Ontology and AI Integration for Real-Time Detection of Cyberbullying Among University Students

Khaliq Ahmed[a], Ashley Mathew[b] and Shajina Anand[c]
Department of Math and Computer Science, Seton Hall University, New Jersey, U.S.A.

Keywords: Cybersecurity, BERT, NLP, Cyberbullying, Ontology, Graph – Based Ontology, Youth, Online Safety.

Abstract: With the increasing of the internet, smartphones, and social media, nearly everyone is a potential target for cyberbullying. Our research introduces an AI-driven approach to detect and address cyberbullying among college students, with a focus on its impact on mental health. We developed a context-specific ontology, drawing from real-time data, publicly available data, surveys, academic literature, and social media interactions to categorize information into domains such as victims, causes, types, environments, impacts, and responses. We collected real-time data from college students through surveys, interviews, and social media, leveraging advanced NLP (Natural Language Processing) techniques and BERT for accurate and efficient detection. By integrating this ontology with AI, our system dynamically adapts to emerging cyberbullying patterns, offering more precise detection and response strategies. Experimental results show that the proposed model achieves 96.2% accuracy, with 95.8% precision, 95.5% recall, and an F1-score of 95.6%. This performance surpasses traditional methods, emphasizing its capability to identify both explicit and implicit forms of abusive behavior. The approach not only introduces a tailored ontology for college students' unique social dynamics but also offers solutions to evolving cyberbullying trends. This research significantly enhances online safety and fosters a healthier digital environment for university students use.

1 INTRODUCTION

The widespread use of social media has significantly increased instances of cyberbullying, with serious implications for students' well-being and academic performance. To address this, researchers have developed automated methods to detect cyberbullying and create safer online environments. Deep learning techniques, particularly transformer models like BERT and DistilBERT, have shown significant promise, outperforming traditional methods by leveraging large datasets (Teng & Varathan, 2023). BERT-based models have also enhanced Aspect Target Sentiment Classification (ATSC), effectively identifying relationships between targets and sentiments in cyberbullying incidents (Chen et al., 2023). Advances in cross-platform detection methods, such as adversarial learning, have improved monitoring across multiple social media platforms, increasing detection flexibility (Yi & Zubiaga, 2022). While traditional approaches often focus on explicit language cues, they can miss subtle or covert abusive behaviors. BERT captures contextual nuances effectively but integrating it with graph-based ontologies further enhances detection by mapping relationships between abusive behaviors and associated concepts (Rogers et al., 2020). Our study combines BERT's contextual understanding with a hierarchical ontology framework, enabling precise detection of both explicit and implicit behaviors. Achieving 96.2% detection accuracy, our system surpasses traditional and state-of-the-art methods, offering a dynamic solution to improve online safety for college students.

2 LITERATURE REVIEW

Recent studies have focused on improving cyberbullying detection using advanced AI language

[a] https://orcid.org/0009-0000-7945-8476
[b] https://orcid.org/0009-0005-0876-0703
[c] https://orcid.org/0000-0001-6721-1150

Ahmed, K., Mathew, A. and Anand, S.
Ontology and AI Integration for Real-Time Detection of Cyberbullying Among University Students.
DOI: 10.5220/0013213500003929

In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 709-716
ISBN: 978-989-758-749-8; ISSN: 2184-4992

models. For example, Alrowais et al. (2024) developed an upgraded RoBERTa model, called RoBERTaNET, which uses GloVe word embeddings to detect cyberbullying tweets with 95% accuracy. While it shows high performance, it demands significant computing power, making widespread adoption challenging, especially in developing countries. Similarly, Ogunleye and Dharmaraj (2023) introduced a new dataset named D2 to enhance RoBERTa's detection capability. This approach provides better accuracy and resists skewed class problems but requires a large dataset, which limits its use in environments with limited data availability.

Teng and Varathan (2023) used transfer learning with DistilBERT to enhance detection, incorporating psycholinguistic factors, but achieved only 64.8% on the F-measure for logistic regression, indicating more work is needed for diverse social media content. The XP-CB model by Yi and Zubiaga (2022) uses adversarial learning to improve cross-platform detection, but it has high processing requirements, limiting its scalability. Sen et al. (2024) combined BERT with CNN and MLP, achieving 87.2% to 92.3% accuracy, outperforming other machine learning methods. However, its complexity makes real-time deployment challenging. Ejaz et al. (2024) developed a dataset that covers multiple aspects of cyberbullying, such as violence, repetition, and peer-to-peer interaction. This makes it more flexible for researchers but lacks detailed performance metrics. In a separate study, Chow et al. (2023) found that BERT achieved the highest accuracy (96%), slightly outperforming Bi-LSTM (95%) and Bi-GRU (94%) in detecting cyberbullying on tweets.

El Koshiry et al. (2024) used a CNN-BiLSTM model with Focal Loss and GloVe embeddings, achieving a 99% accuracy rate. However, the model struggled with recall, indicating a need for further improvements in capturing all instances of cyberbullying. Lastly, Kaur and Saini (2023) conducted a scient metric analysis of AI applications for cyberbullying detection, highlighting trends, contributions, and future research directions in this field, but without evaluating specific model performances. Ontology-based approaches (e.g., Gencoglu, 2020) provide structured domain knowledge, enabling better categorization and semantic understanding but often lack the ability to process implicit language patterns effectively. Transformer-based models, such as BERT and its variants (Chen et al., 2023; Yi & Zubiaga, 2022), excel in capturing linguistic nuances but struggle with representing complex relationships between concepts. Despite the strengths of these approaches, they are generally applied independently, leaving a gap in combining these techniques for tasks like cyberbullying detection.

As of now, there is no existing research that combines graph-based ontologies with BERT or similar transformer models for cyberbullying detection. Our work addresses this gap, providing an opportunity to develop a dual-layered approach that integrates the contextual understanding of BERT with the hierarchical structuring capabilities of ontologies to enhance both detection accuracy and adaptability to evolving abusive behaviors.

3 METHODOLOGY

3.1 Data Description

The dataset used in this study consists of messages labeled as either 'cyberbullying' or 'not cyberbullying.' Data was collected from three main sources: real-time input from university students through surveys, direct interactions, and virtual interviews; publicly available datasets from online platforms; and web-scraped data from public forums to capture diverse language patterns. The dataset contains two primary columns: 'message_text', which includes user-generated content from social media, and 'cyberbullying_type', indicating whether the content qualifies as cyberbullying. A majority of the messages are labeled as 'cyberbullying,' while a smaller portion is labeled as 'not cyberbullying'.

3.2 Data Collection

Participants for this study were voluntarily recruited from the university campus. After signing a consent form, they completed a survey on Qualtrics where they shared their experiences with cyberbullying, detailing its impact on their mental health and academic performance. Eligible participants were 18 years or older, currently enrolled in either bachelor's or master's programs, and had experienced or witnessed cyberbullying within the past year. Those interested in further participation engaged in 30-minute virtual interviews via MS Teams to provide deeper insights into their personal experiences. Additionally, we used existing datasets such as the Cyberbullying Detection Dataset on Twitter (2023), Instagram Cyberbullying Dataset (2022), and the OLID Dataset (2020). These datasets helped refine the ontology model and enhance the accuracy of AI-based detection algorithms.

3.3 Data Preprocessing

The initial stage of data preprocessing and critical component of Natural Language Processing (NLP) involved cleaning and preparing the raw text for analysis. Text preprocessing included tokenization, where sentences were split into individual words or sub words to facilitate analysis. For example, the text "You are so annoying!" was tokenized into ["You", "are", "so", "annoying", "!"]. All text was converted to lowercase to ensure consistency and reduce duplication, so that "annoying" and "Annoying" would be treated the same. Additionally, stop words like "and" "the," and "is", were removed as they did not add any significant meaning. Special characters, HTML tags, and other extraneous symbols were eliminated to ensure only relevant content remained for analysis. Figures 1 and 2 provide a comprehensive view of the dataset characteristics detailing distributions of cyberbullying types.

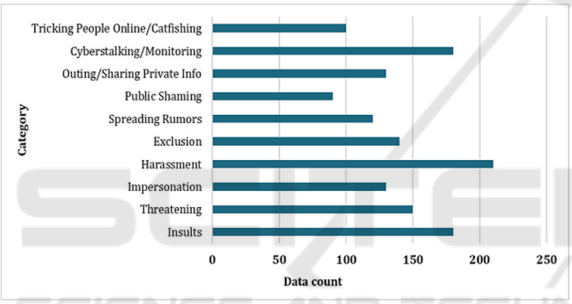

Figure 1: University data collection.

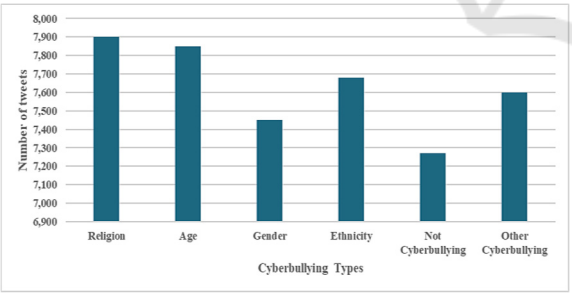

Figure 2: Cyberbullying Types.

3.4 Feature Extraction Using Bert

In the context of detecting cyberbullying, accurately interpreting the text often filled with slang, misspellings, or context-specific terms is crucial. BERT is used for feature extraction, given its ability to understand complex language patterns. The tokenization process is a fundamental step in preparing data for BERT, converting the input text into a format BERT uses Byte-Level Byte-Pair Encoding (BPE) for tokenization, breaking words into sub words to accommodate slang, rare words, and spelling errors commonly seen in user-generated content. that the model can process.

3.5 Graph Based Ontology

This section outlines the creation of a graph-based ontology designed to categorize and structure cyberbullying behaviors for better detection and analysis. By integrating domain knowledge with real-world examples, the ontology addresses limitations of existing models. Combining BERT's contextual understanding with ontology-based reasoning enhances the detection of explicit and implicit forms of cyberbullying. The ontology comprises two key components:

- T-Box (Terminological Box): Represents the schema, including categories (C) and their hierarchical relationships (R).
- A-Box (Assertional Box): Contains specific instances (I) derived from raw data (D), representing real-world examples of cyberbullying.

3.6 Building Ontology

Map each Feature *fj* directly to a Category:
For each identified feature related to cyberbullying behavior, we assign it to a specific category C.

Example: If f_{m+1} = "using derogatory terms", then:
f_{m+1} à " Insults"
If f_{m+2} = creating fake profiles", then:
f_{m+2} à " Impersonation"

Instance Mapping to Categories
Instances (I) identified from raw data are linked to appropriate categories:

If i_{k+1} = *"sending mean messages"*, then: $g(i_{k+1})$ = *"Harassment"*
If i_{k+2} = *"spreading false rumor"*, then: $g(i_{k+2})$ = *"spreading rumors"*

Constructing the Ontology Graph
The ontology graph (G) is built with vertices (V) representing categories and instances, and edges (E) representing relationships:

$$V = C \cup I$$

E={(Spreading Rumors,Gossiping about someone),(Impersonation, Creating fake profiles)}

Example: If “Spreading Rumors” includes “Gossiping about someone,” then

E = Spreading Rumors, Gossiping someone)

Define relationship types for Ontology
Define hierarchical relationships for the new categories:

Hierarchy: h(“Cyberbullying”) =
{“Insults”. “Using derogatory terms”}

Associations: r: I à C, e.g.,
“Creating fake profiles” à Impersonation.

Semantic Validation
Semantic validation ensures consistency between categories and relationships. Categories and instances are vectorized, and semantic similarity measures (SSS) are applied:

$$S(c_i, c_j) = \frac{Vec(c_i) * Vec(c_j)}{||Vec(c_i)|| \ ||Vec(c_j)||}$$

For example:

- S(“Spreading Rumors,” ”Gossiping about some one”): Measures closeness and adjusts categories if needed.

Inference and Reasoning over Ontology
Inference rules are implemented to enhance detection:

- If i_j involves "using derogatory terms," infer it as:
$$\Phi(i_j) = \text{"Insults"}$$
- If i_k involves "Creating fake profiles," infer it as "Impersonation."

The ontology was developed using Protege for designing and visualizing the hierarchical structure, while Owlready2 was employed to integrate the ontology into the detection framework seamlessly. Additionally, Python-based NLP tools were utilized for preprocessing the raw data (D), extracting relevant features (F), and identifying instances (I) of cyberbullying behaviors.
Final Ontology Structure: The final ontology is represented as:

$$O = (C, I, R, G, f, g)$$

where: **C:** Set of defined concepts, **I:** Set of identified instances, **R:** Set of relationships, **G:** Ontology graph structure, **f, g:** Functions mapping features and instances to categories.

The ontology expands detection capabilities by capturing complex relationships, improving accuracy, and adapting dynamically to evolving cyberbullying behaviors.

3.7 Ontology Reasoning and Development

Ontology in computer science categorizes cyberbullying behaviors based on social context (Lembo et al., 2013), aligning data with domain-specific categories for accurate detection. Our approach adapts the ontology to new data, categorizing emerging cyberbullying terms through semantic relations. A hierarchical structure (Figure 3) represents categories like insults, harassment, and catfishing.
Arrows in Figure 3 denote two types of relationships:

- **Hierarchical Relationships:** Solid arrows connect broader categories to subcategories, such as linking "Cyberbullying" to "Insults" or "Impersonation."
- **Associative Relationships:** Dashed arrows connect co-occurring or conceptually related terms, such as "Idiot" and "Ugly," reflecting linguistic patterns from real-world data.

Instances like "sharing false information" under Spreading Rumors and "creating fake profiles" under Impersonation capture explicit and implicit behaviors. BERT maps feature predefined ontology categories, ensuring precise classification of abusive content while capturing nuanced, context-dependent meanings. The ontology dynamically refines categories based on new data, making it a powerful tool for detecting online abuse and supporting students' mental health.

3.7.1 NLP with BERT

The core technology behind our project is Natural Language Processing powered by BERT algorithm. We use both to detect and understand cyberbullying among college students. BERT is particularly well-suited for this task due to its exceptional ability to read and comprehend language contextually, much like a human would. It also can process words in relation to both preceding and succeeding words in a sentence. This enables BERT to capture hidden meanings, which is a crucial feature for detecting cyberbullying, where harmful intent may be implicit or context dependent (Lee et al., 2019).

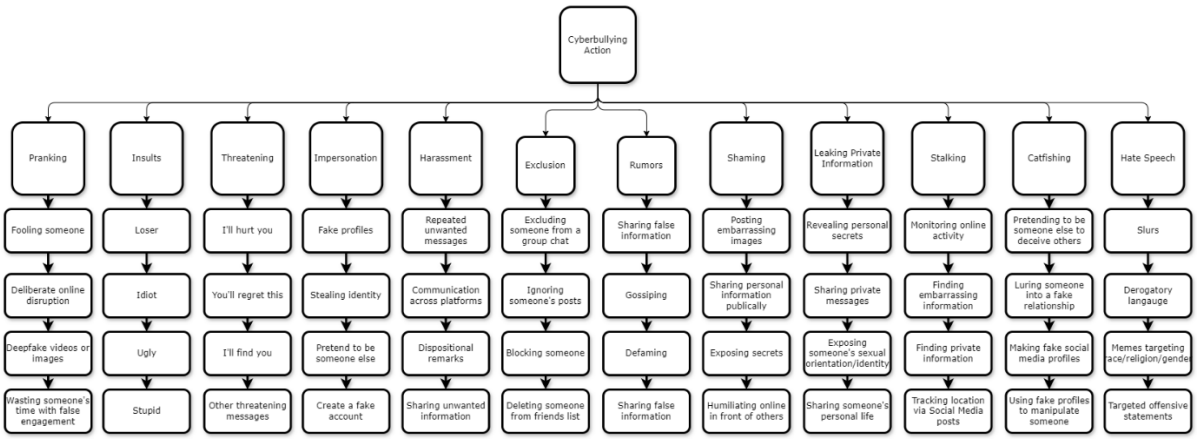

Figure 3: Ontology Graph.

3.8 BERT's Self-Attention Model

BERT processes text using Transformer architecture, which enables it to understand the relationships between cyberbullying words. The mathematical core of BERT lies in the **self- attention mechanism**, which computes relationships between words using this formula:

$$Attention\ (Q, K, V) = softmax(\frac{QK_K}{\sqrt{dk}})\ V \text{ where,}$$

- **Q** (query), **K** (key), and **V** (value) matrices are taken from the input words or sentences.
- The **SoftMax** function is used to confirm the values add up to 1. Allows BERT to focus on the most relevant parts of the sentence specific to the cyberbullying issue.
- $(\frac{QK^K}{\sqrt{d_k}})$ Makes sure the attention is prioritized, and each cyberbullying word is compared with every other word to decide how much focus to give each one (Rogers et al., 2020).

3.9 Integration of NLP with Graph-Based Ontology

The proposed cyberbullying detection model combines BERT's NLP capabilities with a graph-based ontology for semantic comprehension: The algorithm starts by initializing the model with a pre-trained BERT and loading a structured ontology tree. Text preprocessing involves tokenization and embedding generation. These embeddings are then matched with ontology-defined concepts to identify relevant instances of abusive behavior. The system is designed for continuous learning, logging new instances and retraining itself periodically to remain effective against evolving language dynamics.

1. Model Initialization

BERT←Load(BERT_pretrained,TrainingDataset)
If OntologyFile exists then:
Ontology ← Load(OntologyFile)
OntologyTree ← ConstructOntologyTree(Ontology)

2. Text Preprocessing

If RelevantContent is not empty then:
Tokens ← Tokenize(RelevantContent)
BERT_Embeddings ← GenerateEmbeddings(Tokens, BERT)

3. Contextual Analysis

If BERT_Embeddings are valid then:
MatchedConcepts ← MatchConcepts(OntologyTree, BERT_Embeddings)
If MatchedConcepts are found then:
ContentAnalysis ← Integrate(MatchedConcepts, ContentAnalysis)

4. Continuous Learning and Updates

LogInteraction(CyberbullyingType, ContextualUnderstanding)
If LoggedData is sufficient for retraining then:
BERT ← RetrainModel(BERT, LoggedData)
OntologyTree ← UpdateOntologyTree(LoggedData)
End procedure

Algorithm 1: Cyberbullying detection model.

3.10 Model Architecture and Methodology Framework

The proposed cyberbullying detection framework consists of four components: Data Collection, Data Preprocessing, Ontology Integration, and AI/ML

Integration. Data is sourced from public datasets, surveys, and interviews, followed by cleaning, normalization (e.g., lowercasing and punctuation removal), and tokenization to ensure compatibility with BERT. BERT generates contextual embeddings that capture complex language nuances, which are then aligned with a graph-based ontology. This ontology defines abusive behaviors and their relationships, mapping concepts like 'harassment' and 'impersonation' to improve detection accuracy for both explicit and subtle forms of cyberbullying.

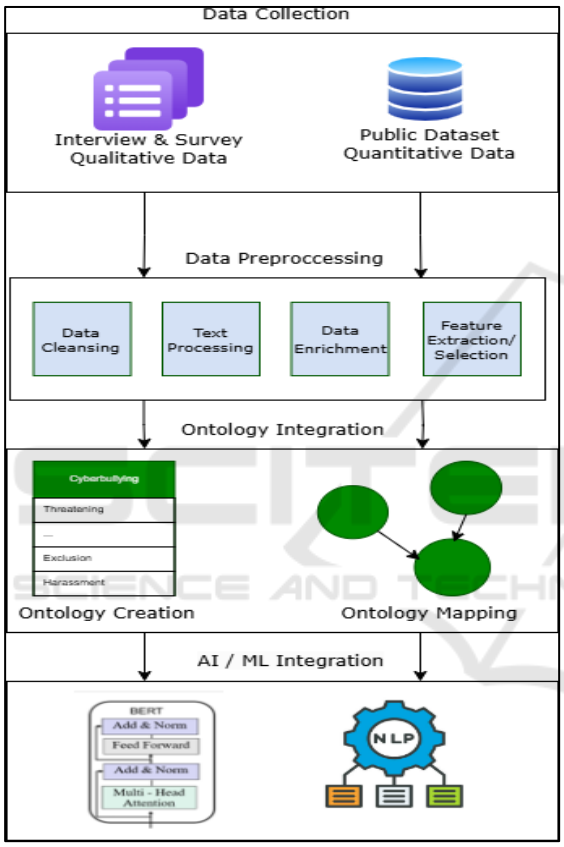

Figure 4: Overall Methodology framework.

The framework undergoes training with cross-validation and hyperparameter tuning to enhance performance. A continuous learning mechanism logs interactions, periodically retrains the BERT model, and updates the ontology to adapt to evolving trends in abusive behavior. By integrating NLP, BERT, and ontology-based reasoning, the system effectively identifies and classifies various forms of online abuse, contributing to safer digital environments through real-time detection and mitigation.

4 EXPERIMENTAL SETUP

The evaluation of cyberbullying detection models enhanced with a graph-based ontology shows notable performance variations between Logistic Regression (LR) and Random Forest (RF) across dimensions like overall detection, ethnicity, gender, age, and religion-related content. The experiment used 10,000 labelled data points collected from public datasets, surveys, and interviews with college students. The data was split into training (70%), validation (15%), and testing (15%) sets. Preprocessing included text normalization, tokenization, and enrichment to ensure BERT compatibility for analysing complex language patterns. The ontology was created and managed using Python's owlready2 library to integrate semantic reasoning into the framework.

4.1 Performance Comparison Models

Logistic Regression (LR) outperformed Random Forest (RF) overall with an AUC of 0.62 versus 0.46, as LR better captured cyberbullying characteristics. However, RF excelled in detecting ethnicity- and gender-related content (AUCs of 0.33 and 0.45, respectively) due to its ability to model non-linear patterns. LR performed slightly better for age-related content (AUC 0.38), while both models were highly effective for religion-related content (LR: 0.99, RF: 0.98).

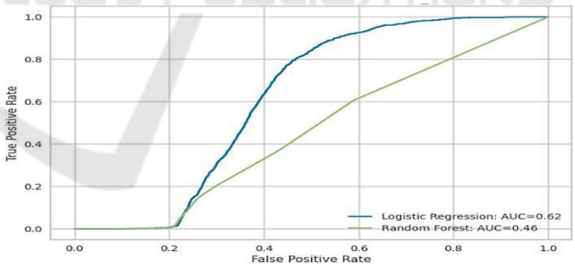

Figure 5: Comparison for class other_cyberbullying.

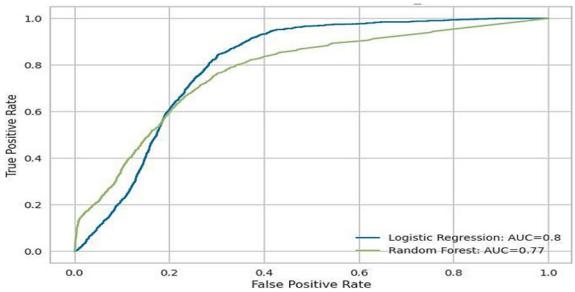

Figure 6: Comparison for class not_cyberbullying.

A BERT-based model integrated with NLP and ontology achieved the best results, with 96.2% accuracy, 95.8% precision, 95.5% recall, and an F1 score of 95.6%, highlighting its strength in detecting both explicit and implicit abusive behaviors. RoBERTa-based models (RoBERTa + CNN and RoBERTa + GRU) also performed well, achieving 95.2% and 94.8% accuracy, respectively. The graph-based ontology further improved detection accuracy by understanding relationships between concepts, making BERT with Ontology a superior solution for identifying diverse abusive behaviors.

Figure 7: Graph Based Ontology - words classification.

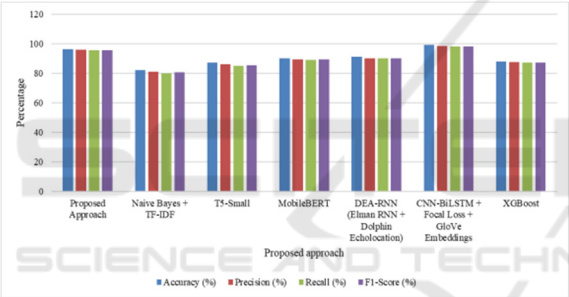

Figure 8: New Proposed Approach.

5 CONCLUSIONS

This study introduces an effective approach for detecting cyberbullying by integrating Natural Language Processing (NLP) with the BERT model and a Graph-Based Ontology framework. The system achieved an accuracy of 96.2%, with precision, recall, and an F1-score all exceeding 95%. These results represent a significant improvement over traditional methods, demonstrating the model's ability to accurately identify both clear and subtle forms of cyberbullying. By using BERT's advanced language understanding for feature extraction along with the structured insights provided by graph-based ontologies, this approach effectively handles complex language patterns often found in abusive behavior online. The strong performance of the combined BERT + Ontology model shows its capability to detect nuanced instances of cyberbullying that other models may overlook. Beyond improving detection accuracy, this method has practical applications in areas like social media monitoring and online safety programs. It offers a comprehensive solution that adapts to real- time language changes, making online spaces safer.

Future work could further enhance this approach by integrating additional language models and expanding the ontology to cover emerging trends in digital interactions. This ongoing development would strengthen the system's ability to detect evolving forms of cyberbullying, ultimately contributing to more effective online safety measures.

REFERENCES

RRoBERTa Transformer Based Model for Cyberbullying Detection with GloVe Features. *IEEE Access*.

Ogunleye, B., & Dharmaraj, B. (2023). The use of a large language model for cyberbullying detection. *Analytics*, 2(3), 694-707.

Pericherla, S., & Ilavarasan, E. (2021). Performance analysis of word embeddings for cyberbullying detection. *IOP Conference Series: Materials Science and Engineering*, 1085(1), IOP Publishing.

Emon, M. I. H., et al. (2022). Detection of Bangla hate comments and cyberbullying in social media using NLP and transformer models. In *International Conference on Advances in Computing and Data Sciences* (pp. 1-12). Cham: Springer International Publishing.

Pericherla, S., & Ilavarasan, E. (2024). Transformer network-based word embeddings approach for autonomous cyberbullying detection. *International Journal of Intelligent Unmanned Systems*, 12(1), 154-166.

Yani, M. A., & Maharani, W. (2023). Analyzing Cyberbullying Negative Content on Twitter Social Media with the RoBERTa Method. *JINAV: Journal of Information and Visualization*, 4(1), 61-69.

Teng, T. H., & Varathan, K. D. (2023). Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access*, 11, 55533-55560.

Chen, C. K., Ramjee, S., & Wang, J. Aspect-Target Sentiment Classification for Cyberbullying Detection.

Yi, P., & Zubiaga, A. (2022). Cyberbullying detection across social media platforms via platform-aware adversarial encoding. *Proceedings of the International AAAI Conference on Web and Social Media*, 16, 123-133.

Pericherla, S., & Egambaram, I. (2021). Cyberbullying detection on multi-modal data using pre-trained deep learning architectures. *Ingeniería Solidaria*, 17(3), 1-20.

Gencoglu, O. (2020). Cyberbullying detection with fairness constraints. *IEEE Internet Computing*, 25(1), 20-29.

Hasan, M. T., et al. (2023). A review on deep-learning-based cyberbullying detection. *Future Internet*, 15(5), 179.

Pericherla, S., & Egambaram, P. (2021). Detecting cyberbullying across multiple platforms using multi-modal deep learning techniques. *Journal of Computer Vision and Multimedia Studies*, 29(4), 102-120.

Gencoglu, O. (2020). Fairness-aware cyberbullying detection using deep learning. In *Proceedings of the IEEE/ACM International Conference on AI Ethics and Fairness* (pp. 25-34). IEEE. https://doi.org/10.1109/MIC.2020.3032461

Akhter, A., et al. (2023). A robust hybrid machine learning model for Bengali cyberbullying detection in social media. *Natural Language Processing Journal*, 4, 100027.

Islam, M. S., & Rafiq, R. I. (2023). Comparative analysis of GPT models for detecting cyberbullying in social media platforms threads. In *Annual International Conference on Information Management and Big Data* (pp. 245-257). Cham: Springer Nature Switzerland.

Ottosson, D. (2023). Cyberbullying detection on social platforms using large language models. *Journal of AI Applications in Social Media*, 35(1), 59-71.

Akinyemi, J. D., et al. (2023). Cyberbullying detection and classification in social media texts using machine learning techniques. In *International Conference on Computer Science, Engineering and Education Applications* (pp. 102-115). Springer Nature Switzerland.

Sen, M., Masih, J., & Rajasekaran, R. (2024). From tweets to insights: BERT-enhanced models for cyberbullying detection. *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*. IEEE.

Ferrer, R., Ali, K., & Hughes, C. (2024). Using AI-based virtual companions to assist adolescents with autism in recognizing and addressing cyberbullying. *Sensors*, 24(12), 3875.

Ejaz, N., Razi, F., & Choudhury, S. (2024). Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm. *Computers in Human Behavior*, 153, 108123.

Murshed, B. A. H., et al. (2022). DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. *IEEE Access*, 10, 25857-25871.

Chow, D. V., et al. (2023). Cyberbullying detection: An investigation into natural language processing and machine learning techniques. *2023 5th International Conference on Cybernetics and Intelligent System (ICORIS)*. IEEE.

Jadhav, R., et al. (2023). Cyber bullying and toxicity detection using machine learning. *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*. IEEE.

El Koshiry, A. M., et al. (2024). Detecting cyberbullying using deep learning techniques: A pre-trained GloVe and focal loss technique. *PeerJ Computer Science*, 10, e1961.

Kaur, M., & Saini, M. (2023). Role of artificial intelligence in cyberbullying and cyberhate detection. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining.

Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842–866. https://doi.org/10.1162/tacl_a_00349

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). brat: A web-based tool for NLP-assisted text annotation. *Conference of the European Chapter of the Association for Computational Linguistics*, 102–107. http://pontus.stenetorp.se/res/pdf/stenetorp2012brat.pdf