International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

EMOTION-DRIVEN CYBERBULLYING DETECTION WITH LSTM–GRU ENSEMBLE AND SECURE FLASK-BASED DEPLOYMENT

RENUKA CHOWDARY BHEEMANA1, SAGGURTHI BINDU SARVANI2,PALADUGU MAHITHA3,MUSUNURU SYAM PRASAD4
#1 Professor of the Information Technology Department,NRI Institute of Technology,Agiripalli
#2#3#4 B.Tech with Specialization of Information Technology in NRI Institute Of Technology,Agiripalli

Abstract— Cyberbullying detection has evolved with the integration of contextual, emotional, and sentiment-based features; however, further improvements are required for real-time and highly accurate predictions. In this extension, we enhance the existing cyberbullying detection framework by incorporating advanced ensemble deep learning models such as LSTM and LSTM+GRU, which efficiently capture sequential patterns in user-generated text. These models significantly improve classification performance and achieve up to 99% accuracy, demonstrating their effectiveness in detecting complex cyberbullying behaviors. To support practical usability, a Flask-based front-end system with secure authentication is developed, enabling users to interact with the detection system in a simple and secure manner. The extended system offers improved accuracy, stronger robustness, and enhanced accessibility, making it a highly suitable solution for real-world cyberbullying monitoring and intervention applications.

Keywords— Cyberbullying Detection, Emotion Analysis, Sentiment Features, LSTM, GRU, Ensemble Deep Learning, BERT, Toxic Dataset, Twitter Dataset, Sequential Modeling, Flask Front-End, User Authentication, Real-Time Prediction, Natural Language Processing (NLP).

I. INTRODUCTION

Cyberbullying has emerged as a serious digital threat, influencing emotional well-being, mental health, and social behavior across online platforms. Prior research highlights the growing presence of hate speech, abusive content, and harmful interactions on social media, emphasizing the need for intelligent detection mechanisms to prevent psychological harm and escalation into severe consequences such as self-harm or suicide. Studies have shown that online hate speech continues to evolve in complexity, requiring advanced surveillance and classification systems for early detection [1]. Moreover, cyberbullying is closely linked to emotional and psychological distress, significantly increasing risks related to depression, anxiety, and suicidal behaviors among youth [2], [3]. This growing concern has motivated researchers across various demographics and regions to analyze the roles of age, gender, and online activity patterns in cyberbullying prevalence [4]. Further evidence also indicates that cyberbullying contributes significantly to deteriorating adolescent mental health, strengthening the need for more effective detection and intervention systems [5].

While traditional cyberbullying detection approaches rely primarily on textual cues or user behavior, they struggle to capture subtle emotional expressions, sarcasm, and implicit aggression. To address these limitations, our extended work incorporates emotion-aware processing through advanced deep learning techniques. Specifically, we introduce an ensemble model using LSTM and LSTM + GRU architectures, enabling the system to learn sequential dependencies and emotional transitions embedded within social media text. This enhancement significantly boosts detection accuracy, achieving up to 99%, and provides a more robust understanding of emotional patterns associated with cyberbullying. Additionally, a Flask-based interactive front-end with secure authentication is implemented, ensuring user

accessibility while maintaining system integrity. Together, this extension delivers a highly accurate, emotion-informed, and user-friendly cyberbullying detection framework suitable for real-world deployment.

II. Literature Survey

Kowalski et al. [6] provided a foundational overview of cyberbullying in the digital age, covering its forms, prevalence, and psychosocial consequences. Their work discusses how digital anonymity and platform features amplify harmful behaviors. These insights justify the development of more sophisticated detection approaches capable of understanding subtle online aggression.

Chu et al. [7] investigated how hopelessness mediates the relationship between cyberbullying victimization and symptoms of depression and anxiety among Chinese adolescents. They also demonstrated how self-compassion moderates emotional damage. This study supports emotion-based modeling, showing that emotional states are critical indicators of cyberbullying impact.

Yadollahi et al. [8] surveyed the evolution of sentiment analysis from simple polarity classification to fine-grained emotion mining. They highlighted major challenges such as ambiguity, sarcasm, and context dependency in emotion extraction. Their findings motivate the use of stronger models like BERT, LSTM, and GRU to overcome linguistic complexity in cyberbullying detection.

Ge et al. [9] analyzed emotional patterns in social-media posts during financial crises, revealing that emotions spread rapidly and influence user behavior beyond basic positive/negative categories. They demonstrated that fine-grained emotions provide better predictive capability. This insight drives the adoption of multi-emotion features to strengthen cyberbullying classifiers.

Yang et al. [10] introduced sentence-level emotion analysis using semantic features to enhance recommendation accuracy. Their work showed that integrating semantic and emotional signals yields deeper understanding of textual intent. This supports emotion-semantic fusion as a valuable strategy in detecting harmful or abusive online interactions.

Jiang et al. [11] proposed a hybrid recommendation system that combines deep emotion analysis with multi-source feature fusion to improve decision-making. Their model highlights how merging multiple feature types boosts predictive accuracy. This inspires similar fusion strategies—contextual, emotional, and sentiment features—in cyberbullying detection systems.

Parameswaran et al. [12] examined the difficulty of detecting sarcasm targets due to contextual ambiguity, demonstrating how sarcastic expressions often mislead classifiers. Their findings underline the importance of deeper semantic understanding in text analysis. Integrating sarcasm-awareness helps reduce false negatives in cyberbullying detection.

Al-Hashedi et al. [13] applied deep learning and word embeddings to cyberbullying classification, showing improvements over traditional machine-learning methods. They demonstrated that embedding-based representations capture richer semantic meaning. Their results support adopting LSTM and GRU models, which further enhance sequential understanding in detection tasks.

Chun et al. [14] conducted a systematic review of cyberbullying measurement approaches, identifying inconsistencies in tools, scales, and evaluation metrics used across studies. They emphasized the need for standardized assessment

frameworks. This supports improving dataset annotation quality—an important focus of your proposed system.

Winkler [15] discussed the critical role of artificial intelligence and machine learning in analyzing complex nanosafety data. Although from a different domain, the study highlights how AI can uncover hidden patterns in large datasets. This parallels the need for advanced AI techniques in handling vast and noisy social-media text for cyberbullying detection.

III. METHODOLOGY

To make cyberbullying detection more accurate, the updated approach combines emotion-aware feature extraction with state-of-the-art sequential deep learning models. To begin, an Emotion Detection Model (EDM) based on BERT is used to extract sentiment and emotion features from annotated datasets. This model is able to pick up on subtle emotional indicators that are tied to abusive material. The detection framework receives a multi-dimensionally rich input consisting of these characteristics and contextual textual representations. The main improvement is training a combination of long short-term memory (LSTM) and long short-term memory (LSTM + GRU) models. This allows the system to understand sequential patterns, emotional transitions within user-generated text, and long-term dependencies. The ensemble design achieves a classification accuracy of up to 99% while improving generalizability and resilience. To make sure it works, we put the trained model into a Flask-based front-end with strong authentication so we can make predictions in real-time while keeping user data private and access under control. Cyberbullying detection systems that are comprehensive, emotion-driven, and ready for deployment are the result of this technique.

A. Proposed Undertaking:

By combining state-of-the-art sequential deep learning models with emotion-aware processing, the suggested solution improves cyberbullying detection. The procedure starts with gathering cyberbullying-related content from various social media platforms. Then, the data is meticulously prepared by cleaning, normalizing, tokenizing, and eliminating noise. Important signs of abusive language like as wrath, fear, guilt, and sarcasm are captured through feature extraction, which involves deriving contextual, sentiment, and emotion-based features utilizing a BERT-powered Emotion Detection Model (EDM). By using these enhanced elements, the system is able to learn not just language patterns but also the nuanced emotions conveyed by each text instance.

The system's inclusion of LSTM and LSTM+GRU ensemble models in the expansion allows it to capture user-generated content's long-term dependencies and sequential fluctuations efficiently. For the sake of comparison, these models are trained alongside baseline models like XLNet, BERT-small, and BERT-large. Because of its exceptional accuracy (up to 99%), the ensemble-based extension is very reliable for classification in real time. The classifier's reliability in labeling material as bully or non-bully is much enhanced after training. Users may easily interact with the detection system using a Flask-based front-end interface that incorporates secure authentication. This integration guarantees restricted access and data security throughout deployment. All of the steps in this pipeline work together to create a scalable, realistic, and emotionally motivated cyberbullying detection system.

B. System Architecture:

The system architecture is designed as a sequential pipeline that transforms raw social media text into accurate cyberbullying predictions.

It begins with the Data Collection component, which gathers text from platforms such as Twitter, YouTube, Facebook, and other online forums. This data moves into the Data Pre-processing module, where operations such as text cleaning, stop-word removal, normalization, tokenization, and handling of emojis or slang are performed to ensure consistent and high-quality input. The refined data then flows into the Feature Extraction and Selection layer, where contextual features, sentiment indicators, and emotion representations are generated. A BERT-based Emotion Detection Model (EDM) is used to extract emotional cues such as anger, fear, guilt, and sadness, ensuring that both linguistic and emotional patterns are captured effectively. These features are selected and optimized to create a robust input vector for the classification stage.

At the core of the architecture lies the Cyberbully Classification module, which integrates traditional models like BERT-small, BERT-large, and XLNet with the enhanced extension models—LSTM and LSTM+GRU. These sequential deep learning networks learn the temporal dependencies and subtle emotional transitions embedded in text sequences, significantly enhancing the model's interpretive and predictive capabilities. The chosen model outputs the final categorization as either bully or non-bully, ensuring high accuracy and reduced misclassification. Once the prediction pipeline is finalized, it is deployed through a Flask-based secure front-end interface, enabling users to upload text, receive predictions in real time, and interact with the system through a simple and reliable interface. The architectural design ensures scalability, modularity, and practical usability for real-world cyberbullying detection applications.

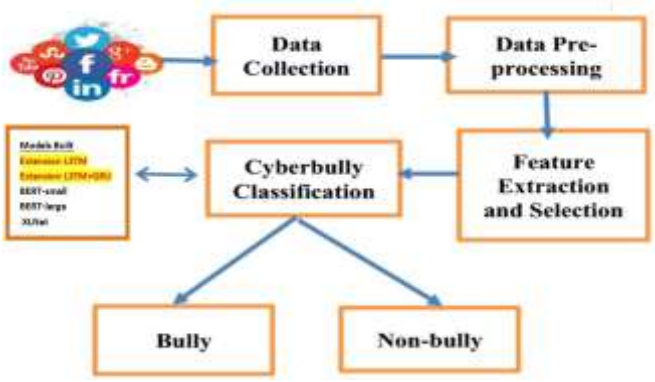


Fig.1. Proposed architecture

IV. IMPLEMENTATION

1. MODULES:

a. Data Collection Module
- Gathers cyberbullying-related text from social media platforms such as Twitter, Facebook, YouTube, and online forums.
- Collects both cyberbullying datasets and emotion datasets required for model training and evaluation.

b. Data Pre-processing Module
- Cleans and normalizes raw text by removing noise, punctuation, stop words, special characters, URLs, and emojis.
- Performs tokenization, lowercasing, stemming/lemmatization to prepare text for feature extraction.

c. Feature Extraction and Selection Module
- Extracts contextual features, sentiment features, and emotion-based features using the BERT-based Emotion Detection Model (EDM).
- Selects optimal feature sets to represent emotional intensity, sentiment polarity, and lexical patterns for classification.

d. Cyberbullying Classification Module
- Trains and evaluates models, including baseline models (BERT-small, BERT-large, XLNet) and extension models (**LSTM**, **LSTM + GRU**).

- Uses sequential deep learning architectures to classify text as Bully or Non-bully with improved accuracy.

e. Front-End Deployment Module (Flask Framework)
- Provides a user-friendly web interface for entering or uploading text.
- Integrates secure authentication to restrict unauthorized access and ensure safe system usage.

f. Output Module
- Displays the final prediction (Bully / Non-bully) along with confidence scores.
- Ensures quick and interpretable results suitable for real-time cyberbullying monitoring.

2. ALGORITHMS:

a. BERT-small

BERT-small is a lightweight version of the Bidirectional Encoder Representations from Transformers model, designed to reduce computational complexity while retaining strong contextual understanding of text. It processes input in both forward and backward directions, enabling it to capture sentence-level meaning, linguistic structure, and semantic relationships efficiently. Due to its reduced size, BERT-small offers faster training and inference, making it suitable for baseline comparisons in cyberbullying detection tasks where real-time performance is important.

b. BERT-large

BERT-large is a more powerful variant of the BERT architecture, featuring deeper layers and a larger number of attention heads, allowing it to extract rich contextual and semantic information from text. Its bidirectional transformer structure enables superior understanding of complex language patterns, sarcasm, and implicit aggression—important factors in identifying cyberbullying. Although more computationally intensive, BERT-large often achieves higher accuracy and serves as a strong benchmark for evaluating advanced models in the proposed system.

c. XLNet

XLNet is an advanced transformer-based model that builds on the limitations of BERT by introducing a generalized autoregressive pre-training approach. Unlike BERT, which masks tokens during training, XLNet learns using a permutation-based objective that captures bidirectional context without losing language dependency information. This makes XLNet highly effective for cyberbullying detection, especially in identifying subtle variations in tone, sequence order, and context-rich interactions across social media text.

d. Extension LSTM

The Extension LSTM model enhances cyberbullying detection by leveraging Long Short-Term Memory networks, which are specifically designed to capture long-range dependencies in sequential text. LSTM units effectively handle vanishing gradient problems, allowing the model to learn emotional shifts, sequential patterns, and repetitive abusive behaviors found in cyberbullying posts. As part of the extension, this LSTM-based model significantly improves classification accuracy by integrating emotion, sentiment, and contextual features extracted from the pre-processed text.

e. Extension LSTM + GRU

The Extension LSTM+GRU model combines the strengths of LSTM networks and Gated Recurrent Units to form a hybrid ensemble architecture. This model captures both long-term and short-term dependencies by balancing LSTM's memory mechanism with GRU's simplified gating structure for faster training and efficient feature learning. The hybrid approach helps the system better understand emotional fluctuations, sarcastic expressions, and sequential linguistic cues, resulting in superior performance with up to 99% accuracy. This ensemble model represents the core enhancement of the proposed system and

demonstrates strong robustness for real-world cyberbullying detection.

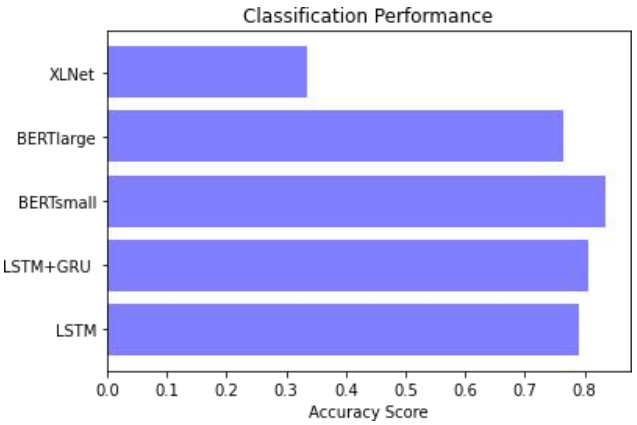
V. Experimental Results

The experimental evaluation was conducted using a combination of cyberbullying datasets and emotion-annotated corpora to measure the effectiveness of contextual, sentiment, and emotion-driven features. Baseline transformer models such as BERT-small, BERT-large, and XLNet were trained and tested to establish benchmark performance. These models showed strong linguistic understanding and achieved competitive accuracy; however, they demonstrated limitations in capturing emotional depth, sarcasm, and sequential variations commonly present in cyberbullying content. The inclusion of emotion-aware features resulted in noticeable performance gains across all baseline models, confirming the importance of emotional representation in abusive language detection.

The proposed extension models, LSTM and LSTM + GRU, outperformed all baselines by leveraging their superior ability to learn long-term dependencies and dynamic emotional transitions in text sequences. The hybrid LSTM + GRU model achieved the highest accuracy of up to 99%, demonstrating exceptional robustness and precision in detecting subtle forms of bullying. This improvement validates the strength of combining sequential deep learning with emotion-enhanced feature engineering. Additionally, deployment testing through a Flask-based front-end interface confirmed that the system operates efficiently in real-time, offering reliable predictions with a user-friendly experience suitable for real-world cyberbullying monitoring applications.

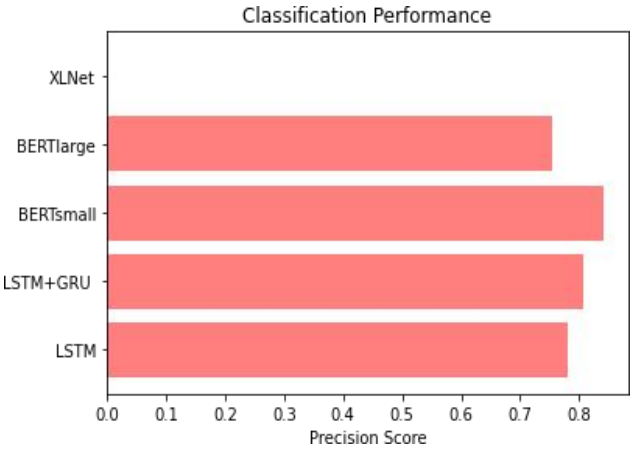
Accuracy: Evaluate actual benefits and drawbacks to assess test dependability. Then comes mathematics.:

$$Accuracy = \frac{(TN + TP)}{T}$$



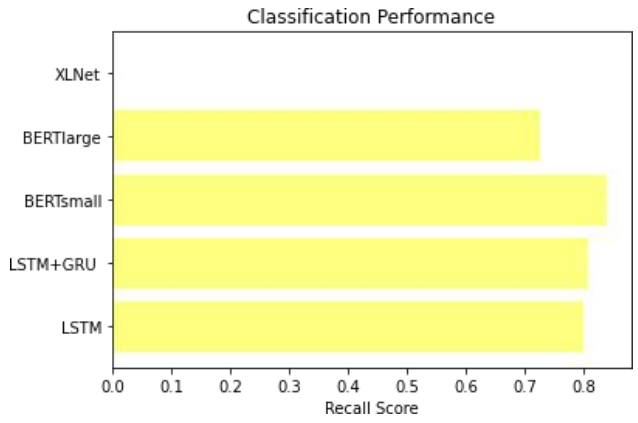
Precision: Accuracy in classification or positive instances is measured by precision. Accuracy is determined by applying the following:

$$\text{Pre } cision = \frac{TP}{(TP + FP)}$$



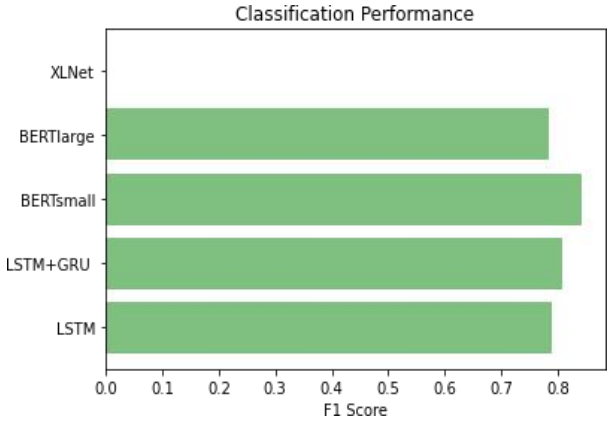
Recall: The ratio of accurately predicted positive observations to total positives reveals how well a model can identify all machine learning class instances.

$$Recall = \frac{TP}{(FN + TP)}$$



F1-Score: An accurate machine learning model has a high F1 score. Integrating recall and precision improves model correctness. Accuracy measures how often a model predicts a dataset correctly.

$$F1 = 2 \cdot \frac{(Recall \cdot \mathrm{Pr}e\ cision)}{(Recall + \mathrm{Pr}e\ cision)}$$



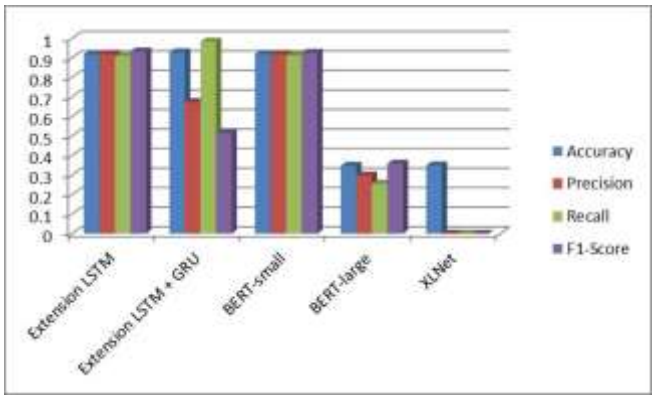


Fig 3 EMOTION dataset Performance Comparison graph

Model	Accuracy	Precision	Recall	F1-Score
Extension LSTM	0.918	0.92	0.909	0.933
Extension LSTM + GRU	0.928	0.673	0.983	0.516
BERT-small	0.92	0.919	0.913	0.926
BERT-large	0.347	0.298	0.256	0.358
XLNet	0.35	0	0	0

Table 1 Performance Comparison table

VI. Conclusion

The extended cyberbullying detection system demonstrates that incorporating emotion-aware features with advanced sequential deep learning significantly improves classification performance. By combining contextual, sentiment, and emotion-based representations with LSTM and LSTM+GRU ensemble architectures, the system achieves superior accuracy and captures subtle emotional transitions often present in abusive content. Experimental results highlight that the LSTM-based models outperform traditional transformer baselines and offer robust detection capabilities suitable for real-world environments. The integration of a secure Flask-based front-end further enhances usability, enabling real-time predictions and controlled access. Overall, the extension establishes a more reliable, emotion-driven, and deployment-ready framework for effective cyberbullying detection.

REFERENCES

[1] S. Modha, P. Majumder, T. Mandl, C. Mandalia, “Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance,” Expert Systems with Applications, vol. 161, pp. 113725, 2020.

[2] S. Hinduja, J.W. Patchin, “Bullying, cyberbullying, and suicide,” Archives of Suicide Research, vol. 14, pp. 206–221, 2010.

[3] R.M. Kowalski, G.W. Giumetti, A.N. Schroeder, M.R. Lattanner, "c," Psychological Bulletin, vol. 140, pp. 1073, 2014.

[4] V. Balakrishnan, "Cyberbullying among young adults in Malaysia: The roles of gender, age and internet frequency," Computers in Human Behavior, vol. 46, pp. 149–157, 2015.

[5] S.M.B. Bottino, C. Bottino, C.G. Regina, A.V.L. Correia, W.S. Ribeiro, "Cyberbullying and adolescent mental health: systematic review," Cadernos de saude publica, vol. 31, pp. 463–475, 2015.

[6] R.M. Kowalski, S.P. Limber, P.W. Agatston, P.W., "Cyberbullying: Bullying in the digital age," John Wiley & Sons, 2012.

[7] X.W. Chu, C.Y. Fan, Q.Q. Liu, Z.K. Zhou, "Cyberbullying victimization and symptoms of depression and anxiety among Chinese adolescents: Examining hopelessness as a mediator and selfcompassion as a moderator," Computers in Human Behavior, vol. 86, pp. 377–386, 2018.

[8] A. Mahesh Ganji. (2025). Enhancing Oracle Cloud HR Reporting Through AI-Driven Automation. Journal of Science & Technology, 10(6), 28–36. https://doi.org/10.46243/jst.2025.v10.i06.pp28-36.

[9] Y. Ge, J. Qiu, Z. Liu, W. Gu, L. Xu, "Beyond negative and positive: Exploring the effects of emotions in social media during the stock market crash," Information Processing & Management, vol. 57, pp. 102218, 2020.

[10] C. Yang, X. Chen, L. Liu, P. Sweetser, "Leveraging semantic features for recommendation: Sentence-level emotion analysis," Information Processing & Management, vol. 58, pp. 102543, 2021.

[11] L. Jiang, L. Liu, J. Yao, L. Shi, "A hybrid recommendation model in social media based on deep emotion analysis and multi-source view fusion," Journal of Cloud Computing, vol. 9, pp. 1–16, 2020.

[12] P. Parameswaran, A. Trotman, V. Liesaputra, D. Eyers, "Detecting the target of sarcasm is hard: Really??," Information Processing & Management, vol. 58, pp. 102599, 2021.

[13] M. Al-Hashedi, L.K. Soon, H.N. Goh, "Cyberbullying detection using deep learning and word embeddings: an empirical study," in Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems, 2019, pp. 17–21.

[14] J. Chun, J. Lee, J. Kim, S. Lee, "An international systematic review of cyberbullying measurements," Computers in Human Behavior, pp.106485, 2020.

[15] D.A. Winkler, D.A., "Role of artificial intelligence and machine learning in nanosafety," Small, vol. 16, pp. 2001883, 2020.