Cyberbullying Classification Using Three Deep Learning models: GPT, BERT, and RoBERTa

Muhammad Abusaqer and Charles Fofie Jr

Department of Math and Computer Science

Minot State University

Minot, ND, USA

muhammad.abusaqer@minotstateu.edu; charles.fofiejr@minotstateu.edu

Abstract

This research paper presents a study on the classification of cyberbullying on social media feeds using deep learning algorithms, including GPT-2, BERT, and RoBERTa Transformers. Cyberbullying is a growing concern in social media, so it is crucial to develop systems for detecting and preventing it. Cyberbullying involves using technology to harass, threaten, embarrass, or target individuals based on age, gender, religion, etc. This paper proposes a system that leverages both deep learning and traditional machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, and Random Forest, to detect cyberbullying and reduce its impact, particularly on teen suicides. The study trains the models on a dataset of 1,000 tweets selected randomly from a larger dataset of 46,692 tweets.

The study compares the performance of these deep learning models to traditional machine learning algorithms in terms of accuracy, precision, recall, and F1-score. The study results demonstrate that the RoBERTa Transformers model outperforms the other models, highlighting the effectiveness of leveraging large-scale pre-trained language models for cyberbullying detection. However, the results also reveal that traditional machine learning algorithms, such as SVM and Random Forest, can still offer competitive performance compared to some transformer-based models, particularly when computational resources are limited.

This study makes significant contributions to the field by providing a performance comparison between state-of-the-art deep learning models and traditional machine learning algorithms for cyberbullying detection. In addition, the results of this study could help develop tools to assist in monitoring social media for cyberbullying feeds and immediately deleting them, thereby ensuring the safety and well-being of online users.

1 Introduction

The advent of social media has revolutionized communication, allowing people from around the world to connect and share their lives. However, this unprecedented access to global communication has given rise to a troubling phenomenon: cyberbullying. Cyberbullying is defined as the use of technology to harass, threaten, embarrass, or target individuals based on factors such as age, gender, or religion. As a growing concern, particularly among adolescents, cyberbullying has been associated with severe psychological consequences, including depression, anxiety, and even suicide. Therefore, it is imperative to develop robust systems for detecting and preventing cyberbullying on social media platforms.

This research paper presents a comprehensive study on the classification of cyberbullying on social media feeds using three state-of-the-art deep learning algorithms: GPT-3 from OpenAI, BERT from Google, and RoBERTa from Facebook AI. These deep learning models are trained on a dataset of 46,692 tweets to detect instances of cyberbullying. Additionally, the study compares the performance of these deep learning models with traditional machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, and Decision Trees, which have been widely used in previous research on cyberbullying detection.

The primary contributions of this study are twofold. First, it is among the first studies to employ the newly released GPT-3, BERT, and RoBERTa deep learning models in the context of cyberbullying detection. Second, it provides a comprehensive performance comparison between these cutting-edge deep learning models and traditional machine learning algorithms. By demonstrating the superiority of deep learning models in detecting cyberbullying, the results of this study have the potential to inform the development of tools that can aid in monitoring social media for cyberbullying content and enable timely intervention, ultimately creating a safer online environment for users of all ages.

2 Related Work

Researchers studied cyberbullying using different machine learning and deep learning algorithms.

A recent study by Hani and his colleagues presents a supervised machine-learning method aimed at detecting and mitigating cyberbullying [1]. Authors employed various classifiers to train and find bullying behavior. Upon evaluation of the proposed technique using a cyberbullying dataset revealed that the Neural Network (NN) model outperformed other classifiers, achieving an accuracy rate of 92.8%. The Support Vector Machine (SVM) model followed closely with an accuracy of 90.3%. Moreover, the NN model proved superior performance compared to classifiers employed in similar research efforts when tested on the same dataset.

In a study focusing on automatic cyberbullying detection in a social media text, researchers explored the feasibility of identifying posts written by bullies, victims, and bystanders in online bullying situations [2]. They developed a fine-grained annotated cyberbullying corpus for both English and Dutch languages and employed linear support vector machines with a rich feature set to perform a series of binary classification experiments. The study aimed to find which information sources contribute the most to the task of automatic cyberbullying detection. The results showed promising outcomes for detecting cyberbullying-related posts, with the optimized classifier achieving F1 scores of 64% and 61% for English and Dutch, respectively, significantly outperforming baseline systems [2].

In a recent study examining the rise of cyberbullying in digital spaces, particularly social media, researchers aimed to detect cyberbullying comments automatically using machine learning and deep learning techniques [3]. They highlighted the various forms of cyberbullying, such as sexual remarks, threats, hate mail, and spreading false information about individuals, and noted the long-lasting impacts on victims, both physically and psychologically. The study revealed an increase in cyberbullying-related suicides in recent years, with India being one of the top four countries with the highest number of cases. To address this issue, the researchers employed metrics like accuracy, precision, recall, and F1-score to evaluate the performance of their models [3]. The study found that the Gated Recurrent Unit, a deep learning technique, outperformed all other techniques considered in the paper, achieving an impressive accuracy of 95.47%.

In a recent study examining the critical role of cybersecurity in safeguarding complex networks of client and organization data, researchers emphasized the importance of cybersecurity for individuals, families, corporations, agencies, and educational institutions [4]. The study highlights the potential of machine learning in advancing the cybersecurity landscape, particularly in light of the massive amounts of data collected by modern businesses and infrastructure systems. As data becomes increasingly central to various business-focused and infrastructure systems, the authors argue that machine learning and artificial intelligence are gaining traction across all domains of today's systems, whether on-premises or in the cloud [4]. By incorporating these advanced technologies, cybersecurity teams may be better equipped to protect sensitive data and maintain the integrity of mission-critical systems.

Trong and his colleagues focused on detecting cybersecurity events. The researchers emphasized the importance of event detection (ED) to identify event trigger words within the cybersecurity domain [5]. To facilitate future research, the authors introduced a new dataset for this problem, comprising manual annotations for 30 significant cybersecurity event types and a large dataset size suitable for developing deep learning models. Compared to previous datasets for this task, the new dataset includes more event types and supports the modeling of document-level information, potentially enhancing performance [5]. The researchers conducted extensive evaluations using current state-of-the-art methods for ED on the proposed dataset, revealing the challenges associated with cybersecurity ED and presenting numerous research opportunities in this area for future work.

3 Proposed Methodology

3.1 Dataset

The study used a dataset from Kaggle. The dataset has 47,693 tweets with cyberbullying labels. Each tweet is labeled to one of these six classes: "not_cyberbullying", "other_cyberbullying", "age", "ethnicity", "gender": 4, and "religion". Because of limited computing power, the authors randomly sampled 1000 rows from the dataset for running the experiments.

3.2 Dataset Cleaning

The tweet data has been cleaned from punctuation, stopwords, and nonalphanumeric text, as they do not contribute to the classification. Also, the text was transferred to lowercase.

3.3 Machine Learning

The study used three machine learning classifiers, Multinomial Naïve Base (MultinomialNB), Support Vector Machine (SVM), and RandomForest (RF).

3.4 Transformers for NLP

Transformers form the underlying architecture for many popular NLP models, such as BERT, RoBERTa, and GPT. They were proposed by a team of researchers from Google in 2017 in the paper, Attention Is All You Need [6]. The study used three transformers models, GPT – 2.0, RoBERTa, and BERT.

3.5 Evaluation Metrics

Accuracy, precision, recall, and F1-score are common evaluation metrics used to assess the performance of classification models. These metrics provide insights into the model's ability to identify and classify instances in the data correctly, and each metric focuses on a different aspect of the classification task.

3.5.1 Accuracy

Accuracy represents the ratio of correct predictions (encompassing both true positives and true negatives) made by the model to the entire number of instances within the dataset. This metric is frequently utilized to evaluate a classifier's overall performance [7].

Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives) [8]

Nonetheless, relying solely on accuracy might be unsuitable when dealing with imbalanced data, as it could produce misleading results when the majority of instances pertain to a single class [9].

3.5.2 Precision

Precision refers to the fraction of true positives out of all instances predicted as positive by the model [7]. In essence, it evaluates the classifier's capability to accurately identify positive instances among all predictions deemed positive.

Precision = True Positives / (True Positives + False Positives) [8]

Precision serves as a valuable metric in situations where the consequences of false positives are significant, such as spam detection. In this context, wrongly classifying a legitimate email as spam could result in the loss of crucial information [10].

3.5.3 Recall

Recall, also referred to as sensitivity or true positive rate, represents the fraction of true positives out of the entire count of actual positive instances in the dataset [7]. It assesses the classifier's capacity to identify all positive instances accurately.

Recall = True Positives / (True Positives + False Negatives) [8]

Recall emerges as a critical metric in scenarios where the repercussions of false negatives are substantial, such as in medical diagnoses where undetected diseases can lead to grave consequences [11].

3.5.4 F1-score

The F1-score serves as the harmonic mean of precision and recall, offering a unified metric that harmonizes both precision and recall [12]. It proves particularly valuable when working with imbalanced datasets since it considers both false positives and false negatives.

F1-score = 2 * (Precision * Recall) / (Precision + Recall) [8]

An F1-score of 1 signifies impeccable precision and recall, while an F1-score of 0 denotes that either precision or recall (or both) amounts to zero [13].

4 Results

In this study, the authors evaluated various machine learning models, including BERT Transformers, RoBERTa Transformers, GPT Transformers, Random Forest Classifier, Multinomial Naïve Bayes, and Support Vector Machine (SVM) to detect cyberbullying in a dataset of tweets. The models were trained and tested on a smaller subset of the dataset (1000 samples) for computational efficiency. The performance of each model was measured using accuracy, precision, recall, and F1-score. The results obtained are shown in Table 1 contains data.

	Accuracy	Precision	Recall	F1-Score
RoBERTa Transformers	0.83	0.82	083	0.82
BERT Transformers	0.78	0.76	078	0.73
GPT Transformers	0.18	0.03	0.18	0.05
RandomForest	0.75	0.79	0.75	0.76
MultinomialNB	0.71	0.73	0.71	0.68
SVM	0.75	0.80	0.75	0.76

Table 1: Evaluation results of ML and Transformers.

5 Discussion

The results indicate that among the evaluated models, the RoBERTa Transformers achieved the highest performance in terms of accuracy, precision, recall, and F-1 score. This demonstrates the potential of using advanced pre-trained transformer models for cyberbullying detection tasks. On the other hand, the GPT Transformers showed significantly lower performance compared to the other models, possibly due to the model's inherent design as a generative language model rather than a classification model.

The traditional machine learning algorithms, such as Random Forest, Multinomial Naïve Bayes, and SVM, showed competitive performance compared to the BERT Transformers, with SVM having similar performance in accuracy, recall, and F1-score. This suggests that, despite the advancements in deep learning and natural language processing, traditional machine learning algorithms still hold potential for cyberbullying detection tasks, especially when computational resources are limited.

In conclusion, the choice of the model for detecting cyberbullying in social media text data depends on the available computational resources and the desired performance metrics. While the RoBERTa Transformers model provides the best overall performance, traditional machine learning algorithms, such as SVM and Random Forest, can still offer competitive results with lower computational requirements.

6 Future work

Future research in this area could explore the use of other advanced transformer models, such as GPT-3, or investigate the benefits of combining multiple models through ensemble techniques to further improve classification performance. Additionally, examining the impact of different data preprocessing and feature engineering methods, as well as incorporating domain-specific knowledge, could provide valuable insights for enhancing the detection of cyberbullying in social media text data.

7 Conclusion

In this research, the authors investigated the performance of various machine learning models, including BERT Transformers, RoBERTa Transformers, GPT Transformers, Random Forest Classifier, Multinomial Naïve Bayes, and Support Vector Machine (SVM), for the task of cyberbullying detection in social media text data. The experiments demonstrated the potential of advanced pre-trained transformer models in achieving high performance for this challenging task.

The RoBERTa Transformers model outperformed the other models in terms of accuracy, precision, recall, and F1-score, highlighting the effectiveness of leveraging large-scale pre-trained language models for cyberbullying detection. Despite the relatively lower performance of the GPT Transformers model, the results emphasize the importance of model selection and fine-tuning strategies to match the characteristics of the specific classification task.

The study also revealed that traditional machine learning algorithms, such as SVM and Random Forest, can still offer competitive performance compared to some transformer-based models, particularly when computational resources are limited. These findings underscore the value of considering a diverse range of classification techniques for cyberbullying detection, depending on the available resources and desired performance metrics.

In conclusion, this research contributes to the growing body of literature on the application of machine learning for cyberbullying detection, offering valuable insights into the performance of various models and guiding the development of more effective and efficient detection systems. As social media continues to play an increasingly prominent role in current days, the ability to accurately and swiftly identify and address instances of cyberbullying is critical for ensuring the safety and well-being of online users.

References

[1] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social Media Cyberbullying Detection using Machine Learning," *International Journal of Advanced Computer Science and Applications*, vol. 10, pp. 703–707, Jan. 2019, doi: 10.14569/IJACSA.2019.0100587.

[2] C. V. Hee *et al.*, "Automatic detection of cyberbullying in social media text," *PLOS ONE*, vol. 13, no. 10, p. e0203794, Oct. 2018, doi: 10.1371/journal.pone.0203794.

[3] A. K. G and D. Uma, "Detection of Cyberbullying Using Machine Learning and Deep Learning Algorithms," in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, Aug. 2022, pp. 1–7. doi: 10.1109/ASIANCON55314.2022.9908898.

[4] R. Kumar and E. Al, “Detection of Cyberbullying using Machine Learning,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 9, Art. no. 9, Apr. 2021, doi: 10.17762/turcomat.v12i9.3131.

[5] H. Man Duc Trong, D. Trong Le, A. Pouran Ben Veyseh, T. Nguyen, and T. H. Nguyen, “Introducing a New Dataset for Event Detection in Cybersecurity Texts,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 5381–5390. doi: 10.18653/v1/2020.emnlp-main.433.

[6] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Mar. 30, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[7] S. Kotsiantis, I. Zaharakis, and P. Pintelas, “Machine learning: A review of classification and combining techniques,” *Artificial Intelligence Review*, vol. 26, pp. 159–190, Nov. 2006, doi: 10.1007/s10462-007-9052-3.

[8] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[10] G. V. Cormack, “Email Spam Filtering: A Systematic Review,” *INR*, vol. 1, no. 4, pp. 335–455, Jun. 2008, doi: 10.1561/1500000006.

[11] T. Fawcett and P. Flach, “A Response to Webb and Ting’s On the Application of ROC Analysis to Predict Classification Performance Under Varying Class Distributions,” *Machine Learning*, vol. 58, pp. 33–38, Jan. 2005, doi: 10.1007/s10994-005-5256-4.

[12] Y. Yang and X. Liu, “A re-examination of text categorization methods,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley California USA: ACM, Aug. 1999, pp. 42–49. doi: 10.1145/312624.312647.

[13] D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.” arXiv, Oct. 10, 2020. doi: 10.48550/arXiv.2010.16061.