BILINGUAL PUBLISHING GROUP

Journal of Electronic & Information Systems
https://journals.bilpubgroup.com/index.php/jeis

ARTICLE

Attribute-specific Cyberbullying Detection Using Artificial Intelligence

***Adeyinka Orelaja*[1*], *Chidubem Ejiofor*[2], *Samuel Sarpong*[3], *Success Imakuh*[4], *Christian Bassey*[5], *Iheanyichukwu Opara*[6], *Josiah Nii Armah Tettey*[7], *Omolola Akinola*[8]**

[1] *Department of Computer Science, Austin Peay State University, Tennessee, 37044, USA*
[2] *Department of Computer Science, Western Illinois University, Macomb, Illinois, 61455, USA*
[3] *Department of Computing, East Tennessee State University, Tennessee, 37604, USA*
[4] *Department of Computing, Teesside University, Middlesbrough, TS1 3BX, United Kingdom*
[5] *Department of Computer Science, Innopolis University, Innopolis, 420500, Russia*
[6] *Department of Oil and Gas Engineering, Robert Gordon University, Aberdeen, AB10 7AQ, United Kingdom*
[7] *Department of Computer Science, Wright State University, Dayton, Ohio, 45435, USA*
[8] *Department of Information System and Analysis, Lamar University, Beaumont, Texas, 77705, USA*

ABSTRACT

Cyberbullying, a pervasive issue in the digital age, poses threats to individuals' well-being across various attributes such as religion, age, ethnicity, and gender. This research employs artificial intelligence to detect cyberbullying instances in Twitter data, utilizing both traditional and deep learning models. The study repurposes the Sentiment140 dataset, originally intended for sentiment analysis, for the nuanced task of cyberbullying detection. Ethical considerations guide the dataset transformation process, ensuring responsible AI development. The Naive Bayes algorithm demonstrates commendable precision, recall, and accuracy, showcasing its efficacy. The Bi-LSTM model, leveraging deep learning capabilities, exhibits nuanced cyberbullying detection. The study also underscores limitations, emphasizing the need for refined models and diverse datasets.

Keywords: Cyberbullying detection; Social media analysis; Artificial intelligence; Naive Bayes; Bi-LSTM; Ethical AI; Machine learning; Digital well-being

*CORRESPONDING AUTHOR:
Adeyinka Orelaja, Department of Computer Science, Austin Peay State University, Tennessee, 37044, USA; Email: aorelaja@my.apsu.edu

ARTICLE INFO
Received: 10 January 2024 | Revised: 31 January 2024 | Accepted: 20 February 2024 | Published Online: 28 February 2024
DOI: https://doi.org/10.30564/jeis.v6i1.6206

CITATION
Orelaja, A., Ejiofor, C., Sarpong, S., et al., 2024. Attribute-specific Cyberbullying Detection Using Artificial Intelligence. Journal of Electronic & Information Systems. 6(1): 10–21. DOI: https://doi.org/10.30564/jeis.v6i1.6206

1. Introduction

1.1 Background

Cyberbullying is the act of harming or harassing someone online through messages or images that are malicious or harmful. Cyberbullying can negatively affect the mental health and well-being of the victims, causing depression, anxiety, low self-esteem, and suicidal thoughts [1].

Another study by Erbic¸er et al. (2023) [2] defines cyberbullying perpetration as a form of harmful behavior, which can be defined as the deliberate, repetitive, and damaging attitude of individuals or groups harming others using the internet, mobile phones, or other communication tools such as e-mail, messages, or social media.

There are many definitions of cyberbullying, however, many definitions are deemed insufficient, often lacking clarity and consistency. The challenges involved typically included variations in the defined electronic methods [3]. Different scholars and organizations may conceptualize cyberbullying in various ways, ranging from explicit threats to subtle forms of harassment. Some definitions encompass the misuse of power differentials, while others focus on the intent to cause harm. By recognizing this diversity, our research seeks to address the multifaceted nature of cyberbullying, encompassing a broad spectrum of aggressive behaviors within the context of online communication.

Social media platforms, such as X (formerly known as Twitter), allow millions of people to share their opinions, thoughts, and feelings online. However, they also enable cyberbullies to target and harass others based on their personal characteristics, such as religion, age, ethnicity, and gender [4]. Therefore, it is crucial to develop effective methods to detect and prevent cyberbullying on social media platforms and to protect the online safety and dignity of the users. A global survey by Microsoft [5] found that 75% of participants agreed that social media companies needed to moderate harmful speech online. Being able to detect cyberbullying on these social media platforms is the first step in achieving this.

Artificial intelligence (AI) is a field of computer science that aims to create machines or systems that can perform tasks that require human intelligence, such as reasoning, learning, and decision-making [6]. AI can be used to analyze and understand large amounts of data, such as text, images, and videos, and to extract useful information or insights from them [7].

AI can be applied to detect cyberbullying on social media platforms by using techniques such as natural language processing (NLP) and machine learning (ML). NLP is a subfield of AI that deals with the interaction between computers and human languages, such as understanding, generating, and translating natural language texts [8] and ML is a subfield of AI that focuses on creating systems that can learn from data and improve their performance without explicit programming [9].

1.2 Motivation

The motivation behind undertaking this research is underscored by the increasing severity and diversity of cyberbullying incidents. In recent years, cyberbullying has transcended traditional forms, branching into attribute-based attacks that target individuals based on characteristics such as religion, age, ethnicity, and gender. The consequences of such attacks are profound, affecting not only the mental well-being of individuals but also perpetuating societal divisions.

The ever-evolving nature of online communication poses a unique challenge. Traditional approaches to cyberbullying detection often struggle to keep pace with the dynamic patterns and expressions of harassment on platforms like Twitter (X). The motivation is thus driven by the need for adaptive, sophisticated algorithms capable of discerning nuanced forms of cyberbullying, particularly those tied to specific attributes. In an era where cyberbullying is a growing concern in the digital era, with notable implications for individuals' mental health [10], the motivation for this research extends beyond academic curiosity to a commitment to foster digital spaces that are free from the detrimental

effects of cyberbullying.

1.3 Problem statement

Cyberbullying encompasses a range of harmful behaviors manifesting in digital spaces, impacting individuals based on attributes such as religion, age, ethnicity, and gender. The lack of effective mechanisms to identify and curb cyberbullying on platforms like X Twitter perpetuates an environment where users may experience online harassment, leading to psychological distress and potential real-world consequences. Existing studies [11] emphasize the need for advanced techniques to automatically detect and mitigate cyberbullying instances, tailoring approaches to the nuanced nature of social media interactions.

1.4 Objectives

This research is anchored in the goal of crafting a machine learning-backed cyberbullying detection tool that utilizes the vast sea of social media data. The following detailed objectives are set to support this ambition:

1) **Data Acquisition and Preprocessing**: To achieve robust data-driven insights, the study aims to collect a diverse dataset from social media posts, particularly Twitter (X). The gathered data will then undergo meticulous preprocessing, encompassing noise reduction, text normalization, and the resolution of challenges inherent to cyberbullying content.

2) **Development of Attribute-Specific Detection Models**: Moving to the development phase, the research will conduct a thorough literature review to inform the creation of advanced machine learning models. Utilizing both Naive Bayes and Long Short-Term Memory (LSTM) algorithms, special attention will be given to attribute-specific detection. This entails tailoring models to recognize cyberbullying instances related to 'religion', 'age', 'ethnicity', and 'gender' categories. The iterative fine-tuning and optimization of these models will be paramount to ensuring their effectiveness.

3) **Comprehensive Evaluation of Model Performance**: Subsequently, the research will shift focus to the comprehensive evaluation of model performance. This involves the selection and justification of appropriate evaluation metrics, considering factors such as accuracy, precision, recall, and F1 score. Real-world testing will be conducted using representative social media data from Twitter (X), and a comparative analysis will benchmark the developed models against each other and existing state-of-the-art cyberbullying detection models.

Ethical considerations will be woven into each stage of the research, with a specific emphasis on addressing bias and ensuring fairness. The objective is to propose strategies that mitigate ethical concerns and enhance the responsible deployment of the developed models.

2. Literature review

The exploration of cyberbullying within the context of social media platforms, notably Twitter (X), has been a subject of significant scholarly inquiry. The extensive body of work in this research area underscores the gravity of the issue and the imperative to comprehend the various facets of online harassment.

2.1 Categorization of aggressive messages

The categorization of aggressive messages is a crucial aspect of understanding and addressing online harassment and cyberbullying. Previous research has delved into various approaches for categorizing aggressive content on social media platforms. Elsafoury et al. (2020) conducted a comprehensive analysis of cyberbullying datasets, including those from Twitter, Kaggle, Wikipedia Talk pages, and YouTube [12]. Their work involved the extraction and classification of over 47,000 tweets, providing insights into different forms of cyberbullying, including age-based, religion-based, ethnicity-based, and gender-based aggression.

The study emphasized the importance of diverse demographic parameters in categorizing

cyberbullying instances. This aligns with the present research, which leverages demographic attributes such as age, ethnicity, gender, and religion in the detection and categorization of cyberbullying tweets. The distribution of instances across classes reflects the varied nature of cyberbullying content within each category.

2.2 Existing approaches to cyberbullying detection

Cyberbullying detection has been addressed through a spectrum of approaches, ranging from traditional rule-based systems to advanced machine learning and deep learning models. Prior studies [10,13] have extensively explored various methodologies for cyberbullying detection. Each methodology brings distinct advantages and challenges to the forefront.

1) Rule-Based Systems: Rule-based systems leverage predefined patterns or heuristics to identify potential instances of cyberbullying. An example is the use of keyword matching where predefined sets of offensive words or phrases trigger an alert. These systems are straightforward to implement and interpret but often struggle with the dynamic and nuanced nature of cyberbullying, as they may not capture context well [14].

2) Machine Learning Models: Machine learning models, including classic algorithms like Support Vector Machines (SVMs) and Random Forests, have demonstrated efficacy in learning patterns from labeled data. SVMs, for instance, have been employed to classify cyberbullying instances based on features extracted from textual data [15]. Random Forests, with their ensemble learning approach, offer robustness against overfitting and have been applied to cyberbullying detection tasks [16]. These models exhibit adaptability to evolving cyberbullying patterns but may require significant labeled data for effective training.

3) Deep Learning Models: Deep learning models, characterized by architectures like recurrent neural networks (RNNs) and convolutional neural networks (CNNs), excel at capturing complex relationships in textual data. For instance, Almomani et al. (2024) [17] proposed a method using a CNN to detect cyberbullying incidents on Instagram, demonstrating the capacity of deep learning models to discern intricate patterns in multimedia-rich content. Long Short-Term Memory (LSTM) networks, a type of RNN, have been employed for sequential modeling, enabling the understanding of temporal dynamics in cyberbullying conversations [18]. Deep learning models showcase a high degree of sophistication in understanding context and semantics, making them well-suited for cyberbullying detection tasks.

The landscape of cyberbullying detection is dynamic, and the effectiveness of each approach depends on the context, the nature of the data, and the specific nuances of cyberbullying instances.

2.3 Ethical considerations in cyberbullying detection

The ethical dimensions surrounding the deployment of cyberbullying detection machine learning models have become increasingly prominent in recent works. Scholars have scrutinized the impact of these models on various ethical aspects, including issues of bias, fairness, and privacy.

Existing works have explored different avenues to address these ethical concerns. Aizenberg and Van Den Hoven (2020) [19] shed light on the broader landscape of big data ethics, emphasizing the need for responsible practices in handling sensitive information. This broader perspective encompasses considerations beyond individual instances of cyberbullying and extends to the overarching ethical responsibilities tied to data usage in the digital sphere.

Sap et al. (2019) [20] delved specifically into the risk of racial bias in hate speech detection, recognizing the profound implications of biases within detection models. Their work highlighted the challenges of ensuring fairness in models that are designed to discern harmful online content. Such biases could potentially exacerbate existing inequalities and societal divisions, necessitating a careful examination of the underlying algorithms.

In the pursuit of advancing cyberbullying

detection models ethically, it is imperative to acknowledge and bridge these gaps. The current research aims to contribute to this ongoing discourse by proposing strategies that not only address biases but also ensure the responsible and fair deployment of cyberbullying detection models within the intricate landscape of social media.

2.4 Gaps in existing literature

Significant strides have been made in understanding and addressing cyberbullying, but notable gaps persist in the current literature, creating avenues for further exploration and refinement.

While existing works touch on biases and fairness in detection models, ethical considerations in the preprocessing of cyberbullying data remain underexplored. This research will fill this gap by scrutinizing the ethical implications of data preprocessing, ensuring the foundation of detection models aligns with ethical standards.

Current literature also often adopts a binary approach, distinguishing between cyberbullying and noncyberbullying instances. However, nuances across attributes like 'religion', 'age', 'ethnicity', and 'gender' call for a more nuanced approach. This research addresses this gap by employing attribute-specific detection, offering a more context-aware understanding of cyberbullying.

This research aims to enhance the existing discourse by addressing these gaps, contributing to a more ethically sound and context-aware cyberbullying detection paradigm.

3. Methodology

3.1 Dataset overview

The dataset utilized in this study originates from a comprehensive collection of cyberbullying data compiled by Elsafoury (2020) [12]. Primarily sourced from various social media platforms, including Kaggle, Twitter (X), Wikipedia Talk pages, and YouTube, the dataset offers a diverse range of cyberbullying instances. For the purposes of this research, the focus was narrowed down to extracting Twitter data (X), resulting in a dataset exceeding 47,000 tweets explicitly labeled as cyberbullying.

Composition and Demographic Parameters: The dataset exhibits a rich composition, including explicit labels for different forms of cyberbullying and demographic parameters such as age, ethnicity, gender, and religion. The classes/labels are moderately balanced as shown in **Table 1**. This multiclass dataset enables a nuanced understanding of cyberbullying phenomena, capturing the intersectionality of various demographic factors with different manifestations of cyberbullying.

Table 1. Distribution of instances across cyberbullying classes.

Class	Count
Religion	7997
Age	7992
Ethnicity	7959
Gender	7948
Not cyberbullying	7937
Other cyberbullying	7823

3.2 Preprocessing

The preprocessing phase is crucial to ensure the integrity and quality of the dataset for cyberbullying detection. This section outlines a series of steps encompassing data loading, duplicate removal, and an intricate set of text-cleaning processes as adhered to in the recommendations outlined in the paper by Bokolo and Liu (2023) [10]. Particular emphasis is placed on avoiding biases and ensuring fairness throughout the cleaning procedures.

1) *Data Loading and Initial Inspection:* The initial step involves loading the raw dataset from a CSV file (cyberbullying_tweets.csv). The dataset is then inspected to comprehend its structure and content, showcasing the first few rows and providing key information such as column names and data types.

2) *Duplicate Removal:* Duplicate tweets can introduce biases and skew the model's performance. This subsection details the identification and removal of duplicate tweets, ensuring the dataset's uniqueness

and integrity.

3) *Column Renaming:* To streamline subsequent references, columns are renamed, adopting more concise names. Specifically, 'tweet_text' is renamed to 'text', and 'cyberbullying_type' to 'sentiment'.

4) *Text Cleaning Functions:* A set of custom-defined functions is introduced for comprehensive text cleaning, with a keen eye on avoiding biases and ensuring fairness:

Emoji Removal: Eliminating emojis to prevent potential bias associated with certain emoticons with the function shown.

Decontraction: Ensuring uniformity in language by expanding contractions, avoiding bias introduced by different writing styles as shown in the 'decontract()' function.

Entity Stripping: Removing links, mentions, and special characters to prevent biased influence from specific entities or symbols.

Hashtag Cleaning: Ensuring fair treatment of hashtags, cleaning them at the end of sentences and removing the '#' symbol within words to prevent unintended biases as shown in the function.

Filtering Specific Characters: Removing characters such as '$' and '&' to avoid biases associated with particular symbols.

Removing Multiple Sequential Spaces: Avoid bias by maintaining consistent spacing throughout the text.

Stemming: Standardizing words to their root form to ensure fairness in words.

5) *Application of Cleaning Functions:* The defined cleaning functions are systematically applied to each tweet in the dataset, resulting in a new column, 'text_clean', containing the cleaned text.

6) *Final Dataset Inspection:* Following the cleaning procedures, a final inspection of the dataset is conducted, revealing changes in size and highlighting any potential improvements in data quality.

7) *Sentiment Labeling:* The sentiment labels are redefined for clarity, mapping 'religion' to 0, 'age' to 1, 'ethnicity' to 2, 'gender' to 3, and 'not_cyberbullying' to 4.

8) *Text Length Analysis:* The distribution of text lengths is analyzed, with visualizations depicting the count of tweets based on their word length. Tweets with lengths exceeding certain thresholds are filtered to ensure data quality and relevance.

3.3 AI techniques

A meticulous split into training and test sets, allocating 20% to the latter and further dividing the remaining 80% into training and validation data, was conducted to monitor model accuracy and mitigate overfitting.

In the subsequent phase of model building, two distinct models—Naive Bayes and Bidirectional Long Short-Term Memory (Bi-LSTM)—were selected and compared for their efficacy in cyberbullying detection. This choice was guided by recommendations from related literature, acknowledging the inherent advantages of these models within the realm of sentiment analysis. The ensuing comparative analysis aims to determine the most effective model for accurately classifying tweets and detecting instances of cyberbullying within the dataset.

Naive Bayes: The Naive Bayes algorithm stands out as a swift and straightforward classification method, particularly adept at handling extensive datasets [21]. Proven effective in various applications, including spam filtering, text classification, public opinion analysis, and recommendation systems, Naive Bayes leverages the Bayes theorem of probability for predicting unknown classes.

In the implementation of the Naive Bayes Model, the model was instantiated using a Count Vectorizer to create a bag of words. Subsequently, TF-IDF (Term Frequency-Inverse Document Frequency) transformation was applied to assign weights to words based on their frequency, enhancing the model's understanding of their contextual significance. **Pytorch-Bi-LSTM Sentimental Analysis**: The Bi-LSTM (Bidirectional Long Short-Term Memory) model plays a pivotal role in the cyberbullying detection framework [10]. Below is a detailed description of the Bi-LSTM model

construction and training.

1) *Model Architecture:* The LSTM model is designed as a subclass of the PyTorch nn.Module class, named BiLSTM_Sentiment_Classifier. It comprises the following key components also shown in **Figure 1**:

Embedding Layer: Converts input tokens into dense vectors of fixed size (embedding).

```
BiLSTM_Sentiment_Classifier(
  (embedding): Embedding(33009, 200)
  (lstm): LSTM(200, 100, batch_first=True, dropout=0.5, bidirectional=True)
  (fc): Linear(in_features=200, out_features=5, bias=True)
  (softmax): LogSoftmax(dim=1)
)
```

Figure 1. Bi-LSTM architecture.

Bidirectional LSTM Layers: The LSTM layers process the embedded tokens, capturing contextual information. The number of layers (lstm), hidden dimension (hidden_dim), and bidirectional nature are configurable.

Fully Connected Layer: A linear layer (fc) transforms the output of the LSTM layers into logits for each sentiment class.

Softmax Activation: The LogSoftmax activation function normalizes the logits to probabilities, facilitating class predictions [22].

Initialization of Hidden States: The init_hidden method initializes the LSTM hidden and cell states.

The model is equipped to handle a dynamic batch size (batch_size), allowing flexibility during training and evaluation.

2) *Model Training:* The model is trained using the 'AdamW' optimizer with a learning rate of 3e-4 and a weight decay of 5e-6. The negative log-likelihood loss (NLLLoss) serves as the criterion for training. The training process is conducted over multiple epochs (5 EPOCHS), with early stopping implemented to prevent overfitting. Training involves iterating through the training dataset, computing gradients, and updating model parameters.

4. Results and discussion

In this chapter, the intricacies of the results obtained during the experimentation phase are delved into. The primary objective is to provide a transparent and detailed overview of the cyberbullying detection framework's effectiveness across various dimensions. This includes an exploration of an in-depth analysis of model performance metrics, and a comparative assessment of diverse machine learning and deep learning models. Furthermore, the chapter addresses the ethical considerations embedded in the models, reflecting on potential biases and fairness aspects.

4.1 Model performance

In evaluating the models for cyberbullying detection, we meticulously examine the performance metrics of two distinct approaches: the Naive Bayes classifier and the Bi-LSTM neural network. These models represent different paradigms, with Naive Bayes relying on probabilistic principles and Bi-LSTM leveraging the power of recurrent neural networks. Performance evaluation metrics were applied following the methods detailed in the research by Bokolo et al. (2023) [23].

1) *Naive Bayes Performance:* The Naive Bayes classifier exhibits commendable performance with precision, recall, and F1 score all hovering around 0.85. This suggests that the model effectively identifies instances of cyberbullying while minimizing false positives. The accuracy of 0.85 underscores its overall correctness in predictions. The Naive Bayes model demonstrated commendable performance across various classes, as illustrated in **Table 2**. Notably, the model excelled in precision for the 'Religion' and 'Age' classes, achieving 85% and 80%, respectively. However, the model showed some challenges in recall for the 'Not Bullying' class, achieving 47%.

Table 2. Class-wise performance of the Naive Bayes model.

	Precision	Recall	F1-score	Support
Religion	0.85	0.97	0.91	1579
Age	0.80	0.98	0.88	1566
Ethnicity	0.90	0.92	0.91	1542
Gender	0.89	0.85	0.87	1462
Not bullying	0.84	0.47	0.60	1274

2) *Bi-LSTM Performance:* Contrastingly, the Bi-LSTM neural network demonstrates superior

performance with precision, recall, and F1 score all surpassing 0.93. This signifies the model's robust ability to capture instances of cyberbullying with high precision while ensuring comprehensive coverage of actual positive instances. The accuracy of 0.93 attests to the model's overall proficiency. The Bi-LSTM model exhibited superior performance across all classes, as shown in **Table 3**. Particularly noteworthy is the high precision and recall for the 'Age' and 'Ethnicity' classes, showcasing the model's effectiveness in detecting cyberbullying related to these attributes.

Table 3. Class-wise performance of the Bi-LSTM model.

	Precision	Recall	F1-score	Support
Religion	0.97	0.93	0.95	1572
Age	0.98	0.97	0.97	1560
Ethnicity	0.98	0.98	0.98	1535
Gender	0.96	0.87	0.91	1456
Not bullying	0.77	0.91	0.83	1269

4.2 Comparative analysis of models

The Naive Bayes model, rooted in probabilistic principles, showcases a balanced performance, effectively distinguishing between cyberbullying and non-cyberbullying content. Its reliance on statistical independence assumptions doesn't hinder its effectiveness in this context.

On the other hand, the Bi-LSTM, a deep learning model, leverages the sequential nature of language, capturing intricate patterns within the text. The superior performance metrics highlight its adeptness in discerning the nuanced language indicative of cyberbullying across various attributes.

Both models, while showcasing high accuracy, precision, recall, and F1 score, do so through distinct mechanisms. The Naive Bayes model excels in probabilistic reasoning, while the Bi-LSTM harnesses the power of neural networks to capture complex patterns. The confusion matrices (**Tables 4 and 5**) provide a visual aid in understanding the models' classification outcomes.

These results underscore the potential of diverse approaches in cyberbullying detection, each with its unique strengths. The choice between these models should be guided by the specific requirements and nuances of the online environment under consideration.

Table 4. Naive Bayes confusion matrix.

	Predicted				
	Religion	**Age**	**Ethnicity**	**Gender**	**Not bullying**
Religion	1536	14	10	9	10
Age	11	1541	6	5	3
Ethnicity	58	50	1417	14	3
Gender	30	31	57	1248	96
Not bullying	164	295	85	129	601

Table 5. Bi-LSTM confusion matrix.

	Predicted				
	Religion	**Age**	**Ethnicity**	**Gender**	**Not bullying**
Religion	1463	3	3	2	31
Age	3	1513	2	2	33
Ethnicity	5	4	1499	5	9
Gender	7	3	26	1263	44
Not bullying	96	37	26	180	1152

4.3 Ethical considerations

It is imperative to underscore the ethical considerations that guided this research. The deployment of AI models in sensitive domains such as cyberbullying detection necessitates a conscientious approach to address potential ethical challenges. Aligning with the ethical considerations posited by Bokolo and Liu (2023) [13], our research critically examines the potential biases and fairness issues in the cyberbullying detection process.

1) *Bias Mitigation and Fairness:* Ensuring fairness in our models is a paramount concern. We meticulously examined the training data to identify and rectify biases that might lead to disparate impacts on different demographic groups. This involved scrutinizing the dataset for imbalances in class distribution and refining the model's training to mitigate potential biases.

2) *Privacy Concerns:* Respecting user privacy is central to ethical AI practices. Our study relied on anonymized data to minimize the risk of identifying individuals involved in social media conversations. Additionally, all personally identifiable information was rigorously stripped from the dataset during the preprocessing phase.

3) *Continuous Monitoring:* Ethical considerations extend beyond the development phase to the entire lifecycle of the models. We advocate for continuous monitoring and evaluation of the models' performance in real-world scenarios. Regular assessments help identify and rectify any unforeseen biases or ethical implications that might arise as the models are deployed.

4) *Informed Consent:* When dealing with user-generated content on social media, obtaining explicit consent for data usage is challenging. However, we acknowledge the importance of transparency and informed consent. Our study emphasizes the use of publicly available, anonymized data to respect user privacy while conducting meaningful research.

By addressing biases, ensuring privacy, promoting transparency, and advocating for ongoing monitoring, we strive to uphold the achievable ethical standards in our research and its practical implications.

5. Conclusions

5.1 Summary of research

1) *Introduction Recap:* This research endeavors to tackle the pertinent issue of cyberbullying through the lens of artificial intelligence. With a focus on Twitter data and utilizing the power of machine learning algorithms, the study aims to detect instances of cyberbullying about attributes such as religion, age, ethnicity, and gender.

2) *Methodology Recap:* Commencing with an in-depth methodology, the research encapsulates data preprocessing steps, model training employing Naive Bayes and Bi-LSTM algorithms, and a meticulous evaluation process. Leveraging the sentiment-labeled Sentiment140 dataset, the study repurposes it for cyberbullying detection while addressing ethical considerations in dataset usage.

3) *Results Overview:* The outcomes of the research present a nuanced understanding of the model performances. Both Naive Bayes and Bi-LSTM models exhibit commendable precision, recall, and accuracy, offering promising tools for cyberbullying detection.

5.2 Achievements and contributions

1) *Model Performance:* The study's principal achievements lie in the models' capability to discern cyberbullying across various attributes. The Naive Bayes algorithm showcases robust performance, and the Bi-LSTM model, with its deep learning capabilities, excels in nuanced cyberbullying detection.

2) *Ethical Considerations:* Ethical considerations take center stage in this research, addressing biases in the dataset and ensuring fairness. The commitment to responsible AI development underscores the ethical dimension as an integral part of the study.

5.3 Limitations and challenges

1) *Data Limitations:* While the Sentiment140 dataset proves valuable, the study acknowledges its limitations, suggesting future research explore more specialized datasets for cyberbullying detection.

2) *Model Limitations:* Despite their effectiveness, both models face challenges in handling nuanced expressions and contextual intricacies. Acknowledging these limitations provides avenues for future research and model refinement.

5.4 Implications for future research

Building on the insights gained from the research by Bokolo and Liu (2023) [10], future research directions may explore more advanced deep-learning architectures for enhanced cyberbullying detection.

1) *Further Model Refinement:* The models, although successful, prompt consideration for refinement. Fine-tuning, exploring advanced architectures, and addressing model limitations are avenues for future exploration.

2) *Exploration of New Data Sources:* Diversifying data sources beyond Sentiment140 could enhance model robustness. Investigating datasets explicitly designed for cyberbullying detection is recommended.

3) *Cross-Domain Applications:* Considering the models' adaptability to various platforms and domains presents exciting opportunities. Future research could explore cross-domain applications for a broader societal impact.

5.5 Practical applications

1) *Real-world Implementation:* With promising results, the models hold potential for real-world implementation on social media platforms, providing timely support for individuals facing cyberbullying.

2) *Policy Recommendations:* While cautious in its suggestions, the research hints at the development of policies or interventions based on their outcomes. Ethical deployment and user well-being should guide any policy considerations.

5.6 Concluding remarks

In conclusion, this research contributes valuable insights and tools for combatting cyberbullying. The models presented showcase promising results, affirming the role of artificial intelligence in addressing societal challenges. The commitment to ethical considerations position this study within the framework of responsible AI development, ensuring the tools created serve societal well-being.

Author Contributions

Authors 1 to 4 collaborated on the technical aspects of the research, including a literature review, data collection, preprocessing, feature engineering, and AI model development. Authors 5 to 8 contributed to the ethical considerations, experimental design, validation, and documentation of the cyberbullying detection study, ensuring a comprehensive approach that addresses both technical and ethical aspects.

Conflict of Interest

There is no conflict of interest.

Acknowledgments

We extend our sincere appreciation to Biodoumoye George Bokolo, a final-year doctoral student in Digital and Cyber Forensics Science, for her exceptional contribution to the publication titled "Attribute-Specific Cyberbullying Detection Using Artificial Intelligence". Biodoumoye served as a dedicated mentor and advisor, leveraging her profound expertise in the field of digital forensics and her specialization in utilizing artificial intelligence to detect cybercrime and enhance cybersecurity measures.

Her insightful guidance, unwavering support, and commitment to advancing the realms of digital and cyber forensics have been instrumental in shaping the direction and depth of our research. Biodoumoye's passion for securing cyberspace through innovative AI solutions has left an indelible mark on this project, and we are grateful for her invaluable contributions that have significantly enriched the quality and relevance of our work.

References

[1] Agustiningsih, N., Yusuf, A., Ahsan, A., 2023. Relationships among self-esteem, bullying, and cyberbullying in adolescents: a systematic review. Journal of Psychosocial Nursing and Mental Health Services. 1–7. DOI: https://doi.org/10.3928/02793695-20231013-01

[2] Erbiçer, E.S., Ceylan, V., Yalçın, M.H., et al., 2023. Cyberbullying among children and youth in Türkiye: A systematic review and meta-analysis. Journal of Pediatric Nursing. 73, 184–195. DOI: https://doi.org/10.1016/j.pedn.2023.09.003

[3] Zhang, W., Huang, S., Lam, L., et al., 2022. Cyberbullying definitions and measurements in children and adolescents: Summarizing 20 years of global efforts. Frontiers in Public Health. 10, 1000504.

DOI: https://doi.org/10.3389/fpubh.2022.1000504

[4] Chan, T.K., Cheung, C.M., Lee, Z.W., 2021. Cyberbullying on social networking sites: A literature review and future research directions. Information & Management, 58(2), 103411. DOI: https://doi.org/10.1016/j.im.2020.103411

[5] Civility, Safety and Interaction Online [Internet]. Microsoft; 2020. Available from: https://news.microsoft.com/wp-content/uploads/prod/sites/421/2020/02/Digital-Civility-2020-Global-Report.pdf

[6] Sarker, I.H., 2022. AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. SN Computer Science. 3, 158. DOI: https://doi.org/10.1007/s42979-022-01043-x

[7] Chowdhary, K.R., 2020. Fundamentals of artificial intelligence. Springer India: New Delhi.

[8] Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition [Internet].

[9] Leekha, G., 2021. Learn AI with Python: Explore machine learning and deep learning techniques for building smart AI systems using Scikit-Learn, NLTK, NeuroLab, and Keras (English Edition). BPB Publications: New Delhi.

[10] Bokolo, B.G., Liu, Q. (editors), 2023. Cyberbullying detection on social media using machine learning. IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS); 2023 May 20; Hoboken, NJ, USA. New York: IEEE. p. 1–6. DOI: https://doi.org/10.1109/INFOCOMWKSHPS57453.2023.10226114

[11] Singh, N., Sinhasane, A., Patil, S., et al. (editors), 2020. Cyberbullying detection in social networks: A survey. 2nd International Conference on Communication & Information Processing (ICCIP); 2020 Nov 27–29; Tokyo, Japan.

[12] Elsafoury, F., 2020. "Cyberbullying Datasets," *Mendeley. Com [Internet] [Accessed: 04-Summer-2021]. Available from:* https://www.kaggle.com/datasets/gbiamgaurav/cyberbullying-detection

[13] Bokolo, B.G., Liu, Q., 2023. Combating cyberbullying in various digital media using machine learning. Combatting cyberbullying in digital media with artificial intelligence. Chapman and Hall/CRC: London. pp. 71–97.

[14] Chong, W.J., Chua, H.N., Gan, M.F. (editors), 2022. Comparing zero-shot text classification and rule-based matching in identifying cyberbullying behaviors on social media. 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET); 2022 Sep 13–15; Kota Kinabalu, Malaysia. New York: IEEE. p. 1–5. DOI: https://doi.org/10.1109/IICAIET55139.2022.9936821

[15] Purnamasari, N.M.G.D., Fauzi, M.A., Indriati, L.S.D., et al., 2020. Cyberbullying identification in Twitter using support vector machine and information gain based feature selection. Indonesian Journal of Electrical Engineering and Computer Science. 18(3), 1494–1500. DOI: https://doi.org/10.11591/ijeecs.v18.i3.pp1494-1500

[16] Talpur, B.A., O'Sullivan, D., 2020. Cyberbullying severity detection: A machine learning approach. PloS One. 15(10), e0240924. DOI: https://doi.org/10.1371/journal.pone.0240924

[17] Almomani, A., Nahar, K., Alauthman, M., et al., 2024. Image cyberbullying detection and recognition using transfer deep machine learning. International Journal of Cognitive Computing in Engineering. 5, 14–26. DOI: https://doi.org/10.1016/j.ijcce.2023.11.002

[18] Obaid, M.H., Guirguis, S.K., Elkaffas, S.M., 2023. Cyberbullying detection and severity determination model. IEEE Access. 11, 97391–97399. DOI: https://doi.org/10.1109/ACCESS.2023.3313113

[19] Aizenberg, E., Van Den Hoven, J., 2020. Designing for human rights in AI. Big Data & Society. 7(2).
DOI: https://doi.org/10.1177/2053951720949566

[20] Sap, M., Card, D., Gabriel, S., et al. (editors), 2019. The risk of racial bias in hate speech detection. The 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. p. 1668–1678.

[21] Gasparetto, A., Marcuzzo, M., Zangari, A., et al., 2022. A survey on text classification algorithms: From text to predictions. Information. 13(2), 83.
DOI: https://doi.org/10.3390/info13020083

[22] Kaushik, M., Prakash, P., Ajay, R., et al. (editors), 2020. Tomato leaf disease detection using convolutional neural network with data augmentation. 2020 5th International Conference on Communication and Electronics Systems (ICCES); 2020 Jun 10–12; Coimbatore, India. New York: IEEE. p. 1125–1132.
DOI: https://doi.org/10.1109/ICCES48766.2020.9138030

[23] Bokolo, B.G., Ogegbene-Ise, E., Chen, L., et al. (editors), 2023. Crime-intent sentiment detection on Twitter data using machine learning. 2023 8th International Conference on Automation, Control and Robotics Engineering (CACRE); 2023 Jul 13–15; Hong Kong, China. New York: IEEE. p. 79–83.
DOI: https://doi.org/10.1109/CACRE58689.2023.10208384