Artificial Intelligence in Cybersecurity

Journal Homepage: sciencesforce.com/aics

SCIENCES FORCE
International Academic Publisher

Artificial Intell. Cyb. Vol. 1 (2024) 52–59

Paper Type: Original Article

Innovative Artificial Intelligence Solution as Game Changer in Cyberbullying Detection and Prevention

Salma A. Walli [1,*], **Byeong-Gwon Kang** [2] **and Yunyoung Nam** [2]

[1] Department of Computer science, Faculty of Computer and Informatics, Zagazig University, 44519, Egypt; 20912019200851@fci.zu.edu.eg.

[2] Department of ICT Convergence, Soonchunhyang University, Asan, 31538, Korea; Emails: bgkang@sch.ac.kr; ynam@sch.ac.kr.

Received: 07 Feb 2024 **Revised:** 12 May 2024 **Accepted:** 10 Jun 2024 **Published:** 14 Jun 2024

Abstract

The proliferation of online social networks has brought forth unprecedented connectivity and communication but has also facilitated the emergence of cyberbullying, a pervasive and harmful phenomenon. Traditional methods for identifying cyberbullying often fall short due to the dynamic nature of online interactions and the sheer volume of data. In response, this study explores the application of deep learning techniques for cyberbullying detection, focusing on the integration of LSTM networks with an attention mechanism. The research leverages a diverse dataset encompassing various forms of cyberbullying across age, ethnicity, gender, religion, and non-bullying content. Key findings reveal that the proposed models achieve high accuracy, precision, recall, and F1 scores, effectively classifying instances of cyberbullying with a comprehensive understanding of contextual nuances. Moreover, the study contributes insights into feature extraction methodologies and model optimization techniques, demonstrating the efficacy of deep learning in addressing the complexities of multi-modal social media data.

Keywords: Cyberbullying Detection, Deep Learning, LSTM Networks, Attention Mechanism, Multi-modal Data, Social Media.

1 | Introduction

The proliferation of online social networks (OSNs) and digital communication platforms has revolutionized how individuals interact and share information globally. This digital transformation, while fostering connectivity and information exchange, has simultaneously given rise to a pernicious phenomenon: cyberbullying. Cyberbullying is defined as deliberate, repeated, and hostile behavior directed at an individual or group using information and communication technology (ICT) platforms, such as social media, emails, and instant messaging services [1]. Unlike traditional bullying, which is confined to physical spaces, cyberbullying transcends geographical boundaries and can perpetuate continuously in the victim's virtual spaces, often leading to severe psychological and emotional distress.

As cyberbullying incidents increase, so does the imperative to develop effective methods for detection and prevention. Traditional approaches to identifying cyberbullying, which often rely on manual moderation and user reports, have proven insufficient due to the sheer volume of content generated daily and the subtleties

Corresponding Author: 20912019200851@fci.zu.edu.eg

https://doi.org/10.61356/j.aics.2024.1299

of digital harassment that can escape human detection. This has necessitated the exploration of automated solutions, with deep learning emerging as a particularly promising approach due to its ability to learn complex patterns and nuances from large datasets [2]. Deep learning, a subset of machine learning, involves algorithms that model high-level abstractions in data through architectures composed of multiple layers. These models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown remarkable efficacy in various natural language processing (NLP) tasks, including sentiment analysis, spam detection, and, increasingly, cyberbullying detection [3]. The adoption of deep learning techniques for cyberbullying classification involves processing and analyzing vast amounts of textual data from social media platforms to identify and categorize instances of abuse, thereby enabling timely and effective intervention.

Recent advancements in deep learning have introduced sophisticated models capable of handling the intricacies of cyberbullying. For instance, the integration of feature subset selection methods with deep learning has been proposed to enhance classification accuracy. Techniques such as the binary coyote optimization (BCO) and salp swarm algorithm (SSA) have been used to optimize feature selection and improve model performance [1]. The BCO-SSA-DBN (deep belief network) model has demonstrated superior accuracy in classifying cyberbullying content by effectively handling the complex, multi-dimensional nature of social media data [4].

This paper aims to explore the application of deep learning for cyberbullying classification, focusing on the development and evaluation of advanced deep learning models. this study is structured to first provide a comprehensive overview of cyberbullying detection challenges and the role of deep learning in addressing these challenges. It then details the methodology employed, including dataset description, experimental setup, and evaluation metrics. The results and discussions section presents the findings of the study, highlighting the performance of the proposed LSTM-based models with attention mechanisms. Finally, the conclusion summarizes the key contributions of the research.

2 | Related Work

The detection and classification of cyberbullying on social media using deep learning techniques have garnered significant attention in recent years. This section provides an overview of the prominent contributions in this domain, highlighting the methodologies and results achieved by various researchers.

Murshed et al. (2022) introduced DEA-RNN, a hybrid deep-learning model designed to detect cyberbullying on Twitter. The DEA-RNN model combines Elman-type Recurrent Neural Networks (RNN) with the Dolphin Echolocation Algorithm (DEA) for parameter optimization. This hybrid approach significantly enhances the performance of cyberbullying detection by fine-tuning the RNN’s parameters, thereby reducing training time and improving classification accuracy. The experimental results indicate that DEA-RNN outperforms several state-of-the-art algorithms, including Bi-LSTM, SVM, and Random Forests, achieving superior accuracy and specificity [2]. In a related study, Kumari et al. (2020) developed a hybrid model for detecting cyberbullying on social media, focusing on the integration of text and image data. Their approach employs a pre-trained VGG-16 network for extracting image features and a Convolutional Neural Network (CNN) for text feature extraction. These features are then optimized using a genetic algorithm, enhancing the overall system's efficiency. This innovative model addresses the limitations of separate text and image processing systems by capturing the complex interactions between text and images. Validated on a multimodal dataset, the model achieves a significant improvement in F1-score by 9%, demonstrating its effectiveness over previously reported methods [3].

Another notable contribution is the work by Fang et al. (2021) proposed a model combining bidirectional gated recurrent unit (Bi-GRU) and self-attention mechanism for text-based cyberbullying detection. Their approach addressed challenges such as the subjective nature of cyberbullying identification, the context-free nature of social media posts, and the lack of standardized data. By leveraging the advantages of Bi-GRU and self-attention, they achieved superior performance compared to traditional machine learning methods and conventional deep learning approaches [4]. Moreover, N. Yuvaraj et al. (2021) proposed a nature-inspired automated cyberbullying classification system, integrating a deep reinforcement learning model with artificial neural networks (ANN). This hybrid system leverages psychological and contextual features extracted from social media text to improve classification performance, showcasing higher accuracy compared to traditional machine learning methods [5]. Additionally, Balakrishnan et al. (2020) employed a Random Forest classifier for cyberbullying detection on Twitter, leveraging personality traits and psychological models such as the Big Five and Dark Triad. Their method highlighted the importance of user-specific features in enhancing the precision of cyberbullying classification [5]. These studies underscore the critical role of feature extraction and deep learning in advancing cyberbullying detection technologies, offering robust solutions for automated content moderation on social media platforms.

3 | Deep Learning Approaches for Mitigating Cyberbullying

In this section, we delve into the specific deep-learning techniques employed in our study to mitigate cyberbullying. The focus is on the implementation of two critical components: the Attention mechanism and the LSTM-based Sentiment Classifier. These approaches leverage the power of sequence modeling and context-aware analysis to effectively identify and classify cyberbullying instances.

3.1 | Long Short-Term Memory Networks

LSTM networks have emerged as a prominent deep learning technique for handling temporal sequences and contextual dependencies, which are crucial in tasks such as cyberbullying detection. The primary advantage of LSTM networks lies in their unique architecture designed to address the limitations of traditional RNNs (Recurrent Neural Networks), particularly the vanishing gradient problem, which hampers the learning of long-term dependencies in sequential data [6].

LSTM networks operate by maintaining a series of memory cells and gates that regulate the flow of information, enabling the model to learn which parts of the data to retain and which to discard over long sequences. The fundamental components of an LSTM cell include the input gate i_t , forget gate f_t , cell state c_t , and output gate o_t.These gates are mathematically represented as follows [6]:

- The forget gate decides which information from the previous cell state C_{t-1} should be discarded:

$$f_t = \sigma(U_f x_t + W_f h_{t-1} + b_f) \tag{1}$$

- The input gate determines which new information should be added to the current cell state:

$$i_i = \sigma(U_i x_t + W_i h_{t-1} + b_i) \tag{2}$$

$$C_t^{\sim} = \tanh(U_g x_t + W_g h_{t-1} + b_g) \tag{3}$$

- The cell state is updated as:

$$C_t = f_t \odot C_{t-1} + i_t \odot C_t^{\sim} \tag{4}$$

- The output gate decides what the next hidden state should be:

$$o_t = \sigma(U_o x_t + W_o h_{t-1} + b_o) \tag{5}$$

$$h_t = o_t \odot \tanh C_t \tag{6}$$

3.2 | Attention Mechanism

Attention mechanisms have revolutionized the field of Natural Language Processing (NLP) by enabling models to dynamically focus on the most relevant parts of the input data. This is particularly beneficial for tasks like cyberbullying detection, where the importance of certain words or phrases may vary depending on the context. The core idea behind attention mechanisms is to compute a set of weights that highlight the significance of different parts of the input sequence. This allows the model to selectively focus on the elements that are most relevant to the task at hand, such as identifying abusive language in a social media post [7].

In the context of a sequence-to-sequence model like RNN search, which is often used for tasks such as machine translation, the attention mechanism works as follows [7]:

1. Encoding: The input sequence $x = \{x_1, x_2, \dots \dots, x_t\}$ is encoded into a set of annotations $h = \{h_1, h_2, \dots \dots, h_t\}$ using a BiRNN (Bidirectional Recurrent Neural Network).

$$(h_1, \dots, h_t) = \text{BiRNN}\ (x_1, \dots, x_t) \tag{7}$$

2. Computing Attention Scores: At each time step t, an attention score e_{ti} is computed to assess the relevance of each input annotation h_i with respect to the current state s_{t-1} of the decoder.

$$e_{ti} = f(s_{t-1}, h_i) \tag{8}$$

3. Normalizing Attention Scores: The attention scores are then normalized using a softmax function to produce the attention weights a_{ti}.

$$a_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{t} \exp(e_{ti})} \tag{9}$$

4. Context Vector: The context vector c_t is computed as a weighted sum of the input annotations.

$$c_t = \sum_{i=1}^{t} a_{ti}\, h_i \tag{10}$$

4 | Experiment Design

In this section of our research, we provide a detailed overview of the experimental setup and methodology used for developing our deep learning-based cyberbullying detection system. This encompasses a thorough description of the datasets and their origins, the implementation environment, and the performance evaluation metrics employed.

4.1 | Dataset Description

In the context of cyberbullying detection, especially given its increased relevance during the COVID-19 pandemic, the dataset utilized in this study is instrumental in developing and evaluating robust machine-learning models. This dataset comprises over 47,000 tweets, systematically labeled to reflect the diverse nature of cyberbullying. The six distinct classes include Age, Ethnicity, Gender, Religion, Other types of cyberbullying, and a category for tweets that do not qualify as cyberbullying. Each class contains approximately 8,000 samples, ensuring balanced representation across categories [8].

Cyberbullying on social media is a pressing concern, exacerbated by the widespread reliance on these platforms for communication across all age groups. According to a study, 36.5% of middle and high school students have experienced cyberbullying, while 87% have witnessed it. This pervasive issue has significant psychological impacts, including diminished academic performance, depression, and, in extreme cases, suicidal ideation. Given these statistics, the balanced nature of the dataset is particularly valuable. It enables the development of models that are not biased towards a single category, ensuring a more accurate and generalized detection capability. The dataset's tweets span various forms of bullying, from direct attacks to discriminatory remarks, offering a comprehensive view of the issue [8].

The data collection methodology was designed to ensure a diverse and comprehensive dataset. Tweets were collected from a wide range of public social media profiles, reflecting the diverse nature of the online population. The labeling process involved multiple reviewers to ensure accuracy and consistency in the annotations. This rigorous approach helps minimize biases and ensures that the dataset reflects a broad spectrum of cyberbullying behavior. To facilitate the handling of potentially sensitive content, a trigger warning is included in the dataset description, acknowledging that the tweets may contain descriptions of bullying or be offensive themselves. This caution underscores the importance of ethical considerations when dealing with real-world data that involves personal and potentially distressing content [8].

Table 1. Statistical analysis of the Cyberbullying classification dataset.

	Sentiment	text_len
Count	36273.000000	36273.000000
Mean	1.888457	14.107022
std	1.391364	7.113452
Min	0.000000	0.000000
25%	1.000000	8.000000
50%	2.000000	12.000000
75%	3.000000	20.000000
max	4.000000	31.000000

4.2 | Implementation Setups

For our implementation setup, we utilized a system operating on Microsoft Windows 10 Pro version 10.0.19045, with specifications incorporating an Intel64 Family 6 Model 60 Stepping 3 Genuine Intel processor clocked at approximately 2800 Mhz and a total physical memory of 8,073 MB. The system, an HP ZBook 15 G2, provided a stable and capable environment for our experimentation. Furthermore, we configured our software environment with Python version 3.9.1 and Scikit-learn version 1.4.2. Leveraging the capabilities of Python and sci-kit-learn, we ensured compatibility with the latest tools and libraries for machine learning experimentation. This setup allowed us to efficiently develop and evaluate our intrusion detection algorithms while maintaining consistency and reproducibility throughout the experimentation process.

4.3 | Evaluations Measures

Accuracy stands as a broadly applied metric in multi-class classification, serving as an indicator of the model's proficiency in correctly assigning classes to individual data units. Its derivation hinges on insights stemming from the confusion matrix, a tabular representation detailing both the accurate and inaccurate classifications made by the model [9]. The accuracy formula accentuates the significance of true positives and true negatives, which signify successfully identified instances, against the backdrop of false positives and false negatives, indicating misclassifications.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{13}$$

$$\text{F1} - \text{Score} = \left(\frac{2}{\text{precision}^{-1}+\text{recall}^{-1}}\right) = 2\,.\frac{\text{precision}\,.\,\text{recall}}{\text{precision}+\,\text{recall}} \tag{14}$$

5 | Results & Discussions

In this section, we present and discuss the results of our study on cyberbullying classification using deep learning techniques, specifically focusing on the application of custom Unidirectional and Bidirectional LSTM networks with an attention mechanism. The evaluation metrics used for assessing the performance of our models include the confusion matrix and classification report, providing insights into their accuracy and robustness. The performance of our models is evaluated using a comprehensive set of metrics, including precision, recall, F1-score, and accuracy. The evaluation is carried out on a diverse dataset encompassing various forms of cyberbullying categorized into different classes, such as religion, age, ethnicity, gender, and non-bullying content.

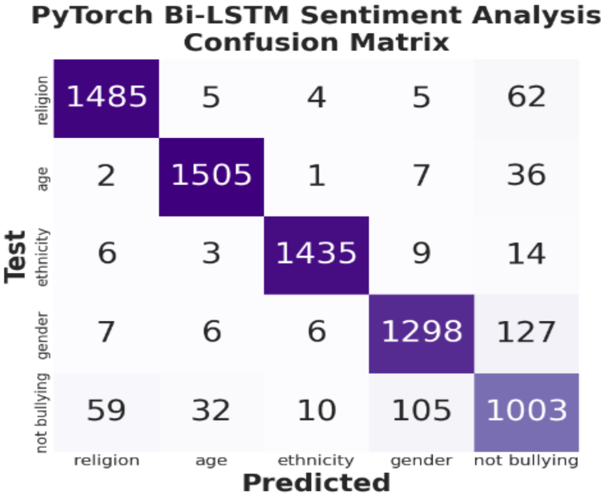

Figure 1. Confusion Matrix for Bi-LSTM model.

The confusion matrix depicted in Figure 1 highlights the classification performance of the Bidirectional LSTM model across five categories: religion, age, ethnicity, gender, and non-bullying. Each entry in the matrix represents the count of instances correctly or incorrectly classified by the model. The diagonal entries indicate the correctly classified instances, while the off-diagonal entries represent the misclassifications.

From the confusion matrix, we observe the following:

- The model shows a high degree of accuracy in classifying instances related to age and ethnicity, with 1505 and 1435 correctly classified instances respectively.
- There is a notable number of misclassifications between religion and non-bullying categories, with 62 instances of religion misclassified as non-bullying and 59 instances of non-bullying misclassified as religion.
- Gender-related cyberbullying instances have the highest misclassification rate, with 127 instances misclassified under the non-bullying category.

Table 2 provides a detailed classification report that includes precision, recall, F1-score, and support for each class. This report is generated to comprehensively evaluate the performance of the Bi-LSTM model on each category of cyberbullying.

Table 2. Classification report for Bi-LSTM model.

Class	Precision	Recall	F1-Score
Religion	0.94	0.94	0.94
Age	0.97	0.97	0.97
Ethnicity	0.98	0.98	0.98
Gender	0.88	0.88	0.88
Not Bullying	0.83	0.83	0.83
Overall	0.92	0.92	0.92

The classification report reveals that:

- The model achieves the highest precision for the ethnicity category (0.98) and the lowest for the non-bullying category (0.83). Precision measures the proportion of true positive predictions in relation to the total number of positive predictions.

- Similarly, the recall is highest for the ethnicity category (0.98) and lowest for the non-bullying category (0.83). Recall measures the proportion of true positive predictions in relation to the total number of actual positives.
- The F1-score, which is the harmonic mean of precision and recall, follows a similar pattern, indicating the model's balanced performance across different classes.

The results indicate that our custom Unidirectional and Bidirectional LSTM models, augmented with an attention mechanism, effectively classify cyberbullying content with notable precision, recall, and overall accuracy of 93%. The high F1 scores across various categories demonstrate the model's robustness in handling diverse forms of cyberbullying. Despite the overall strong performance, certain areas require further improvement. The misclassification rates in the gender and non-bullying categories suggest that the model could benefit from additional training data and potentially more sophisticated feature extraction techniques to better distinguish these categories.

The findings from this study contribute significantly to the field of cyberbullying detection by demonstrating that deep learning models with attention mechanisms can effectively handle the complexities of multi-modal data. This approach not only enhances the classification accuracy but also provides a more nuanced understanding of the contextual elements that contribute to cyberbullying.

6 | Conclusion

This research underscores the pivotal role of deep learning methodologies, specifically LSTM networks enhanced with attention mechanisms, in mitigating the pervasive issue of cyberbullying on social media platforms. The study's findings highlight the models' robust performance in accurately identifying diverse forms of cyberbullying, ranging from discriminatory remarks to direct attacks, across different demographic categories. The achieved high accuracy and balanced evaluation metrics underscore the models' effectiveness in handling the intricacies and evolving nature of online interactions. Moving forward, further advancements in model refinement and the exploration of additional contextual features are recommended to enhance detection capabilities and contribute towards safer digital environments.

Acknowledgments

This research was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0012724, HRD Program for industrial Innovation) and the Soonchunhyang University Research Fund.

Author Contribution

All authors contributed equally to this work.

Funding

This research was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0012724, HRD Program for industrial Innovation) and the Soonchunhyang University Research Fund.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Neelakandan, S., Sridevi, M., Chandrasekaran, S., Murugeswari, K., Pundir, A. K. S., Sridevi, R., & Lingaiah, T. B. (2022). Deep learning approaches for cyberbullying detection and classification on social media. Computational Intelligence and Neuroscience, 2022, Article ID 2163458. https://doi.org/10.1155/2022/2163458.

[2] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif and H. D. E. Al-Ariki, "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform," in IEEE Access, vol. 10, pp. 25857-25871, 2022, doi: 10.1109/ACCESS.2022.3153675.

[3] Kumari, K., & Singh, J. P. (2020). Identification of cyberbullying on multi-modal social media posts using genetic algorithm. Transactions on Emerging Telecommunications Technologies, DOI: 10.1002/ett.3907.

[4] Fang, Y., Yang, S., Zhao, B., & Huang, C. (2021). Cyberbullying Detection in Social Networks Using Bi-GRU with Self-Attention Mechanism. Information, 12(4), 171. https://doi.org/10.3390/info12040171.

[5] Yuvaraj, N., Srihari, K., Dhiman, Gaurav, Somasundaram, K., Sharma, Ashutosh, Rajeskannan, S., Soni, Mukesh, Gaba, Gurjot Singh, AlZain, Mohammed A., Masud, Mehedi, Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking, Mathematical Problems in Engineering, 2021, 6644652, 12 pages, 2021. https://doi.org/10.1155/2021/6644652

[6] Mekruksavanich, S., & Jitpattanakul, A. (2021). LSTM networks using smartphone data for sensor-based human activity recognition in smart homes. Sensors, 21(5), 1636. https://doi.org/10.3390/s21051636

[7] A. Galassi, M. Lippi and P. Torroni, "Attention in Natural Language Processing," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 10, pp. 4291-4308, Oct. 2021, doi: 10.1109/TNNLS.2020.3019893.

[8] J. Wang, K. Fu, C.T. Lu, “SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection,” Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), December 10-13, 2020

[9] Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for Multi-Class Classification: An Overview. CRIF S.p.A. arXiv:2008.05756v1 [stat.ML].

[10] Sripa, K., Thummaphan, P., Ninphet, T., & Gulabutr, V. (2024). Combating cyberbullying in Thai youth: Learning innovation as a possible solution. International Journal of Educational Development, 108, 103065.

[11] Orrù, G., Galli, A., Gattulli, V., Gravina, M., Micheletto, M., Marrone, S., ... & Sansone, C. (2023). Development of Technologies for the Detection of (Cyber) Bullying Actions: The BullyBuster Project. Information, 14(8), 430.

[12] Cedillo, P., Bermeo, A., Betancourth, A., Espinosa, F., Illescas, L., & Jadán, J. (2022). A Systematic Literature Review on Technological Solutions to Fight Bullying and Cyberbullying in Academic Environments. CSEDU (1), 413-420.

[13] Hasan, M. T., Hossain, M. A. E., Mukta, M. S. H., Akter, A., Ahmed, M., & Islam, S. (2023). A review on deep-learning-based cyberbullying detection. Future Internet, 15(5), 179.

[14] Neha, M. V., Muhammad, S., Indu, V., & Thampi, S. M. (2023). Detection and Prevention of Cyberbullying in Social Media Using Cognitive Computational Analysis. In Combatting Cyberbullying in Digital Media with Artificial Intelligence (pp. 18-34). Chapman and Hall/CRC.

[15] Ptaszynski, M. E., & Masui, F. (Eds.). (2018). Automatic cyberbullying detection: Emerging research and opportunities: Emerging research and opportunities.

Disclaimer/Publisher’s Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.