A Comparative Analysis of Transformer and Traditional ML Models for Cyberbullying Detection on Twitter (now X)

Muhammad Abusaqer
Department of Mathematics & Computer Science
Minot State University, Minot, ND, USA
muhammad.abusaqer@minotstateu.edu

Jamil Saquer
Department of Computer Science
Missouri State University, Springfield, MO, USA
jamilsaquer@missouristate.edu

***Abstract*—Cyberbullying on social media poses critical risks to mental health and public safety. This paper investigates advanced computational approaches for detecting and classifying cyberbullying in a large Twitter dataset of 47,692 tweets, labeled into six categories. We compare three transformer-based models (GPT-3.5, BERT, and RoBERTa) against three traditional machine learning algorithms (Naïve Bayes, SVM, and Random Forest), evaluating accuracy, precision, recall, F1-score, and computational efficiency. Our results indicate that RoBERTa achieves the highest overall performance (87–88% accuracy) but at a higher computational cost (13 hours on CPU), while Random Forest offers a strong balance between speed and performance (85.36% accuracy in 83 seconds). In contrast, the experiment using GPT-3.5 in a batched, zero-shot configuration achieved accuracy of 25.41%, an F1-score of 23.61%, and elapsed time of 5.14 hours, highlighting the challenges of applying generative models to cyberbullying detection without fine-tuning. These findings inform model selection for real-world deployment of cyberbullying detection systems, illuminating the trade-offs between transformer-based and traditional methods for automated social media monitoring.**

***Index Terms*—cyberbullying detection, machine learning, deep learning, transformers, text classification, natural language processing.**

I. Introduction

Cyberbullying, a pervasive form of harassment in the digital age, has intensified with the widespread use of social media platforms, where anonymity and global reach allow perpetrators to act without accountability, intensifying its psychological and societal impact [1]. Studies indicate that cyberbullying is particularly detrimental to adolescents, contributing to severe psychological consequences such as anxiety, depression, and, in extreme cases, suicide [1]–[3]. As reliance on digital communication escalates, robust, real-time detection tools are essential for reducing the prevalence and impact of cyberbullying content.

Traditional machine-learning algorithms, such as Naïve Bayes, Support Vector Machines (SVM), and Random Forest, remain popular for text-classification tasks, including automatic detection of offensive or hateful content, because of their implementation simplicity and computational efficiency [4]–[6]. Practical deployment typically relies on prior domain literature, comparative studies, and subsequent empirical testing to identify the most suitable classifier for a given dataset [4]–[6]. Nevertheless, these methods often fail to capture the context-dependent linguistic subtleties, such as sarcasm, irony, colloquialisms, and rapidly evolving slang, that characterise social-media discourse, thereby limiting accuracy in cyberbullying detection [7]–[9]. Such shortcomings motivate research on transformer-based architectures, whose deep contextual representations model nuanced semantics more effectively [7].

In contrast, transformer-based models leverage large-scale pre-training and context-aware embeddings to capture semantic subtleties in language, leading to superior performance on various NLP tasks [10]. Recent transformer architectures such as BERT [11], GPT [12], and RoBERTa [13] have demonstrated state-of-the-art accuracy in sentiment analysis, hateful speech detection, and other text classification tasks [10], [14]–[16]. Despite these performance gains, transformer models also demand significant computational resources, posing challenges for large-scale or real-time applications.

This research aims to:

1) Compare the performance of transformer-based models (GPT-3.5, BERT, and RoBERTa) and traditional ML classifiers (Naïve Bayes, SVM, Random Forest) for detecting cyberbullying on a large Twitter dataset of nearly 47,692 tweets.
2) Evaluate models in terms of accuracy, precision, recall, F1-score, and computational efficiency (elapsed time), highlighting the trade-offs between performance and cost.
3) Recommend practical strategies for deploying accurate and scalable cyberbullying detection systems in real-world contexts.

The study's empirical results show that RoBERTa outperforms other models in classification metrics, while Random Forest offers a viable accuracy–efficiency balance for resource-constrained environments. The final choice of the model depends on available computational resources and application-specific requirements (e.g., real-time vs. offline detection).

Accepted version of the paper published in Proc. 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), Toronto, Canada, 2025, pp. 1607 to 1612. Published version: https://doi.org/10.1109/COMPSAC65507.2025.00216. © 2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

II. Related Work

A. Increasing Cyberbullying During Global Crises

Researchers have recently observed that cyberbullying intensifies during global crises, such as the COVID-19 pandemic [17]. For instance, Wang et al. [17] introduced a multi-category detection system to classify bullying behaviors targeting various protected attributes (age, gender, religion, ethnicity). While this work highlighted fine-grained detection, it gave limited attention to comprehensive data preprocessing (e.g., removing tweets beyond 280 characters).

B. Traditional Models for Cyberbullying Detection

Prior studies utilized Support Vector Machines (SVM), Naïve Bayes, and Random Forest to detect harassing language [9], [18]. Van Hee et al. [9] developed a multilingual detection strategy emphasizing sentiment lexicons and topic modeling, achieving F1-scores (64.32% for English, 58.72% for Dutch). Similarly, Hani et al. [18] compared Neural Networks and SVM on a small dataset of 1,608 samples, revealing that advanced feature extraction with TF-IDF and sentiment cues can significantly improve detection. However, traditional ML classifiers often underperform when faced with sarcasm, indirect bullying, and evolving slang [19]–[21].

C. Deep Learning Approaches for Cyberbullying Detection

Recent studies have employed deep learning methods, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, for cyberbullying detection. CNNs have been utilized to capture local contextual features in text by applying convolutional filters, while LSTM architectures have been used to model sequential dependencies inherent in social media language. Abusaqer et al. [22] demonstrated that a RoBERTa-based sentiment model, fine-tuned on nearly one million tweets about large-scale hospital cyberattacks, can reliably track user attitudes toward security incidents, further underscoring the value of transformer architectures for safety-related discourse. Hybrid CNN-LSTM models have been shown to outperform traditional classifiers in capturing subtle indicators of bullying [17]. Moreover, LSTM networks have proven effective in modeling the temporal dynamics and contextual nuances of cyberbullying content [23]. Although these methods require substantial computational resources and large training datasets, their ability to automatically extract hierarchical features from raw text renders them promising for addressing the complexities of cyberbullying detection.

D. Advancements in Transformer Models

Transformer-based models, including BERT [11], GPT [12], and RoBERTa [13], employ deep contextual embeddings and attention mechanisms to grasp nuanced language structures. Several works leverage these properties for improved hateful content detection [14], [15], achieving higher F1-scores than traditional approaches. More recent endeavors combine multi-task learning and sarcasm detection to address subtle bullying cues [10], [16]. Nevertheless, these models often require substantial computational resources, raising questions about deployment feasibility for real-time or large-scale social media monitoring.

E. Summary and Research Gaps

The literature confirms that transformer-based methods can deliver high accuracy but introduce computational hurdles for large-scale or real-time detection. The present study systematically compares advanced transformers and traditional algorithms on a large dataset (47k tweets), focusing on classification performance and resource constraints to identify the trade-offs relevant for practical deployment.

III. Methodology

A. Description

The dataset comprises 47,692 tweets labeled into six classes: *not_cyberbullying*, *other_cyberbullying*, *age*, *ethnicity*, *gender*, and *religion* [20], [24]–[29]. In this multi-class classification framework, the label *not_cyberbullying* denotes the absence of any cyberbullying behavior and is treated as one class among the six possible outcomes. Table I outlines the distribution of instances across these classes.

Each label represents a specific aspect of cyberbullying:

- **Age:** Cyberbullying based on the age of the target.
- **Ethnicity:** Cyberbullying targeting individuals based on their ethnic background.
- **Gender:** Cyberbullying directed at individuals based on their gender.
- **Religion:** Cyberbullying associated with an individual's religious beliefs.
- **other_cyberbullying:** General instances of cyberbullying that do not specifically target the aforementioned attributes.
- **not_cyberbullying:** Tweets that do not exhibit any form of cyberbullying.

This dataset, further elaborated by Wang et al. in SOS-Net [17], was constructed using a semi-supervised Dynamic Query Expansion (DQE) technique to address class imbalances and to provide insights into evolving patterns of online harassment during the COVID-19 pandemic. Preliminary analysis revealed a relatively balanced class distribution, with an average tweet length of 136.25 characters. Tweets exceeding the 280-character Twitter limit were trimmed to ensure uniformity and relevance.

TABLE I
Class Distribution of the Six Cyberbullying Types

Cyberbullying Type	Count
Religion	7998
Age	7992
Gender	7973
Ethnicity	7961
not_cyberbullying	7945
other_cyberbullying	7823

B. Dataset Preprocessing

Preprocessing steps included expansion of contractions (e.g., "don't" → "do not"), and duplicate removal reduced the dataset from 47,692 to 45,020 unique entries. Punctuation, hyperlinks, extra spaces, and numeric data were stripped during text cleaning, while stopwords were removed using the NLTK library to enhance semantic value of the text. Repetition handling eliminated repeating characters (for example, "heyyy" → "hey"). These steps follow standard practices in text mining and natural language processing (NLP) [30], [31].

C. Classification Models

The study investigates supervised classification models, which learn from labeled data to predict outcomes. Two categories of models were considered.

1) Traditional Machine Learning Models: Traditional machine learning refers to algorithmic approaches that learn statistical relationships from manually engineered features. These models, such as Multinomial Naïve Bayes, Support Vector Machine (SVM), and Random Forest, are characterized by their reliance on predefined mathematical formulations and heuristic feature extraction, yielding better interpretability and computational efficiency than deep learning models [23], [32], [33]. For example, SVM employs a radial basis function (RBF) kernel that implicitly maps input data into an infinite-dimensional space [34].

2) Transformer-Based Models: Transformer-based models are deep neural architectures that utilize self-attention mechanisms to capture contextual and semantic dependencies in text. Pre-trained on large corpora and subsequently fine-tuned for specific tasks, these models (e.g., BERT, GPT-3.5, RoBERTa) have achieved state-of-the-art performance on numerous natural language processing tasks [11]–[13]. Their ability to automatically learn hierarchical representations contrasts with the feature engineering required by traditional methods.

In summary, while traditional ML models rely on handcrafted features and are computationally efficient, transformer-based models learn complex representations automatically but require substantial computational resources and longer training times.

D. Evaluation Metrics

Model performance was assessed using standard evaluation metrics—accuracy, precision, recall, and F1-score—which are widely accepted in classification tasks. In addition, computational efficiency was evaluated by measuring the elapsed time required for model training and inference. Although elapsed time is not a conventional metric for classification performance, it is crucial for real-world applications that demand rapid deployment and real-time processing [35].

IV. Preliminary Analysis

This section provides insights into the dataset characteristics and distributions before model training. Understanding these properties clarifies the linguistic context and complexity of the data, thereby guiding model selection and hyperparameter tuning.

A. Frequency Analysis

Table II shows the ten most common words in the entire corpus. Personal pronouns like "I" appear frequently, suggesting that tweets often adopt a personal or direct tone. Offensive or condescending terms (e.g., "dumb," "bullied") also surface in the top ten, hinting at the corpus's aggression or hateful undertones. This frequency-based snapshot highlights the self-referential and emotional quality of many tweets, which can significantly impact how ML or transformer models differentiate legitimate conversation from bullying.

TABLE II
THE 10 MOST COMMON WORDS IN THE CORPUS

Term	Frequency
I	12,637
school	7,326
like	4,826
high	4,542
dumb	4,321
bullied	4,317
people	4,177
RT	3,830
rape	3,380
f**k	3,203

V. Experimental Setup and Results

A. Data Splitting and Feature Extraction

Stratified sampling was applied to partition the dataset, ensuring balanced representation across classes. For transformer-based models that require training (BERT and RoBERTa), an 80:20 training-to-testing split was employed. In contrast, GPT-3.5 was deployed in a zero-shot configuration, where classification was performed on the entire dataset using prompt-based inference in batched mode. For traditional ML approaches, an 80:20 split was used for Naïve Bayes and SVM, while a 75:25 split was adopted for Random Forest. The 75:25 split for Random Forest was chosen because ensemble methods such as Random Forest have been observed to overfit when trained on small training sets; using a slightly smaller training set and a larger testing set provides a more robust estimate of its generalization performance [23]. For traditional ML methods, text was first transformed into numerical features using term Frequency-Inverse Document Frequency (TF-IDF) vectorization with a maximum feature set of 5,000 tokens, thereby balancing efficiency with representativeness [7], [8].

B. Model Configurations and Hyperparameters

The experimental settings and hyperparameters for each model were as follows:

- **Multinomial Naïve Bayes:** Implemented with scikit-learn's default parameters.
- **Support Vector Machine (SVM):** Utilized with a radial basis function (RBF) kernel that implicitly maps data to an infinite-dimensional feature space. The default hyperparameters were employed.

- **Random Forest:** Configured with 100 decision trees (`n_estimators=100`) to mitigate overfitting via majority voting.
- **BERT ("bert-base-uncased"):** Fine-tuned with a batch size of 16 over 5 epochs using the AdamW optimizer at a learning rate of 2×10^{-5}.
- **GPT-3.5 ("gpt-3.5-turbo", zero-shot):** A zero-shot prompt was used (i.e., no few-shot examples were provided). The simplified system message was:

  > "You are an expert tweet classifier specialized in detecting cyberbullying. Your task is to analyze each of the following tweets and determine whether it involves cyberbullying. If it does, assign it exactly one of the following categories: not_cyberbullying, other_cyberbullying, age, ethnicity, gender, religion. If no cyberbullying is detected, label it as not_cyberbullying. Output the category for each tweet on a new line, in the same order as the tweets appear."

  To efficiently process the dataset and manage API rate limits, tweets were grouped into batches of 20. For each batch, a prompt was constructed that enumerated the tweets, and GPT-3.5 was instructed to return one category per tweet on separate lines. An exponential backoff strategy (applying 5- or 10-second sleep delays upon rate limit errors) was incorporated, along with a fixed 1-second delay between batches; the cumulative sleep time was tracked to differentiate between active processing time and overall elapsed time [12].
- **RoBERTa ("roberta-base"):** Fine-tuned under similar conditions as BERT, omitting the next-sentence prediction objective [13].

C. Hardware and Software Environment

Experiments were conducted on a Windows 11 system equipped with a 13th Gen Intel® Core™ i9-13950HX CPU (24 physical/32 logical cores, up to ∼2.42 GHz) and 63.61 GB of RAM. All transformer-based models were trained on CPU, significantly impacting their training times compared to traditional ML models.

D. Experimental Results

The traditional ML models achieved the performance shown in Table III.

TABLE III
PERFORMANCE METRICS FOR TRADITIONAL ML MODELS

Model	Accuracy (%)	Precision	Recall	F1-Score
Random Forest	**85.36**	0.8560	0.8536	**0.8545**
SVM	84.01	0.8499	0.8402	0.8440
Naïve Bayes	78.01	0.7658	0.7801	0.7646

In contrast, the transformer-based models yielded the performance metrics shown in Table IV.

E. Hyperparameter Tuning

While initial hyperparameters (e.g., learning rate, batch size) for traditional models and for fine-tuning transformer-based models (BERT and RoBERTa) were derived from established defaults and refined via a small-scale tuning procedure on a 5% subset of the training data, no gradient-based fine-tuning was performed for GPT-3.5. Instead, we leveraged GPT-3.5 in a zero-shot configuration.

Preliminary experiments investigated various batching strategies to balance classification focus and processing efficiency. These experiments revealed that grouping tweets into batches of 20 yielded the best trade-off, as larger batches tended to dilute the model's focus on individual tweets. The GPT-3.5 setup employed a fixed prompt with a simplified system message (i.e., without few-shot examples), a temperature of 0.0, and a maximum token limit of 300 per API call. Since no fine-tuning was performed for GPT-3.5, hyperparameters such as learning rate, number of epochs, and task-specific batch size are not applicable.

This preliminary tuning ensured that each model was given comparable opportunities to reach its best performance within resource constraints (13th Gen Intel® Core™ i9-13950HX CPU, 24 cores, 63.61 GB RAM).

VI. DISCUSSION

A. Comparative Analysis of Model Performance

The comparative analysis indicates that transformer-based models, particularly BERT and RoBERTa, outperform traditional ML classifiers in predictive accuracy and F1-score. However, these models require substantially higher computational resources, rendering them less practical for real-time applications without GPU acceleration. In contrast, traditional methods such as Random Forest deliver a favorable balance between accuracy (approximately 85%) and computational efficiency (with training times around 83 seconds), making them attractive for resource-constrained environments. Naïve Bayes and SVM provide acceptable performance but fall short of the transformer-based models in overall predictive power.

GPT-3.5, employed in a zero-shot configuration through prompt-based batched inference, achieved an accuracy of 25.41% and an F1-score of 0.2361 over 18,521 seconds of total elapsed time. This result highlights the limitations of using a zero-shot prompt for cyberbullying detection without further fine-tuning or prompt engineering. Table V summarizes the overall performance metrics across all models.

B. Performance vs. Efficiency

RoBERTa emerged as the top-performing classifier in terms of both F1-score (approximately 0.87) and accuracy (87–88%). However, its 13-hour CPU training time restricts its use in real-time or resource-constrained deployments. In contrast, Random Forest, achieving an F1-score of about 0.85 with an accuracy of 85%, required only 83 seconds, offering a strong compromise for large-scale or near-real-time pipelines.

TABLE IV
PERFORMANCE METRICS FOR TRANSFORMER-BASED MODELS

Model	Accuracy (%)	Precision	Recall	F1-Score	Epochs	Batch Size	Max Len	Time (sec)
GPT-3.5 (zero-shot)	25.41	48.23	25.41	23.61	0	20	300	18,521
BERT	86.88	87.25	86.88	86.82	5	16	128	47,355
RoBERTa	**87.01**	86.82	87.01	**86.85**	5	16	128	46,800

TABLE V
OVERALL COMPARISON OF MODEL PERFORMANCE

Model	Accuracy (%)	F1-Score	Time (sec)
Naïve Bayes	78.01	0.76	0.04
SVM	84.01	0.84	82
Random Forest	85.36	0.85	83
GPT-3.5	25.41	23.614	18521
BERT	86.88	0.87	47,355
RoBERTa	**87.01**	**0.87**	46,800

C. Implications for Model Selection

The high F1-scores obtained by RoBERTa and BERT indicate their robustness in capturing nuanced features such as sarcasm and subtle indicators of cyberbullying. For instance, the tweet

"Rebecca Black Drops Out School Due Bullying"

was correctly classified as *not_cyberbullying* by both RoBERTa and BERT, whereas other classifiers misclassified it as *ethnicity*, *other_cyberbullying*, or *age*.

D. Explaining GPT-3.5's Underperformance

Although GPT-3.5-turbo has demonstrated robust performance in a variety of natural language tasks, the used model of zero-shot configuration for cyberbullying classification achieved an F1-score of approximately 0.24. In this experiment, no few-shot examples or fine-tuning were employed; instead, a simplified prompt was used to facilitate efficient batch processing. This minimal context appears to have limited GPT-3.5's ability to capture the nuanced linguistic cues necessary for accurately classifying subtle instances of cyberbullying. Future work should investigate the incorporation of few-shot examples or task-specific fine-tuning to better harness GPT-3.5's potential for cyberbullying detection.

E. Real-Time Constraints and Ethical Considerations

Certain applications, such as automated moderation systems, demand sub-second or near-real-time performance. Although inference is generally much faster than training, transformer-based models like RoBERTa and BERT can still incur non-negligible inference delays—especially when deployed on CPUs or when processing a high volume of concurrent requests. The size of transformer models can lead to higher inference latency due to the large number of parameters and complex computations involved. For example, even if a model's inference time per tweet is only a few hundred milliseconds, such a delay may be unacceptable in systems that require sub-second responses.

F. Statistical Significance

All reported metrics, including the F1-scores, are based on single runs. Future research should incorporate repeated trials and statistical significance tests (e.g., McNemar's test, confidence intervals) to validate the reliability of the observed differences.

VII. ETHICAL AND SOCIETAL IMPLICATIONS

Automated cyberbullying detection systems have significant implications for online moderation. Potential issues include algorithmic bias arising from unrepresentative training data, which may lead to disparate false positive or false negative rates across demographic groups, and concerns regarding user privacy when continuous content monitoring is implemented. Furthermore, reliance solely on automated decisions risks penalizing legitimate discourse. Incorporating a human-in-the-loop framework, along with periodic bias audits and transparent appeal processes, is essential for mitigating these ethical risks.

VIII. FUTURE WORK

Future efforts may focus on lightweight transformer variants such as DistilBERT or apply model compression techniques (e.g., quantization, pruning) to BERT and RoBERTa to enhance scalability for real-time tasks. Investigating advanced large language models (e.g., GPT-4, LLaMA-3.1, Claude-3.5) and ensemble methods may further improve classification performance as measured by F1-score. Additionally, improvements in preprocessing (e.g., the incorporation of domain-specific dictionaries) and ethical frameworks (e.g., bias auditing, user privacy safeguards) are recommended. Cross-platform evaluations and rigorous significance testing will be critical in confirming the generalizability of the results.

IX. CONCLUSION

A comparative analysis was conducted on traditional ML models (Naïve Bayes, SVM, Random Forest) and transformer-based models (GPT-3.5, BERT, RoBERTa) for cyberbullying detection on 45,020 tweets. BERT and RoBERTa achieved high F1-scores ($\approx$ 0.87) and accuracies (87–88%), albeit with extended CPU training times ($\approx$ 13 hours). In contrast, Random Forest delivered a competitive performance ($\approx 85\%$

accuracy and 0.85 F1-score) in only 83 seconds, highlighting its efficiency for resource-constrained applications. GPT-3.5, employed in a zero-shot, batched configuration, attained significantly lower performance (25.41% accuracy, 0.2361 F1-score), indicating that further fine-tuning or prompt optimization is necessary to improve generative model outcomes. These results underscore the trade-offs between classification performance and computational cost, informing future research on optimization techniques such as model compression and fine-tuning for enhanced scalability and accuracy.

REFERENCES

[1] Comparitech staff. (2024, jan) Cyberbullying statistics and facts for 2024. Accessed 2024-01-22. [Online]. Available: https://www.comparitech.com/internet-providers/cyberbullying-statistics/

[2] NBC News. (2024, jan) New jersey family to sue school district after 12-year-old daughter's suicide. Accessed 2024-01-22. [Online]. Available: https://www.nbcnews.com/news/us-news/new-jersey-family-sue-school-district-after-12-year-old-n788506

[3] L. Korte. (2017) Youth suicide rates are rising. school and the internet may be to blame. USA TODAY. Accessed 2024-01-22. [Online]. Available: https://www.usatoday.com/story/news/nation-now/2017/05/30/youth-suicide-rates-rising-school-and-internet-may-blame/356539001/

[4] M. A. Al-Garadi and et al., "Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges," *IEEE Access*, vol. 7, pp. 70 701–70 718, 2019.

[5] V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," *International Journal of Artificial Intelligence and Applications*, vol. 3, no. 2, pp. 85–99, 2012.

[6] M. Fernandez-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real-world classification problems?" *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[7] M. T. Hasan, M. A. E. Hossain, M. S. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on deep-learning-based cyberbullying detection," *Future Internet*, vol. 15, no. 5, p. 114, 2023.

[8] L. M. Al-Harigy, H. A. Al-Nuaim, N. Moradpoor, and Z. Tan, "Building towards automated cyberbullying detection: A comparative analysis," *Computational Intelligence and Neuroscience*, vol. 2022, p. 4794227, 2022.

[9] C. Van Hee and et al., "Automatic detection of cyberbullying in social media text," *PLOS ONE*, vol. 13, no. 10, p. e0203794, 2018.

[10] N. A. Helal, A. Hassan, N. L. Badr, and Y. M. Afify, "A contextual-based approach for sarcasm detection," *Scientific Reports*, vol. 14, no. 1, p. 15415, 2024.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805, may 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Blog, 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pre-training approach," arXiv:1907.11692, jul 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[14] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained BERT model," in *2020 International Conference on Electronics & Sustainable Communication Systems (ICESC)*, 2020, pp. 1096–1100.

[15] S. Paul and S. S. Saha, "CyberBERT: BERT for cyberbullying identification," *Multimedia Systems*, vol. 28, no. 6, 2020.

[16] Y. Y. Tan, C.-O. Chow, J. Kanesan, J.-H. Chuah, and Y. Lim, "Sentiment analysis and sarcasm detection using deep multi-task learning," *Wireless Personal Communications*, vol. 129, no. 3, 2023.

[17] J. Wang, K. Fu, and C.-T. Lu, "SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 1699–1708.

[18] 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, no. 5, 2019.

[19] M. Dadvar, F. M. G. de Jong, R. Ordelman, and R. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop*, 2012, pp. 23–25. [Online]. Available: https://pure.knaw.nl/portal/en/publications/improved-cyberbullying-detection-using-gender-information

[20] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics*, 2012, pp. 656–666. [Online]. Available: https://aclanthology.org/N12-1084.pdf

[21] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," in *Proceedings of the 2017 ACM on Web Science Conference*, 2017, pp. 13–22.

[22] M. Abusaqer, M. B. Senouci, and K. Magel, "Twitter user sentiments analysis: Health system cyberattacks case study," in *Proc. 2023 Int. Conf. on Artificial Intelligence in Information and Communication (ICAIIC)*, 2023, pp. 18–24.

[23] H. Chen, L. Wu, J. Chen, W. Lu, and J. Ding, "A comparative study of automated legal text classification using random forests and deep learning," *Information Processing & Management*, vol. 59, no. 2, p. 102798, 2022.

[24] D. Chatzakou and et al., "Detecting cyberbullying and cyberaggression in social media," *ACM Transactions on the Web*, vol. 13, no. 3, pp. 1–51, 2019.

[25] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Advances in Information Retrieval*, ser. LNCS, vol. 10772. Springer, 2018, pp. 141–153.

[26] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2017, pp. 512–515. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14955

[27] U. Bretschneider, T. Wöhner, and R. Peters. (2014) Detecting online harassment in social networks. [Online]. Available: https://core.ac.uk/download/pdf/301363213.pdf

[28] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL Student Research Workshop*, 2016, pp. 88–93.

[29] Z. Waseem, "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter," in *Proceedings of the First Workshop on NLP and Computational Social Science*, 2016, pp. 138–142. [Online]. Available: https://aclanthology.org/W16-5618.pdf

[30] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text preprocessing for text mining in organizational research: Review and recommendations," *Organizational Research Methods*, vol. 25, no. 1, pp. 114–146, 2022.

[31] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, "Normalization techniques in training dnns: Methodology, analysis and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 173–10 196, 2023.

[32] E. Frank and R. Bouckaert, "Naive bayes for text classification with unbalanced classes," in *Knowledge Discovery in Databases: PKDD 2006*. Springer, 2006, pp. 503–510.

[33] S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine-learning-based algorithms for text classification," *Sustainable Operations and Computers*, vol. 3, pp. 238–248, 2022.

[34] C. Campbell and Y. Ying, *Learning with Support Vector Machines*. Springer Nature, 2022.

[35] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 50–60, 2017.