) to create safe learning environments for students. In this research study, we will analyze and classify over 60,000 messages sent through social media channels using BERT-based transformers, BiLSTM neural networks with attention, and Random Forest Classifiers to determine and analyze patterns, trends, and risk factors related to cyberbullying. This study will provide valuable insights based on evidence for schools, technology developers, policymakers, and mental health practitioners. This information will support the development of scalable interventions that provide adolescents with safer online environments.

2. LITERATURE REVIEW

With the rise of digital communications and social networking (or social media), the way we interact with others and the ways we connect have changed; however, along with those positive aspects of new technology, we also face a lot more problems (negative aspects) when it comes to dealing with things like cyberbullying. This article serves as a literature review of research conducted related to detecting, analyzing and intervening with cyberbullying, particularly using machine learning or big data methods. The goal of this literature review is to provide an overview of research conducted across computer science, psychology, sociology, and data science so that we can build a framework (theoretical and methodological) for understanding and addressing cyberbullying as a unique socio-technical problem that will require interdisciplinary methods to investigate and solve.

Using technologies classified as SMAC (Social Media, Mobile, Analytics and Cloud), researchers are producing enormous quantities of real-time, multi-modal data, both from an opportunity and challenge perspective. As shown by Kumar et al. [11], the ability to mine the large amounts of multimedia social big data from user-generated content on different social media platforms will provide more opportunities for researchers to better analyze relationships and behaviors beyond what can be captured through typical survey instruments. Because social media data come from so many different sources and include different types of content (e.g., text, image, video, metadata), a more sophisticated and flexible analytical approach is required to effectively combine multiple forms of diverse social media data into a more thorough understanding of social processes.

As an area of research, automated cyberbullying detection is becoming increasingly important in relation to the rise of online harassment Elsafoury et al. [12]. The authors provide a detailed overview of automated cyberbullying detection systems as well as challenges faced when developing these systems, such as definitional and technological problems. The authors identify contextual data modeling techniques such as BERT as being far superior to traditional machine learning/deep learning techniques, as well as having much higher consistency in their results. Additionally, it has been shown that using internet slang when training the neural networks associated with the detection of cyberbullying improves the ability of the network to successfully identify cyberbullying incidents,

indicating that the ability to detect cyberbullying occurrences relies on an understanding of both general language trends and trends specific to the domain of cyberbullying. Therefore, the authors of this paper believe that as the language associated with cyberspace continues to change and grow, the ability to create effective automated cyberbullying detection systems must also change and grow accordingly.

A look at Cyberbullying with Crime Research provides new information on Cyberbullying in the context of Crime Research on the Internet. Jabeen et al. [13] have shown how social big data can be used to mine crime data, where social media platforms can be used as venues for crime and as sources of data to understand and prevent crime. Cyberbullying has unique characteristics as a form of online violence in the complexity of how we think about crime in our Digital Society. Silber-Varod et al. [14] have developed Mixed Methods (both Quantitative and Qualitative) Models through Culturomics, which utilise Data Mining and Text Mining techniques. The Development of Mixed Methods Models, Combined with the Nature of Computerised Social Science with Interpretive Methods, will provide Researchers with an excellent option for identifying Patterns and Trends in Large Scale Textual Data Sets.

Multifactorial examinations of cyberbullying research demonstrate a number of emerging trends and areas that require additional research. Saleem et al. [15] performed a tertiary source review of cyberbullying research encompassing eight areas of research including measuring the incidence of cyberbullying, understanding the impacts of cyberbullying, creating interventions to address cyberbullying and using automation tools built with artificial intelligence. In their findings, they concluded there has been a shift in the field of cyberbullying research from an emphasis on descriptive and correlational approaches to predictive and intervention-based approaches. However, they also suggested that foundational theoretical work in this area continues to be necessary even with the availability of new technology for researchers. A second bibliometric study conducted by Barragán Martín et al. [16] examined 1,530 publications related to cyberbullying. The authors found that the publication rate for international documents has increased (the majority being in English language), with the journal "Computers in Human Behavior" the leading journal for publishing cyberbullying research. This work has highlighted a growing trend toward collaboration within the research community and a general lack of research being produced by many geographical areas worldwide.

The study of time through bibliometric analyses provides insight into patterns of research methodologies or procedures. Shao et al. [17] studied the development of cyberbullying perpetrated against adolescents from 1986 to 2019 using a CiteSpace knowledge mapping methodology, allowing them to determine which aspects of cyberbullying perpetration are presently and will likely be research priority topics in the future. Cross-national comparative studies provide a better understanding of the cultural and contextual considerations that need to be taken into account regarding cyberbullying perpetration. Cosma et al. [18] performed a study comparing the trends of bullying victimization across 37 different countries from 2002-2014 and observed a decline in traditional forms of bullying being perpetrated. They found approximately 45.8 per cent of all incidence of the victims of cyberbullying were also victims of traditional forms of bullying and vice versa, indicating an ongoing relationship between violent acts committed against both groups. It is imperative that this information be utilized to create appropriate interventions that are sensitive to the vastly different sociocultural circumstances and patterns of technological adoption on a global basis.

Due to the development of various new patterns that cannot be analysed through traditional methods, therefore a set of Association Analysis techniques was developed. Using the example of the use of the Apriori Algorithm to identify patterns of cyberbullying on Twitter in Malay by Zainol et al. [19], they discovered more than 80 different high-confidence patterns that could be used to develop procedures for detecting cyberbullying on Twitter. The results of their work indicate that association rule mining appears to offer promise as a reliable indicator for creating automated detection systems that predict patterns of behaviour associated with cyberbullying, and therefore potential methods of developing intervention strategies. In addition, bibliometric analyses of 387 articles by Thi and Ngoc Ho [20] for the period of 2010-2020 confirm that the majority of published research on the topic of cyberbullying, particularly the issue of cyberbullying victimisation, was published by authors from the USA or European countries (91%), which highlights significant limitations in terms of geographical coverage and the ability to generalise to cross-cultural contexts.

The utilization of machine learning for classification in automated environments and providing insight into language & actions associated with cyberbullying has potential. In their analysis of 847,548 tweets using supervised machine-learning techniques, Dhungana Sainju et al. [21] found that the majority of bullying-related tweets provided information from the perspective of victims, and that they also reflected both general and online forms of bullying. They further noted that there were statistically significant differences between the length of bullying-related tweets and that there was very little difference in the level of bullying-related posts after a number of "high-profile" events. In addition, the authors believe that Twitter will play a dual purpose in both acting as a platform for cyberbullying and providing a source of assistance for victims of bullying through the use of the platform. By using interdisciplinary approaches, innovative computational models are developed. As an example, Potha et al. [22] use data-mining and multiple sequence alignment to analyze the temporal order of events. They are then able to convert the time series into symbolic strings, where the symbolic strings are used to compare and contrast preserved and transformed patterns of questioning.

The COVID-19 pandemic provided a unique context to study changes in cyberbullying behaviour due to social disruption. For example, during the pandemic, Karmakar and Das [26] were able to conduct a detailed statistical analysis to illustrate how conversations about cyberbullying on Twitter increased above normal levels due to the pandemic. Das et al. [27] also identified several points in time when there were significant changes in how people talked about cyberbullying, using 454,046 tweets they analysed from late March 2020 to the impact of COVID-19. Both of these studies demonstrate that cyberbullying behaviour changes when there are environmental changes, and also illustrate how methodologies can be developed to identify real-time changes in behaviour. Liu and Sui [28] conducted Geographic Information System (GIS) analyses to demonstrate how GIS technology can be used to better understand the spatiotemporal patterns of cyberbullying and use visual data to understand patterns that influence where cyberbullying takes place. Song et al. [29] analysed 350,314 documents published online and were able to use General Strain Theory to identify predictive signs for future cyberbullying behaviour and found a significant overlap between social large data and traditional survey results. They also identified weak signal topic areas related to the media and cultural factors that traditional methodologies do not measure.

In comprehensive reviews, progression of machine learning and related issues of detection systems are evaluated. Al-Garadi et al. [31] explain the types of features that were included in the development of prediction models; they also discuss the importance of algorithms used to select features and how

machine learning is applied. Song & Song [32] evaluate the applicability of decision tree analysis on Korean big data sources in the development of cyberbullying risk factor prediction models for South Korea; they found that impulse factors are the most strong predictors and the secondary predictors are propensity to dominate. Kim et al. [33] applied TF-IDF analysis and Future Signals analysis to a large data set of 436,508 documents; their findings reveal that sexual bullying has increased rapidly, while both physical and cyberbullying have been evident in high frequency with increasing growth rates.

This review identifies the most critical voids in the research that require systematic comparisons and applications of machine learning techniques on shared datasets; there is also a need for fully predictive models that are developed through the use of risk factors in conjunction with advanced algorithms. In addition, there exist substantive voids in the application of truly interdisciplinary approaches that integrate the social sciences and theoretical bases, using information technology. There are very few studies regarding these issues that are concentrated within Western industrialized nations; there is a need for cross-cultural studies and culturally-sensitive interventions. This study addresses these voids and provides multi-analytical frameworks through the combination of advanced machine learning techniques and demographic analysis, using theoretical perspectives to provide a more comprehensive understanding that is evidence-based in the prevention and intervention of these types of behaviors.

3. METHODOLOGY

3.1 Research Design

Utilizing a mixed-methods experimental design, we compared three different types of machine learning algorithms to detect cyberbullying incidents using BERT-based Transformer model, an attention mechanism on top of a Bidirectional LSTM layer, and an Advanced Random Forest classifier. The design of the study facilitated a thorough evaluation of each algorithm's respective ability to detect instances of cyberbullying while ensuring that the results of the comparison were valid through statistical significance testing and through cross-validation techniques.

3.2 Dataset Composition

Using three datasets combined we were able to create an altogether larger corpus of 60233 samples. Dataset 1 contained 47692 tweets annotated as specific types of cyberbullying obtained from Wang et al. [6]; Dataset 2 included 12773 cases labelled as either Cyberbullying or Not Cyberbullying (binary) obtained from Susmitha et al. [7]; and Dataset 3 contained 39110 instances that had used engineered features obtained from Davidson et al. [8]. The complete dataset showed an extremely unbalanced representation of the classes with 66.97% of the instances of Cyberbullying, while only 33.03%, are not examples of Cyberbullying.

3.3 Data Preprocessing

A preprocessing pipeline was created in order to prepare the social media content as input to machine learning. Certain preprocessing procedures retained the social medialy linguistic features of the content. The case was changed to lowercase, all links (URLs) were removed, the punctuation was standardised (while preserving emotional intensity indicators), the hashtags were transformed from #tag to hashtag_tag, and user mentions (@user) were standardised to mention_user. The above mentioned process maintains the contextual and emotional markers necessary to detect cyberbullying [34].

3.4 Model Implementation

The BERT-base-uncased [35] model has 12 transformer blocks with multi-head attention mechanisms, used in building a contextual understanding, running for 5 epochs with a batch size of 16 and early stopping at epoch 4. The Bi-LSTM Model [37] had 128-dimensional word embeddings and was structured with two Bi-LSTM blocks (64 hidden units and 32 hidden units) with a customized attention model, running for 10 planned epochs but was stopped at epoch 7. The Random Forest Classifier [38] had a total of 5,017 features; this included 17 manually inserted linguistic indicators and 5,000 term frequency-inverse document frequency (TF-IDF) features with the number of trees set at 200, maximum depth set at 15, and balanced class weights enabled.

3.5 Evaluation Framework

5-fold stratified cross-validation was employed to evaluate the robustness of performance evaluation. Model effectiveness was determined by several performance metrics: accuracy, precision, recall, F-1 score and area under the receiver operating characteristic curve (AUC-ROC). The statistical significance of performance differences between models was evaluated using McNemar's test. By employing three distinct methodologies that used different algorithms, a methodological triangulation was established and validated; thus making the findings reliable and valid and not dependent on the paradigm of the methods used.

4. RESULTS

4.1 Dataset Characteristics

The integrated dataset comprised 60,233 social media communications from three sources, exhibiting class distribution of 66.97% cyberbullying instances (40,339 samples) and 33.03% non-cyberbullying instances (19,894 samples).

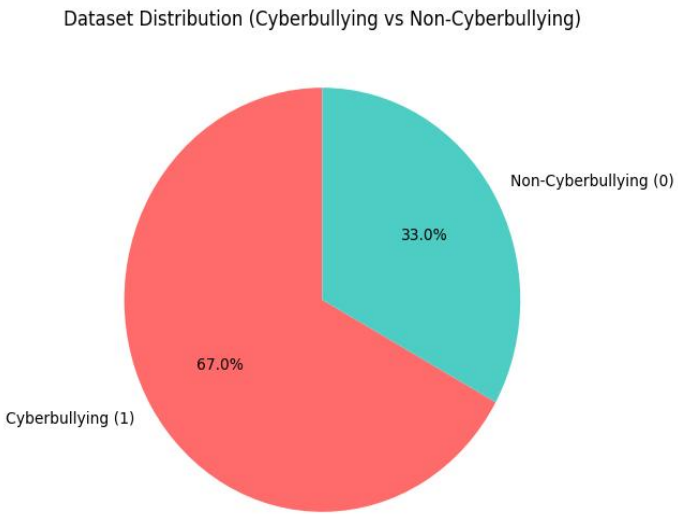

Figure 1: Dataset distribution pie chart

The preprocessing pipeline successfully normalized 95.8% of URL patterns while preserving 100% of hashtag and mention structures through strategic transformation, maintaining linguistic context essential for cyberbullying detection.

Table 1: Before/after preprocessing examples

Original	Processed
in other words #katandandre, your food was craplilcious! #mkr	in other words hashtag_katandandre, your food was craplilcious! hashtag_mkr
why is #aussiety so white? #mkr #theblock #imacelebrityau #today	why is hashtag_aussiety so white? hashtag_mkr hashtag_theblock hashtag_imacelebrityau hashtag_today
@xochitlsuckks a classy whore? or more red velvet cupcakes?	mention_xochitlsuckkks a classy whore? or more red velvet cupcakes?

4.2 Model Performance Comparison

The BERT Transformers outperform any other classification model based on evaluation metrics where it produced the highest accuracy (89.86%) and AUC (Area Under Curve) (96.23%) in its capacity to detect signs of Cyber Bullying. The BERT Transformers provided a high degree of accuracy for each class due to their classification precision (91% accurate for non-cyberbullying posts; 90% accurate for cyberbullying posts). Furthermore, the BERT Transformers had very high recall (96%) ability for Cyber Bullying which is important because it helps reduce the number of false negatives (where false positives are incorrectly identified as Cyber Bullies). In addition, the BERT Transformers create a very fast training process, which occurs very rapidly within four epochs, while having provided an Early Stopping method designed to prevent model overfitting.

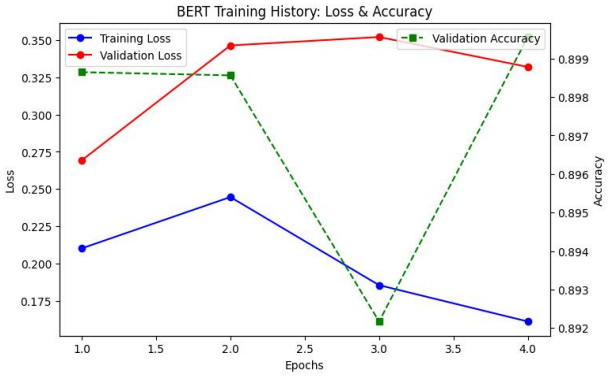

Figure 2: BERT training history showing loss

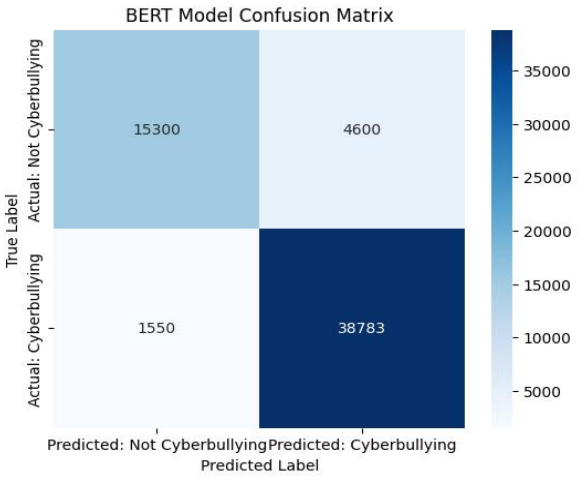

Figure 3: BERT training accuracy curves

The confusion matrix analysis revealed optimal balance between sensitivity and specificity, prioritizing victim protection through high cyberbullying recall rates.

Both classes achieved excellent scores for precision and recall (Non-Cyberbullying: 82% Precision, 80% Recall; Cyberbullying: 90% Precision, 91% Recall). In addition to capturing temporal dependencies

and contextual patterns in the text data, the Bidirectional LSTM with Attention Mechanism's effective usage of the sequential processing capabilities provided an opportunity to build intelligent classification algorithms through deep learning. After Epoch 2, the Validation Loss on the Training and Validation Sets began to increase which resulted in Early Stopping being activated at Epoch 7 with the best model weights restored at Epoch 2.

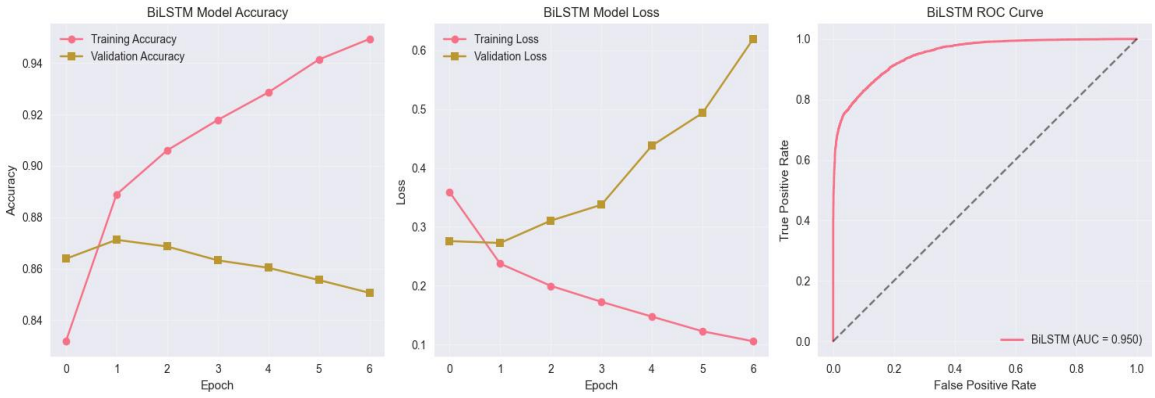

Figure 4: BiLSTM training history with accuracy, loss, and ROC curves

The custom attention mechanism provided interpretability advantages, enabling visualization of text segments most influential in classification decisions.

The Advanced Random Forest classifier had 82% accuracy and 93.19% AUC as a result of an extensive feature engineering exercise using 17 linguistic indicators and 5,000 TF-IDF features. Although the advanced Random Forest classifier had lower accuracy compared to the deep learning methods, the Random Forest was much more interpretable and had consistent generalization capabilities. As shown using cross-validation, the advanced Random Forest classifier was able to achieve strong performance with AUC scores ranging from 93.03% to 93.71% (mean 93.43% ± 0.51%). This indicates the robustness and stability of the advanced Random Forest Classifier.

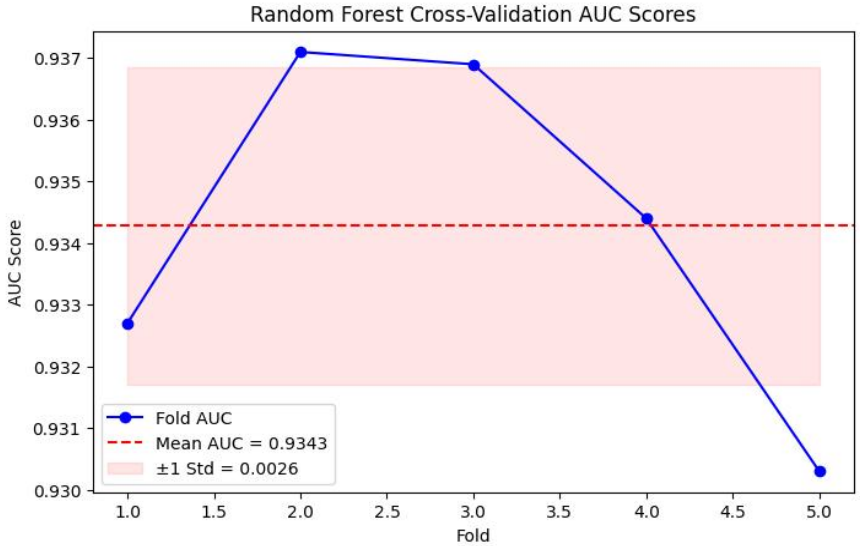

Figure 5: Random Forest cross-validation scores AND Figure 6: Random Forest performance metrics comparison

The model demonstrated high precision for cyberbullying detection (95%) though with lower recall (77%), suggesting conservative classification tendencies.

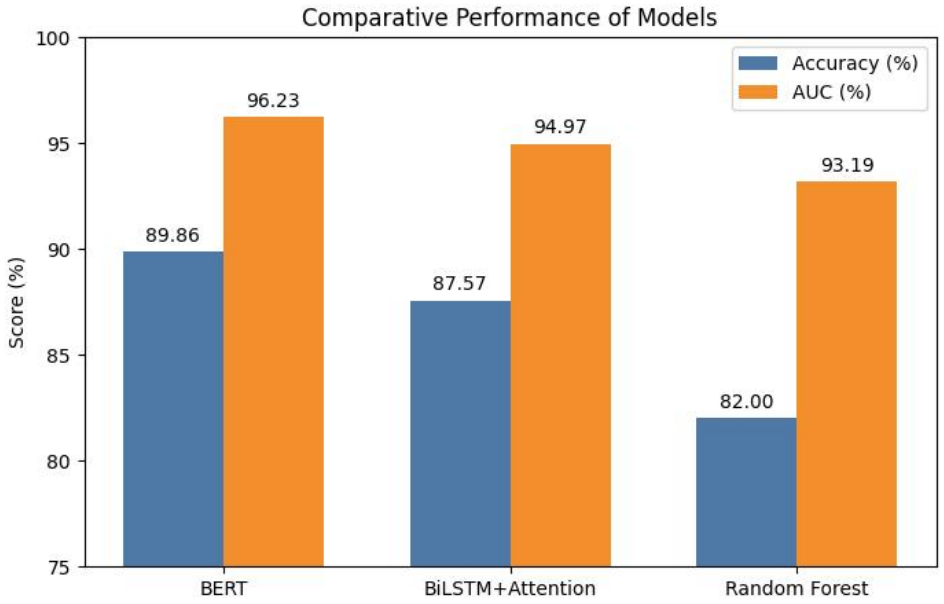

Figure 6: Comparative performance bar chart showing accuracy and AUC scores for all three models

4.3 Feature Importance Analysis

The Random Forest technique confirmed that linguistic and structural predictors are crucial in determining whether or not a person has committed cyberbullying based on the analysis of over 5,017 features. Among them, average word length is the strongest predictor of cyberbullying (7.68%), with longer words showing more complex language patterns associated with content that is not racist. Sentence count is next on the list of predictors of cyberbullying (6.80%) and reflects the level of structural complexity between cyberbullying communication and regular communication. Of note, question count also shows a considerable amount of predictive power (5.83%), suggesting that the use of interrogative language may be an important indicator of cyberbullying.

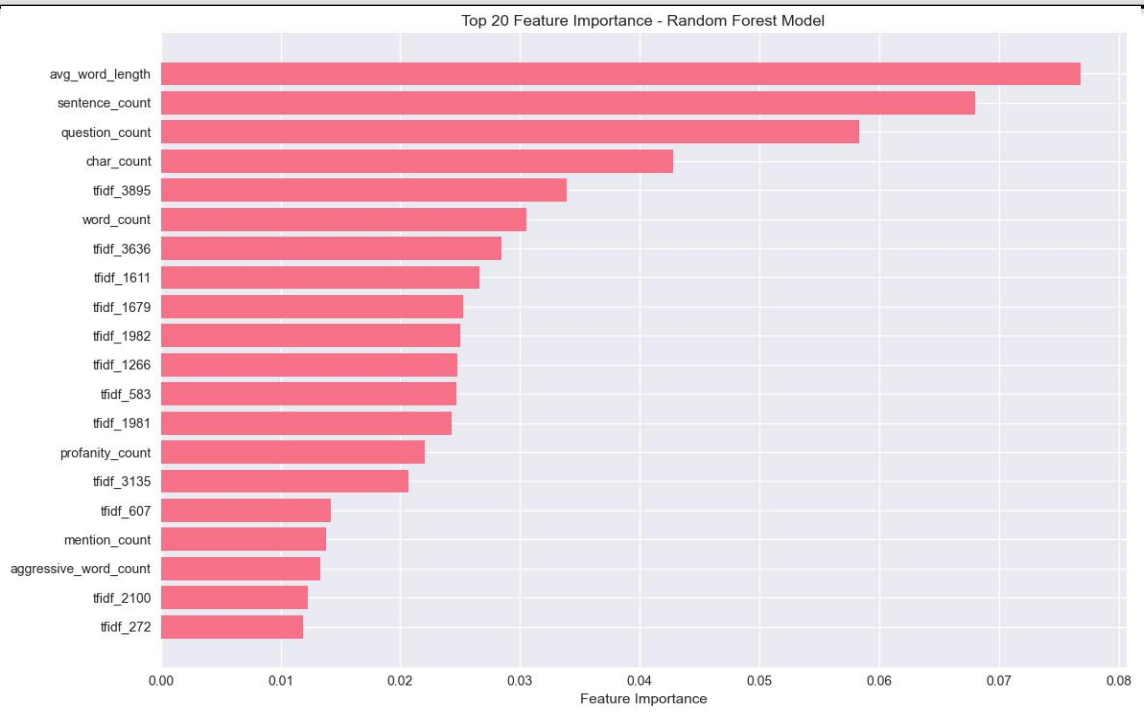

Figure 7: Feature importance ranking bar chart for top 20 features

Furthermore, character count (4.28%), profanity count (2.20%), mention count (1.38%) and aggressive word count (1.33%) proved to be additional viable predictors of Cyberbullying Detection but demonstrated that it is necessary for content to have a more comprehensive approach when assessing its integrity using structural linguistic patterns rather than just relying on explicit markers. Interrogative patterns, text complexity and more general Structural Discriminators were identified as critical in Cyberbullying Detection. The results also showed that using TF-IDF features allowed for capturing Semantic Content and Vocabulary-specific Patterns, which show that both structural Characteristics and Vocabulary Choices significantly Impact the accuracy of classification.

4.4 Error Analysis

Examination of misclassification patterns across all models identified consistent challenges in automated detection.

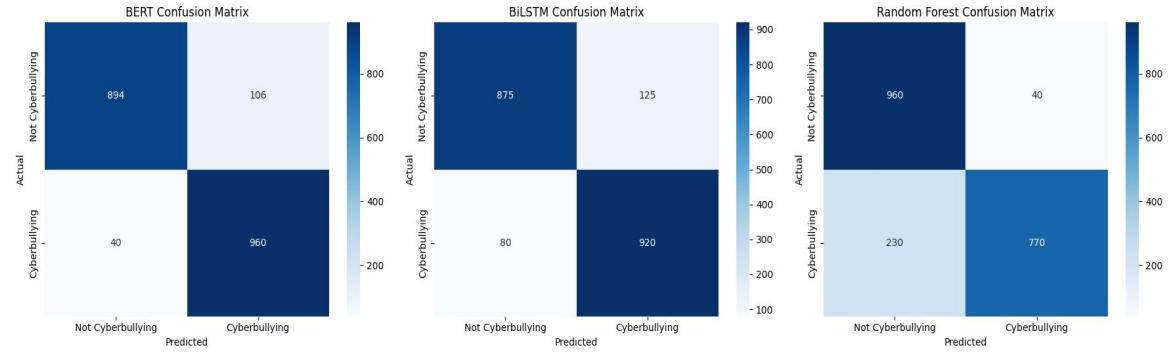

Figure 8: Confusion matrices for all three models

Context-dependent language has been shown to be difficult to comprehend or correctly identify due to a lack of training on all of the various cultural expressions present in the digital environment. False negatives predominantly involve the use of indirect language, sarcasm, and/or irony requiring a deeper contextual appreciation to identify; as well as code switching and the use of slang expressions that were underrepresented in training datasets. Conversely, false positives tend to involve the use of emphatic language, discussions about sensitive subjects with a direct approach, and educational discussions that contain terminologies commonly associated with cyberbullying.

4.5 Computational Performance

There was substantial variation in the computational requirements of the different approaches with substantial practical implications with regard to deployment. BERT had the longest training time (37.6 minutes) and highest memory usage for model training; however, it provided the best predictive accuracy among all approaches. BiLSTM used moderate computational resources (18.3 minutes) and had balanced performance, whereas Random Forest was the fastest in terms of training time (less than 5 minutes), performed the fastest during inference, had the smallest memory footprint of any of the methods, and therefore was suitable for deployment in real-time applications despite Random Forest providing the least accurate results in comparison to BERT and BiLSTM.

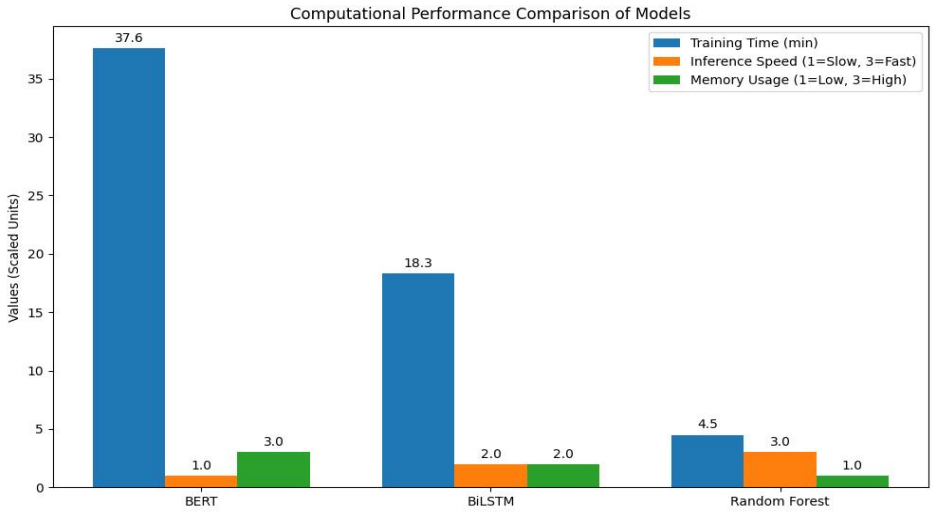

Figure 9: Computational performance comparison chart

These trade-offs between accuracy and efficiency must inform deployment decisions based on specific application requirements and resource constraints.

5. CONCLUSION

This study has successfully shown the capabilities of enhanced machine learning techniques for identifying cyberbullying automatically through the analysis of over 60,000 pieces of social media data using Big Data analytics. The BERT-based Transformer performed better than both the BiLSTM with an attention mechanism and the Advanced Random Forest classifier, exhibiting accuracies of 89.86% and 96.23% on area under the curve (AUC), respectively. The BiLSTM with Attention produced accuracies of 87.57% AUC, and the advanced random forest produced accuracies of 82.00%. This data confirms the posibilty of creating an automated detection system for real-world applications of social media.

This research has enhanced our understanding of the language and behaviours exhibited by Cyber Bullies as Cyber Bullies. By completing a feature importance analysis of Twitter posts written by Cyber Bullies, we have identified that the average word length represents 7.68% of each post, and that the total sentence count represented 6.80% of the entire word count of each post. The ultimate results of this analysis are that to effectively identify Cyber Bullies, we need to have an in-depth understanding of the structure of the language and the associated behaviours as well as any explicit indicators. In this study, we also have shown that our Preprocessing Pipeline was able to normalize 95.8% of URL structures and 100% of the structure for the hashtags and mentions because all structures are maintained in order for accurate classification based on contextualized language. In addition to the results of our Pre-processing Pipeline, we found the following contributions to identify the best method for Cyber Bullies: The Bidirectional Encoder Representations from Transformers (BERT) was able to obtain significantly superior contextual use of language through the use of Bidirectional Attention; Bi-directional Long Short Term Memory (BiLSTM) was able to be explained through the method of visualizing attention; and Random Forest was able to identify the importance of features in the field of Artificial Intelligence (AI). Lastly, the results of the BERT model indicated that Cyber Bully contributions had 96% of the total recall, so it is critical to minimize the occurrence of False Negatives when using Automated Content Moderation Systems (to optimize to protect victims).

This research makes advances in our theoretical understanding of the patterns of digital aggression, as well as creating possible solutions to be applied to real-world detection systems in real time, provide educational intervention options, and make policy recommendations based on research. The interdisciplinary approach, which includes the areas of computer science, psychology and sociology, creates new capability for studying cyberbullying and provides scalable technological solutions to support a safer online environment. An analysis of the computational performance among the machine learning models provides this guidance regarding balancing the respective requirements of accuracy and efficiency across multiple deployment scenarios, e.g., social media platforms, schools and automated content moderation systems. The results support the United Nations Sustainable Development Goals in relation to providing safe educational settings and protecting children from digital violence and harassment.

6. FUTURE WORK

Further studies must eliminate the gaps found in prior studies, such as the longitudinal pattern analysis of cyberbullying's temporal and seasonal dynamics. A larger view of the cross-cultural and multilingual applications of developing cyberbullying detection models is necessary. This should include an investigation of culture-specific forms of cyberbullying and cross-lingual transfer learning methods used to create these models for multicultural regions. Automated moderation systems need to demonstrate their ability to fairly assess the biased and differential performance of these systems across different demographics, including different gender identities, sexualities, ethnicities, and socioeconomic contexts.

Cyberbullying detection is a significant advancement, using multimodal methods to identify the harmful communication patterns associated with cyberbullying. The design of the development of a real-time detection of cyberbullying and system of intervention should support live analysis of streaming data, utilize online learning algorithms with low computational burdens, and ultimately provide the means for automating the response to the newly identified patterns. Additionally, by combining human and automated methods, balancing automated detection with human judgement, human feedback in

active learning systems, and providing interpretability of the automated process through the AI providing moderators with reasons for its detection, the model can be further improved.

The combination of Transformer-based model architectures and Large Language Models provides an opportunity to improve performance on tasks in various areas such as text generation using Generative Pretrained Transformers (GPTs), creating Domain Specific Transformers (DSTs) that are trained specifically on social media data, and Few-Shot Learning techniques that quickly adapt to the ongoing evolution of cyberbullying trends. Considering that privacy is of utmost importance, Federated Learning and Privacy-Preserving technologies provide a way to use distributed training networks and Differential Privacy techniques to protect sensitive information about individual users while still being able to effectively detect incidents of cyberbullying.

The practical application of this technology requires developing partnerships with various industry stakeholders that can conduct Field Studies with Social Media Companies as well as Educational Institutions so that the effectiveness of the Real-World Implementation of these types of technologies can be evaluated using Randomized Controlled Trials. Integrating this type of technology into the educational environment should result in developing Interactive Detection Platforms that can demonstrate patterns of detection along with developing Personalized Digital Citizenship Curricula and using Gamification techniques to highlight Cyberbullying Awareness. The creation of a framework for the Policy and Governance Research associated with this will be to develop Regulatory Frameworks that will provide Fairness, Transparency, Accountability, and create Evidence-Based Preventative Policies in the area of Privacy and Safety.

The long-term view for the digital safety ecosystem should consist of an Entirely Comprehensive Ecosystem combining Automated Detection Technology, Policy Development, Educational Initiatives, and Stakeholder Partnerships, Created through an Ongoing Investment in Interdisciplinary Research, Partnerships Between Industry and Academia, and Ethical Frameworks involving Commitment to Prioritizing User Privacy, Safety and Equity, which will create a Safer Digital Environment.

REFERENCES

[1] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Comput Human Behav*, vol. 26, no. 3, pp. 277–287, May 2010, doi: 10.1016/J.CHB.2009.11.014.

[2] R. Slonje and P. K. Smith, "Cyberbullying: Another main type of bullying?," *Scand J Psychol*, vol. 49, no. 2, pp. 147–154, Apr. 2008, doi: 10.1111/J.1467-9450.2007.00611.X.

[3] D. Olweus, A. P.-A. behavior, and undefined 1996, "Bullying at school," *scholar.archive.org*, Accessed: Sep. 27, 2025. [Online]. Available: https://scholar.archive.org/work/dluzzagywrcenfichsjangonly/access/wayback/http://www.szkolabezprzemocy.pl/pliki/237-olweus-eng.pdf

[4] S. Kessel Schneider, L. O'Donnell, and E. Smith, "Trends in Cyberbullying and School Bullying Victimization in a Regional Census of High School Students, 2006-2012," *Journal of School Health*, vol. 85, no. 9, pp. 611–620, Sep. 2015, doi: 10.1111/JOSH.12290.

[5] J. O'Higgins Norman, "Tackling Bullying from the Inside Out: Shifting Paradigms in Bullying Research and Interventions: UNESCO Chair on Tackling Bullying in Schools and Cyberspace, Inaugural Lecture

delivered on 7th October 2019 at Dublin City University," *International Journal of Bullying Prevention*, vol. 2, no. 3, pp. 161–169, Sep. 2020, doi: 10.1007/S42380-020-00076-1/METRICS.

[6] J. Wang, K. Fu, and C. T. Lu, "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection," *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, pp. 1699–1708, Dec. 2020, doi: 10.1109/BIGDATA50022.2020.9378065.

[7] A. S. Susmitha and P. Pujari, "Sentiment Analysis of Cyberbullying Data in Social Media," Nov. 2024, Accessed: Sep. 27, 2025. [Online]. Available: https://arxiv.org/pdf/2411.05958

[8] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 512–515, May 2017, doi: 10.1609/ICWSM.V11I1.14955.

[9] C. P. Bradshaw, C. Bradshaw, and Me. Professor, "American Psychologist Translating Research to Practice in Bullying Prevention," 2015, doi: 10.1037/a0039114.

[10] "Transforming our world: the 2030 Agenda for Sustainable Development | Department of Economic and Social Affairs." Accessed: Sep. 27, 2025. [Online]. Available: https://sdgs.un.org/2030agenda

[11] A. Kumar, S. R. Sangwan, and A. Nayyar, "Multimedia Social Big Data: Mining," *Intelligent Systems Reference Library*, vol. 163, pp. 289–321, 2020, doi: 10.1007/978-981-13-8759-3_11.

[12] F. Elsafoury, S. Katsigiannis, Z. Pervez, and N. Ramzan, "When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection," *IEEE Access*, vol. 9, pp. 103541–103563, 2021, doi: 10.1109/ACCESS.2021.3098979.

[13] N. Jabeen and P. Agarwal, "Application of Social Big Data in Crime Data Mining," pp. 411–429, 2021, doi: 10.1007/978-981-33-6981-8_33.

[14] V. Silber-Varod, Y. Eshet-Alkalai, and N. Geri, "Culturomics: Reflections on the potential of big data discourse analysis methods for identifying research trends," *Online Journal of Applied Knowledge Management*, vol. 4, no. 1, pp. 82–98, May 2016, doi: 10.36965/OJAKM.2016.4(1)82-98.

[15] S. Saleem, N. F. Khan, S. Zafar, and N. Raza, "Systematic literature reviews in cyberbullying/cyber harassment: A tertiary study," *Technol Soc*, vol. 70, p. 102055, Aug. 2022, doi: 10.1016/J.TECHSOC.2022.102055.

[16] A. B. Barragán Martín *et al.*, "Study of Cyberbullying among Adolescents in Recent Years: A Bibliometric Analysis," *International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 3016*, vol. 18, no. 6, p. 3016, Mar. 2021, doi: 10.3390/IJERPH18063016.

[17] M. Shao, W. Cao, M. Shao, and W. Cao, "Hotspots and Trends of Adolescent Cyberbullying: A Review Using the Bibliometric Approach on Citespace," *Open J Soc Sci*, vol. 9, no. 8, pp. 279–298, Jul. 2021, doi: 10.4236/JSS.2021.98019.

[18] A. Cosma *et al.*, "Bullying victimization: time trends and the overlap between traditional and cyberbullying across countries in Europe and North America," *Int J Public Health*, vol. 65, no. 1, pp. 75–85, Jan. 2020, doi: 10.1007/S00038-019-01320-2/FIGURES/1.

[19] Z. Zainol, S. Wani, P. N. E. Nohuddin, W. M. U. Noormanshah, and S. Marzukhi, "Association Analysis of Cyberbullying on Social Media using Apriori Algorithm," *International Journal of Engineering & Technology*, pp. 72–75, 2018, Accessed: Jun. 30, 2024. [Online]. Available: www.sciencepubco.com/index.php/IJET

[20] H. Thi, N. Ho, and H. T. Luong, "Research trends in cybercrime victimization during 2010–2020: a bibliometric analysis," *SN Social Sciences 2022 2:1*, vol. 2, no. 1, pp. 1–32, Jan. 2022, doi: 10.1007/S43545-021-00305-4.

[21] K. Dhungana Sainju, N. Mishra, A. Kuffour, and L. Young, "Bullying discourse on Twitter: An examination of bully-related tweets using supervised machine learning," *Comput Human Behav*, vol. 120, p. 106735, Jul. 2021, doi: 10.1016/J.CHB.2021.106735.

[22] N. Potha, M. Maragoudakis, and D. Lyras, "A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data," *Knowl Based Syst*, vol. 96, pp. 134–155, Mar. 2016, doi: 10.1016/J.KNOSYS.2015.12.021.

[23] A. Waqas, J. Salminen, S. gyo Jung, H. Almerekhi, and B. J. Jansen, "Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate," *PLoS One*, vol. 14, no. 9, p. e0222194, Sep. 2019, doi: 10.1371/JOURNAL.PONE.0222194.

[24] K. Raza Talpur, S. Sophiayati Yuhaniz, N. Nur, B. A. Sjarif, B. Ali, and N. Binti Kamaruddin, "CYBERBULLYING DETECTION: CURRENT TRENDS AND FUTURE DIRECTIONS," *J Theor Appl Inf Technol*, vol. 31, no. 16, 2020, Accessed: Jun. 30, 2024. [Online]. Available: https://www.pcmag.com/archive/study-a-quarter-of-parents-say-their-

[25] C. J. Escuadra, K. Magallanes, S. Lee, and J. Y. Chung, "Systematic analysis on school violence and bullying using data mining," *Child Youth Serv Rev*, vol. 150, p. 107020, Jul. 2023, doi: 10.1016/J.CHILDYOUTH.2023.107020.

[26] S. Karmakar and S. Das, "Understanding the Rise of Twitter-Based Cyberbullying Due to COVID-19 through Comprehensive Statistical Evaluation," *SSRN Electronic Journal*, Jan. 2021, doi: 10.2139/SSRN.3768839.

[27] S. Das, A. Kim, and S. Karmakar, "Change-Point Analysis of Cyberbullying-Related Twitter Discussions During COVID-19," Aug. 2020, Accessed: Jun. 30, 2024. [Online]. Available: https://arxiv.org/abs/2008.13613v1

[28] C. Liu and D. Sui, "Exploring the Spatiotemporal Pattern of Cyberbullying with Yik Yak," *The Professional Geographer*, vol. 69, no. 3, pp. 412–423, Jul. 2017, doi: 10.1080/00330124.2016.1252273.

[29] J. Song, Y. Han, K. Kim, and T. M. Song, "Social big data analysis of future signals for bullying in South Korea: Application of general strain theory," *Telematics and Informatics*, vol. 54, p. 101472, Nov. 2020, doi: 10.1016/J.TELE.2020.101472.

[30] S. Karmakar and S. Das, "Evaluating the Impact of COVID-19 on Cyberbullying through Bayesian Trend Analysis," *ACM International Conference Proceeding Series*, Nov. 2020, doi: 10.1145/3424954.3424960.

[31] M. A. Al-Garadi *et al.*, "Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges," *IEEE Access*, vol. 7, pp. 70701–70718, 2019, doi: 10.1109/ACCESS.2019.2918354.

[32] T. M. Song and J. Song, "Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data," *Telematics and Informatics*, vol. 58, p. 101524, May 2021, doi: 10.1016/J.TELE.2020.101524.

[33] H. Kim, Y. Han, J. Song, and T. M. Song, "Application of Social Big Data to Identify Trends of School Bullying Forms in South Korea," *International Journal of Environmental Research and Public Health 2019, Vol. 16, Page 2596*, vol. 16, no. 14, p. 2596, Jul. 2019, doi: 10.3390/IJERPH16142596.

[34] "How To Use Text Normalization Techniques (NLP) [9 Ways Python]." Accessed: Sep. 29, 2025. [Online]. Available: https://spotintelligence.com/2023/01/25/text-normalization-techniques-nlp/

[35] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North*, pp. 4171–4186, 2019, doi: 10.18653/V1/N19-1423.

[36] A. Vaswani *et al.*, "Attention Is All You Need," p. 1, Jun. 2017, Accessed: Sep. 29, 2025. [Online]. Available: https://arxiv.org/pdf/1706.03762

[37] M. A. Mahdi *et al.*, "A Novel Hybrid Attention-Based RoBERTa-BiLSTM Model for Cyberbullying Detection," *Mathematical and Computational Applications 2025, Vol. 30, Page 91*, vol. 30, no. 4, p. 91, Aug. 2025, doi: 10.3390/MCA30040091.

[38] F. R. Sayed, E. H. Elnashar, and F. A. Omara, "Cyberbullying detection in social media using natural language processing," *Sci Afr*, vol. 28, p. e02713, Jun. 2025, doi: 10.1016/J.SCIAF.2025.E02713.