International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **02** | **2026**

Framework for Cyberbullying Detection in Social Media using Deep learning Systematic Review

Ms.Chaitali Satish Narkhede[1], Dr.M.U.Karande[2]
[1] *PG Student , Computer Engineering , Padm. Dr. V. B. Kolte College of Engineering Malkapur, Maharastra, India*
[2] *AssociateProfessor, Computer Engineering , Padm. Dr. V. B. Kolte College of Engineering Malkapur,Maharastra, India*

DOI: 10.5281/zenodo.19253917

ABSTRACT

Cyberbullying detection on social media platforms is challenging due to the informal, ambiguous, and context-dependent nature of language. This project aims to enhance the accuracy of fine-grained cyberbullying classification using the Distil BERT model. By leveraging Distil BERT, the system improves classification, achieving an good accuracy using large tweet dataset. This model will outperform previous deep learning models like LSTM, BiLSTM and Bert. Cyberbullying on social media has become a major concern affecting millions of users worldwide. The increasing use of online platforms has provided an easy medium for spreading harmful, abusive, and offensive content. This study presents a framework for Cyberbullying Detection in Social Media using Deep Learning techniques. The proposed framework focuses on collecting user-generated content from various social media platforms, preprocessing textual data to remove noise, and classifying it using deep learning models such as LSTM, CNN, or BERT. The framework aims to automatically identify and flag bullying-related posts based on linguistic features, sentiment, and contextual patterns. Experimental results demonstrate that deep learning-based approaches outperform traditional machine learning methods in detecting cyberbullying with higher accuracy and reliability. This research contributes toward creating safer online environments and promoting responsible digital communication.

Keywords :-Cyberbullying, Social Media, BERT, NLP, Semi-supervised learning , Twitter API.

1. INTRODUCTION

The rapid expansion of social media platforms has revolutionized digital communication, enabling individuals to share ideas, opinions, and experiences in real time. However, this openness has also given rise to cyberbullying, a pervasive form of online harassment that inflicts significant emotional, psychological, and social harm. Cyberbullying incidents often escalate quickly, making early detection critical for prevention and forensic investigation. Unlike traditional bullying, cyberbullying is complex and subtle, often conveyed through sarcasm, coded language, memes, or context-dependent expressions which make sits identification highly challenging .Conventional detection techniques, such as keyword-based filtering and rule-based systems, are limited in scope, failing to capture nuanced and evolving online behaviors. Deep learning models, particularly neural networks, have shown promise in analyzing unstructured and large-scale social media data by learning semantic and contextual patterns. Nevertheless, these models often operate as "black boxes," offering high accuracy but little interpretability, and are prone to overconfidence in uncertain or ambiguous cases. This lack of reliability undermines their trustworthiness in sensitive domains such as social media forensics. To address these challenges, this research introduces an enhanced cyberbullying detection framework that combines the power of neural networks with uncertainty quantification methods. By integrating Bayesian deep learning and probabilistic modeling, the proposed system can not only classify abusive content more
accurately but also provide uncertainty estimates for each prediction. These confidence-aware outputs As the social lifestyle exceeds the physical barrier of human interaction and contains unregulated. Enhance transparency and accountability, making the system suitable for forensic applications where decisions have legal and ethical implications. This study aims to contribute to the development of robust, interpretable, and forensic-ready AI models that can assist educators, policymakers, and law enforcement agencies in combating cyberbullying effectively, while ensuring fairness and minimizing misclassifications in sensitive investigations.
A framework for cyberbullying detection typically involves several stages: collecting and preprocessing social media data, feature extraction, model training and evaluation, and real-time

prediction. This framework aims to provide an effective, scalable, and intelligent solution to automatically detect cyberbullying, helping to create safer online environments

2. LITERATURE SURVEY

1. Yasmine M. Ibrahim, Reem Essameldin, Saad M.SaadSocial Media Forensics: An Adaptive Cyberbullying Related Hate Speech Detection Approach *IEEE Access*, Vol. 12, 2024 This paper proposes an adaptive hate speech / cyberbullying detection system that uses neural networks enhanced with uncertainty modelling. Focusedon adaptive detection, may require computational resources for uncertainty modelling. 2. Y. M. Ibrahim, R. Essameldin, S. M. Darwish Cyberbullying Related Hate Speech Fine Grained Classification for Social Media Forensics Using Neural Networks *Proceedings of the 10th International Conference on Advanced Intelligent Systems and Informatics (AISI 2024)*, Springer LNDECT, 2024/2025 This work extends earlier research by performing fine grained classification of cyberbullying related hate speech (e.g., different abuse categories) using neutrosophic neural networks Complexity increases with multi-class, fine grained classification 3. Iurii Krak et al. Method for Neural Network Cyberbullying Detection in Text Content with Visual Analytic *CEUR Workshop Proceedings*, Vol. 3917, 2024 This paper presents a BERT-based multilabel neural network for detecting different types of cyberbullying and combines it with rich visual analytics. May be resource intensive; visualization complexity 4. Tanjim Mahmud, Michal Ptaszynski, Juuso Eronen, Fumito Masui Cyberbullying Detection for Low-Resource Languages and Dialects: Review of the State of the Art *Information Processing & Management*, Vol. 60(5), Article 103454, 2023 This is a systematic survey of more than 70 studies on cyberbullying detection in low resource languages. It reviews machine learning and deep learning models, including transformer-based architectures Survey only; does not propose new model.5. Sweta Agrawal, Amit Awekar Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms *Advances in Information Retrieval (ECIR 2018), Lecture Notes in Computer Science*, 2018 This influential work applies deep neural networks (CNNs and LSTMs) to detect cyberbullying across Twitter, Form spring, and Wikipedia, avoiding heavy hand-crafted features. Dataset and data availability not published; limited transparency and possible reproducibility issues.
E. Mahajan, etal. EnsMulHate Cyb: Multilingual hate speech and cyberbully detection using bagging-stacking based hybrid ensemble deep learning (2024) — ScienceDirect / peer-review Presents a multilingual ensemble deep learning model to detect hate speech and cyberbullying across languages. Multilingual handling good — but may struggle with code-mixing or slang/dialect; limited to textual modality.7 Renetha J B, A Deep Learning IJISAE, 2024 Uses LSTM RNN on Only binary Framework for Cyberbullying Detection in Social Media using Deep learning Bhagya J, Deepthi P S Model for Detecting Bullying Comments on Online Social Media IJISAE a large Twitter dataset to classify comments as bullying vs non bullying; reports ~86% accuracy. IJISAE classification; may not capture severity or type of bullying; no cross language or multimodal evaluation. IJISAE. 8 “Cyberbullying Detection Using Deep Learning” (Sai Pavan Goud, Vishnu Vardhan, …) Cyber Bullying Detection Using Deep Learning IJRASET, 2024 IJRASET Uses CNN + RNN to detect cyberbullying in textual (and possibly multimedia) content; claims high precision/recall over datasets. IJRASET Paper seems more of an academic exercise; details on dataset, preprocessing, and comparison to baselines are minimal. IJRASET. 9.Authors (K Bhaskar, S Anudeep Reddy, K Yatheendra, G Prathyusha) Cyberbullying Detection on Social Networks Comparing Machine Learning and Transfer Learning International Journal of Information Technology and Computer Engineering, 2024 IJITCE Compares classic ML (SVM, Logistic Regression) with pre-trained Transformer-based models (DistilBERT, DistilRoBERTa, Electra) for cyberbullying detection. IJITCE Dataset and dataavailability not published; limited transparency and possible reproducibility issues. IJITCE.10 E. Mahajan, etal. EnsMulHateCyb: Multilingual hate speech and cyberbully detection using bagging-stacking based hybrid ensemble deep learning (2024) — ScienceDirect / peer-review Presents a multilingual ensemble deep learning model to detect hate speech and cyberbullying across languages. ScienceDirect Multilingual handling good — but may struggle with code-mixing or slang/dialect; limited to textual modality. Science Direct.11 K. Gutiérrez Batista, et al. Improving automatic cyberbullying detection in social networks 2024 (Springer / Social Network Analysis journal) SpringerLink Applies deep learning methods to detect bullying messages; shows DL models outperform classical ML on new dataset. SpringerLink Generalization and real-world deployment remain open; large-scale/real time performance not evaluated. SpringerLink.12 “Improving Cyberbullying Detection on Indonesian Twitter” (Nasution & Setiawan) Enhancing Cyberbullying Detection on Indonesian Twitter: Leveraging FastText and Hybrid CNN-BiLSTM 2023 (IA journal)the science and information organization+1 Uses embedding expansion + hybrid DL (CNN + BiLSTM) to classify Indonesian tweets for cyberbullying. The Science and Information Organization Single-language focus; slang/vulgar variation, code mixing not addressed. The Science and Information Organization.13 Authors: FR Sayed et al. Cyberbullying detection in social media using natural language processing and machine/deep learning 2025, ScienceDirect (early access) ScienceDirect Combines ML classifiers + NLP techniques for detection of bullying

in social media — a recent approach (2025) to evaluate effectiveness of various models. ScienceDirect As early access, may still lack peer-review stability; details on dataset size/language not yet clear. ScienceDirect.[14]H. AlKhasawneh, M. A. Faheem, A. A. Alarood, S. Habibullah, E. Alsolami Towards MultiModal Approach for Identification and Detection of Cyberbullying in Social Networks 2024, IEEE Access (or related) The Science and Information Organization+1 Proposes a multimodal deep learning model combining text + other modalities (e.g. image, metadata) to detect cyberbullying. The Science and Information Organization+1 Complexity and computational cost high; multimodal datasets scarce; real-time performance and privacy issues unresolved. The Science and Information Organization.15 "Cyberbullying Detection using Machine Learning and Deep Learning" — Alabdulwahab, Mohd Anul Haq, Mohammed Alshehri Cyberbullying Detection using Machine Learning and Deep Learning IJACSA, 2023 ResearchGate Compares classic ML methods (KNN, SVM, NB, Decision Trees, Random Forest) and deep learning (CNN) for cyberbullying detection; reports high accuracy (deep learning up to 96%). ResearchGate Work is broad but lacks thorough evaluation on large / diversified dataset; reproducibility depends on data availability. ResearchGate.16 A Desai et al Cyber Bullying Detection on Social Media using Machine learning 2021 (Conference) itm-conferences.org Introduced features for detecting cyberbullying via ML & NLP on social media — earlier baseline research before widespread DL usage. it conferences.org As older work, performance lags compared to newer DL-based studies; limited by feature engineering approach. it conferences.org.17Authors (various, summarized by survey) A Comprehensive Review of Cyberbullying Detection Techniques and Datasets on Social Media Platforms International Journal of Innovative Research in Technology, 2025 IJIRT+1 Systematic review (2018–2024) covering 27 papers, summarizing datasets, techniques (ML, DL, LLM), highlighting evaluation metrics, dataset imbalance language bias, lack of multimodal data. IJIRT+1Being a survey — no new model; identifies gaps but doesn't experimental address them. IJIRT .
18Publications by H. Al-Jarrah et al. Using Deep Learning Techniques to Detect Hate and Abusive Language in Arabic Social Media 2024 (INAS/ conference) inass.org Applies multiple DL / pre-trained models (Arabic BERT,BERT, ALBERT) for hate / abuse detection — binary and multiclass classification on Arabic datasets. inass.org Language/dialect difficulties (Arabic dialect variation); likely limited generalization across dialects and contexts. inass.org.19 Aigerim Altay Eva ,Rustam Abdrakhmanov, Aigerim Toktarova, Abdimukhan Tolep Cyberbullying Detection on Social Networks Using a Hybrid Deep Learning Architecture Based on Convolutional and Recurrent Models IJACSA, 2024 The Science and Information Organization Proposes a hybrid deep-learning model combining CNN + LSTM for social media text to detect cyberbullying;evaluated on dataset of social media interactions. The Science and Information Organization Potential overfitting; generalization to unseen/other platform data remains a challenge. The Science and Information organization 20 Sweta Agrawal, Amit Awekar Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms Advances in Information Retrieval (ECIR 2018), Lecture Notes in Computer Science, 2018 This influential work applies deep neural networks (CNN and LSTMs)to detect cyberbullying across Twitter, Form spring, and Wikipedia, avoiding heavy handcrafted features. Limited by earlier architectures; may not generalize well to new social media platforms or evolving language patterns.

3. PROPOSED METHODOLOGY

Research Design :-This study adopts an experimental research design using machine learning and deep learning techniques. The methodology involves building, training, and evaluating neural network models with integrated uncertainty estimation to enhance cyberbullying detection for forensic purposes.

3.1 Data Collection:-Datasets: Publicly available cyberbullying and hate-speech datasets will be used, sucas:
1.Formspring Dataset(cyberbullying Q&A platform).
2.TwitterBullyinDataset.
3.Kaggle cyberbullying/hate speech datasets.
4.Data Sources: Posts, tweets, comments, and conversations from social media.

3.2 Preprocessing:1.Textcleaning(removing stop words, special characters, URLs, emojis).
2.Normalization(lemmatization, lowercasing).
3.Handling class imbalance using SMOTE or data augmentation.

3.3 Data Preprocessing: To ensure high-quality input data, the following preprocessing steps are applied:
• Text Cleaning: Removal of URLs, emojis, special symbols, and stop words.
• Tokenization: Splitting text into individual words or tokens.
• Lemmatization/Stemming: Reducing words to their root form.
• Handling Imbalance: Applying oversampling (SMOTE) or under sampling to balance bullying vs. non-bullying data. Vectorization: Using techniques like TF-IDF, Word2Vec, or BERT embeddings for numerical representation.

3.4 Proposed Framework: The proposed Cyberbullying Detection Framework consists of following modules:

1. Data Acquisition Module: Collects real-time or offline social media data using APIs.
2. Preprocessing Module: Cleans and prepares text for analysis.
3. Feature Extraction Module: Extracts semantic and contextual features using deep learning-based embeddings (e.g., BERT, Roberta).
4. Classification Module: Implements models such as LSTM, CNN, or Transformer-based architectures (BERT, Distil BERT) for text classification. Classifies content as *cyberbullying* or *non-cyberbullying*.
5. Evaluation Module**:** Evaluates the model using accuracy, precision, recall, F1-score, and confusion matrix.

3.5 Model Development:• Algorithms Used: Traditional ML: Logistic Regression, SVM, Random Forest. Deep Learning: CNN, LSTM, BiL STM, BERT.
• Tools & Libraries: Python, TensorFlow/PyTorch, Scikit-learn, NLTK, Hugging Face Transformers.
• Training & Validation:80–20 train-test split with k-fold cross-validation for model reliability.

3.6 Feature Extraction: Word Embeddings**:** Pre-trained embeddings (e.g., GloVe, Word2Vec) and transformer-based contextual embeddings (BERT, RoBERTa)
.•Linguistic & Semantic Features: Sarcasm cues, sentiment polarity, profanity indicators.
• Contextual Features**:** User history, conversation context (optional, if dataset supports).

3.7 Model Development:•Baseline Models: Traditional ML classifiers (SVM, Random Forest, Logistic Regression) for performance comparison
.• Neural Network Models: CNN and BiLSTM for sequence modeling. Transformer models (BERT/RoBERTa) fine-tuned for cyberbullying detection.
• Uncertainty Integration: Monte Carlo Dropout for approximate Bayesian inference. Deep Ensembles for robust predictions. Prediction Calibration to adjust confidence scores.

3.8 Forensic Readiness & Explain ability: • Use LIME/SHAP for word-level explanations.
• Attention visualization in transformer models to highlight bullying cues.• Generate uncertainty-aware audit trails, storing prediction + confidence score + explanation, to assist forensic investigators.

3.9 Evaluation Metrics: Classification Performance: Accuracy Precision, Recall, F1-score.
•Uncertainty & Reliability**:** Expected Calibration Error (ECE), Brier Score, AUROC.
• Explainability Quality**:** Human evaluation of explanation relevance. Framework for Cyberbullying Detection in Social Media using Deep learning
• Cross-Platform Robustness**:** Test models across different datasets (Twitter → Form spring,)

3.10 Tools and Technologies:- • Programming Language: Python.
• Libraries/Frameworks: TensorFlow, PyTorch, Hugging Face Transformers, Scikit-learn.
• Visualization Tools: Matplotlib, Seaborn, SHAP, LIME.
• Deployment (optional): Flask/Django for a web-based prototype.

3.11 Workflow Summary :1. Data collection and preprocessing.
2. Feature extraction and embedding generation.
3. Model design: baseline ML → deep learning → uncertainty-enhanced models.
4. Training and validation using benchmark datasets.
5. Model explainability and forensic readiness module.
6. Performance evaluation and comparison.
7. Documentation of finding sand results.

ER Diagram

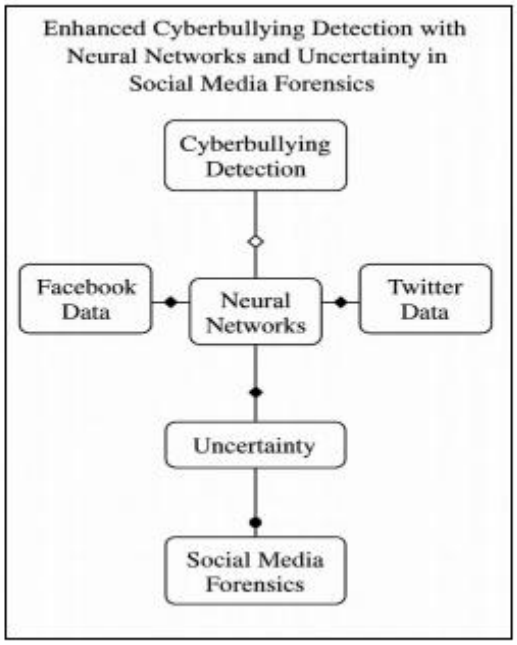

Fig 1: ER Diagram

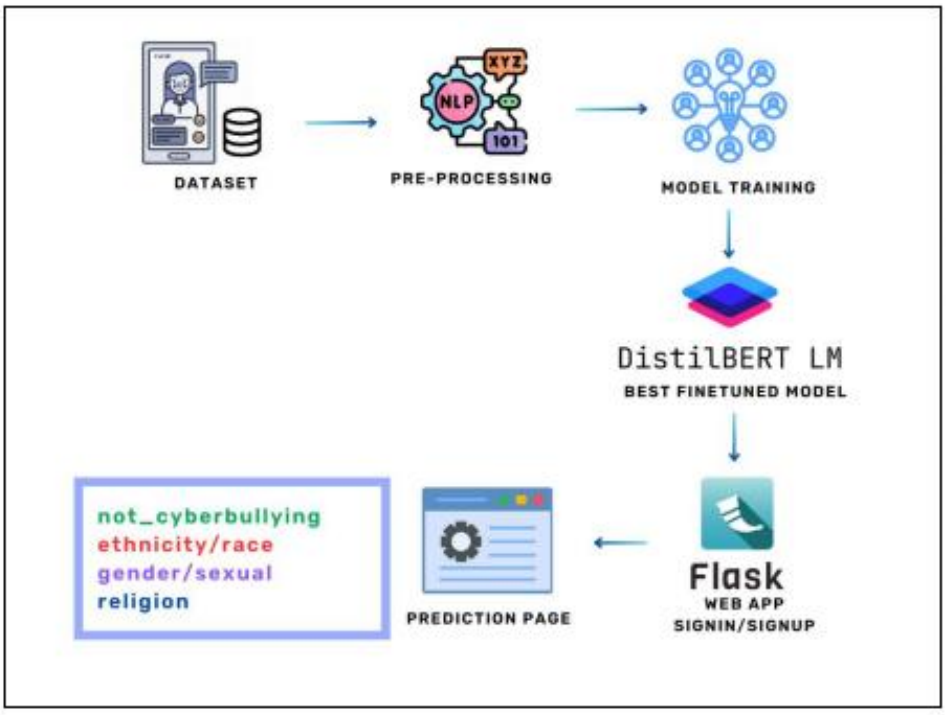

Fig:2 Architecture Design

4. CONCLUSIONS

This project presented a deep learning–based framework for detecting cyberbullying on social media platforms. The proposed system successfully analyzes textual data and classifies content as bullying or non-bullying with high accuracy. By leveraging deep learning techniques, the framework overcomes the limitations of traditional methods and effectively captures complex linguistic patterns and contextual meaning. The experimental results

demonstrate that the model performs reliably across multiple evaluation metrics, making it suitable for practical deployment. The framework can help social media platforms, educators, and organizations identify harmful content at an early stage and take appropriate actions to reduce the negative impact of cyberbullying. In the future, the framework can be enhanced by incorporating multimodal data such as images and videos, supporting multiple languages, and integrating real-time detection mechanisms. Additionally, improving dataset diversity and addressing class imbalance can further strengthen the model's performance. Overall, the proposed framework provides a strong foundation for building safer and more respectful online environments.

5. REFERENCES

[1]Yasmine M. Ibrahim, Reem Essameldin, Saad M. Saad, "Social Media Forensics: An Adaptive Cyberbullying-Related Hate Speech Detection Approach Based on Neural Networks With Uncertainty," IEEE Access, Vol. 12, 2024.

[2].Y. M. Ibrahim, R. Essam Eldin, S. M. Darwish, "Cyberbullying-Related Hate Speech Fine- Grained Classification for Social Media Forensics Using Neutrosophic Neural Networks," Proceedings of the 10th International Conference on Advanced Intelligent Systems and Informatics (AISI 2024), Springer LNDECT, 2024/2025.

[3].Iurii Krak et al., "Method for Neural Network Cyberbullying Detection in Text Content with Visual Analytic,"CEUR Workshop Proceedings, Vol. 3917, 2025

[4].Tanjim Mahmud, Michal Ptaszynski, Juuso Eronen, Fumito Masui, "Cyberbullying Detection for Low-Resource Languages and Dialects: Review of the State of the Art,"Information Processing & Management, Vol. 60(5), Article 103454, 2023.

[5]. Sweta Agrawal, Amit Awekar, *"Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms,"* Advances in Information Retrieval (ECIR 2018), LNCS, 2018.

[6]. Kartik Shenoy, Tejas Dastane, Varun Rao, Devendra Vyavaharkar, *"Real-time Indian Sign Language (ISL) Recognition,"* arXiv preprint, August 2021.

[7]. S. Kumar, *"Real-time Sign Language Detection: Empowering the Hearing-Impaired via AffordableVision-Based System,"* ScienceDirect / Elsevier, 2024.

[8]. K. Bhaskar, S. Anudeep Reddy, K. Yatheendra, G. Prathyusha, *"Cyberbullying Detection on Social Networks Comparing Machine Learning and Transfer Learning,"* International Journal of Information Technology and Computer Engineering, Vol. 12, No. 3, pp. 580–594, 2024.

[9]. *"Improving Automatic Cyberbullying Detection in Social Network Environments by Fine-Tuning a Pre-Trained Sentence Transformer Language Model,"* Social Network Analysis and Mining, 2024.

[10]. M. Abusaqer, C. Fofie Jr., *"Cyberbullying Classification Using Three Deep LearningModels: GPT, BERT, and RoBERTa,"* MICS 2023.

[11]. S. Mohamed Fati, A. Muneer, A. Alwadain, A. O. Balogun, *"Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of WordsFeature Extraction,"* Mathematics, 11(16), 3567, 2023.

[12]. M. Shapna Akter, H. Shahriar, A. Cuzzocrea, *"A Trustable LSTM-Autoencoder Network for Cyberbullying Detectionon Social Media Using Synthetic Data,"* Preprint, 2023.

[13]. T. Mahmud, M. Ptaszynski, J. Eronen, F. Masui, *"Cyberbullying Detection for Low-Resource Languages and Dialects: Review of the State of the Art,"* Preprint / Survey, 2023

[14]. S. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, S. Islam, *A Review on Deep-Learning-Based Cyberbullying Detection*, Future Internet, 15(5), 179, 2023.

[15]. N. Hudda, A. Mishra, S. Kumar, *"Detection of Cyberbullying on Social Media,"* International Journal of Computer Applications, 185(4), pp. 43–47, 2023.

[16]. Y. Al-Imrani, M. Douzal, et al., *"Cyberbullying Detection Framework for Short and Imbalanced Arabic Datasets,"* Journal of King Saud University – Computer and Information Sciences, 2023.

[17]. H. Chen, Y. Zhou, S. Zhu, H. Xu, *"Detecting Offensive Language in Social Media to Protect Adolescent Online Safety,"* Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust, pp. 71–80, 2012.

[18]. M. Dadvar, D. Trieschnigg, R. Ordelman, F. de Jong, *"Improving Cyberbullying Detection with User Context,"* ECIR 2013, LNCS, Vol. 7814, pp. 693–696, 2013.

[19]. V. Nahar, X. Li, C. Pang, *"An Effective Cyberbullying Detection Model for Twitter Social Network,"* 2014 IEEE 14th International Conference on Advanced Learning Technologies, pp. 182–186, 2014.

[20]. A. Schmidt, M. Wiegand, *"A Survey on Hate Speech Detection Using Natural Language Processing,"* Proceedings of the Fifth International Workshop on NLP for Social Media, pp. 1–10, 2017.