Enhancing Cyberbullying Detection Through Sarcasm Recognition and Advanced Text Preprocessing Using RoBERTa

Ángel M. Sánchez García
Master in Computer Science
Advisor: Nelliud Torres Batista, DBA
Polytechnic University of Puerto Rico
Graduate Project EXPO, May 2025

***Abstract* —** *Cyberbullying is an increasingly common problem in digital realms where offenders often use sarcastic language to disguise insult messages and mock the automated moderation system. This study looks at how sarcasm detection models can improve the detection of cyberbullying, which can identify abusive texts identified as sarcastic. The study suggests creating a thorough pre-processing of data, addressing the limitations reported in the literature review, such as text cleanliness and the use of advanced models. Techniques such as replacing abbreviations, removing emojis and emoticons, and filtering multilingual and nonsensical texts are applied. In addition, RoBERTa, a Transformer-based model, is used to detect sarcasm and cyberbullying due to its greater contextual understanding. The results aim to improve the detection of aggressive content hidden behind sarcasm, reduce false negatives, and strengthen content moderation systems. This approach helps to create a safer digital environment and reduces the psychological impact on victims of online harassment.*

***Keywords* —** *Cyberbullying, Deep Learning, RoBERTa, Sarcasm detection*

INTRODUCTION

Communication is how information is transmitted between two or more individuals using a shared instrument, whether verbal or written, to exchange ideas, thoughts, and emotions. This is a fundamental process in human interaction, fostering connection and mutual understanding between individuals [1]. These verbal or non-verbal messages can convey a broad spectrum of information, encompassing emotions and intentions [2]. Social media platforms have emerged as one of the main channels for long-distance communication, facilitating the exchange of messages almost instantaneously [3], [4]. Despite their ability to facilitate communication, they are also susceptible to being used as mechanisms of aggression and harassment [2], [5]. Cyberbullying, conceptualized as the use of technological tools to intimidate, harass, or humiliate one or more individuals, represents an emerging problem that poses challenges on social media platforms [5], [6], [7]. Therefore, social media platform moderators have turned to research in natural language processing (NLP) and deep learning models to identify abusive messages, thus minimizing human intervention automatically [3], [6], [7].

BACKGROUND AND MOTIVATION

As previously indicated, cyberbullying represents an emerging issue on digital platforms, manifesting in multiple forms [4]. These can range from explicit insults to subtle aggressions that may be disguised under the guise of humorous or ironic content [1], [2]. Contemporary research and studies indicate that victims of these cyberbullying incidents may experience psychological problems over time, such as social isolation, anxiety, and depression [4], [7]. Therefore, the creation and implementation of algorithms for detecting abusive content is essential in the moderation field [3], [5], [6]. However, written human communication presents considerable complexity [1], [8]. This phenomenon is because such communication is frequently conditioned by dynamics and context, which hinders these models from distinguishing between humorous comments and genuine harassment [1], [2]. One of the issues that has been the subject of study in this research is sarcasm [1], [2], [7]. Sarcasm constitutes a communicative expression that seeks to convey ironic irony to

offend or mistreat an individual [2]. These sarcastic comments can hide aggressive expressions under a humorous, ironic, and subtle appearance [1], [2]. Such communicative ambiguity results in NLP models being unable to identify patterns of cyberbullying adequately [7], [8]. This leads to many offensive messages that go undetected in content moderation on social media platforms [5], [6].

Sarcasm can also be conceptualized as communication in which the meaning aims to convey the opposite of what is articulated [2], [8]. Frequently, this humor can be used as a tool for verbal confrontation [2]. In the context of the interplay between cyberbullying and sarcastic communication, perpetrators can use sarcastic expressions with abusive content to ridicule or humiliate their victims without the need to resort to insulting terms [2], [7]. Research has shown that the coexistence of sarcasm can intensify the use of abusive language [1], [2]. A textual analysis on the Reddit platform revealed that user groups use sarcastic language to mask harassment, thus facilitating the evasion of detection systems and human moderators [2], [5]. The emerging issue of the absence of effective detection mechanisms requires attention, given that numerous comments categorized exclusively as sarcastic content can be justified by the aggressors as humorous content or misinterpretation [1], [7]. Therefore, the inability of moderation systems to identify this type of content could result in perpetrators persisting in harassing their victims under the pretext of irony [5], [7].

A review of the literature regarding the identification of cyberbullying and sarcasm using NLP and deep learning models reveals that contemporary approaches exhibit fundamental methodological discrepancies [3], [6], [7]. The absence of meticulous text analysis is fundamental for effectively training machine learning models [7], [8]. The research employs data cleaning methods; however, their application is limited to removing punctuation marks or special characters without considering other components that can generate noise in NLP models [7], [8]. Therefore, up to the time of the research, no study has implemented a pipeline that includes the replacement of abbreviations (for example, "u" → "you," "gonna" → "going to"), the removal of emojis and text faces that can alter the interpretation of a message, the filtering of meaningless words or typographical noise (for example, "asdsad," "lofasz"), the removal of sentences in other languages within multilingual sets to prevent bias, and the elimination of punctuation marks and symbols, considering that it can alter the syntactic understanding of sarcasm and, consequently, generate bias [7], [8].

Another revelation in the literature conducted so far indicates that most studies use machine learning models such as Support Vector Machine (SVM), Logistic Regression, BERT, hateBERT, and cyberBERT, among others, which have proven to be effective in classifying offensive text [4]–[7]. However, no study has explored the potential of RoBERTa in this task despite this model demonstrating superior performance in NLP when it comes to capturing context [8], [9].

RoBERTa (A Robustly Optimized BERT Pretraining Approach) is a language model developed by Facebook AI as an improvement of the BERT model, proposed with the purpose of improving contextual representation by modifying the structure and training of the model. This model eliminates the task of next-sentence prediction, integrates dynamic masking during training, and uses 160 GB of diverse textual data to capture complex linguistic nuances more accurately [10], [11]. While models like BERT work from the logical sequence between sentences, RoBERTa is designed to understand relationships at the contextual level without needing a linear narrative structure [10]. This feature makes it ideal for reading text snippets as they are represented on social media. Although its good accuracy in standard natural language processing tests is evident, its contribution to detecting cyberbullying hidden behind sarcastic expressions is almost unknown, which constitutes a gap this study aims to fill.

This study aims to investigate how sarcasm detection models can optimize cyberbullying detection models by identifying abusive texts that appear to be sarcastic, which may go unnoticed by cyberbullying models [7]. Despite advancements in the field of natural language processing (NLP), many contemporary cyberbullying models face challenges in identifying aggressions that are disguised as ironic texts, which limit the effectiveness of content moderation on social media [1], [2], [7]. This study suggests establishing a meticulous preprocessing process and using RoBERTa as a fundamental model for detecting cyberbullying characterized by sarcasm [8], [9]. The implementation of RoBERTa in these studies has been little explored, which offers the possibility of evaluating its feasibility in the current context [8], [9]. The suggested pipeline aims to optimize data quality through advanced preprocessing techniques. It incorporates the normalization of abbreviations, the eradication of textual emoticons, and the purification of texts in various languages [8], [9]. A comparison will be made between RoBERTa's performance in categorizing cyberbullying with sarcasm and other methods to evaluate whether it reduces bias in identification and optimizes the interpretation of linguistic and semantic indicators in texts of an aggressive nature [8], [9]. It is expected to contribute to the advancement of more precise tools for detecting cyberbullying in digital contexts [7], [9]. This approach will contribute to creating safer digital spaces and mitigating the psychological impact on victims [4], [5], [7].

While RoBERTa is designed to extract complex patterns from language, its base structure was not trained on data from informal environments such as social media, where abbreviations, emojis, or misspellings occur continuously. This type of content provides noise that can negatively contaminate the model's performance because, within its operation, the model interprets each token in isolation if they are not within the pretraining vocabulary [11], [12]. An example such as "going to" or "u" may have some meaning in the text that the model cannot capture or recognize adequately without a previous normalization phase. This is why this study introduces a preprocessing process that involves the normalization of texts.

PROBLEM STATEMENT

Cyberbullying is one of the relevant and growing problems within digital environments, directly affecting the mental and emotional health of millions of users around the world [4], [7]. Many aggressors use sarcasm to disguise offensive messages, making them difficult for automated content moderation systems to detect [1], [2], [7].

Research has shown that current models for detecting cyberbullying are limited in their ability to adequately identify offensive content when the text is expressed through sarcasm [1], [7]. This is because the studies conducted do not consider sarcasm as part of content moderation on social media [1], [2], [7].

Another significant shortcoming is the preprocessing of data used for training classification models. The analyzed studies have only used strategies partially, such as removing punctuation marks, URLs, or emojis [7], [8]. However, these studies do not combine these techniques by removing texts written in other languages, textual emoticons, meaningless words, substituting abbreviations with equivalent meanings, and correcting misspelled words [7], [8]. The lack of integrity of these techniques in data processing introduces noise into the data, which limits the models' ability to identify linguistic patterns associated with sarcasm and creates bias when classifying the text correctly [7], [8].

The literature explored so far has not analyzed the potential of advanced models like RoBERTa for classifying sarcastic texts as cyberbullying, even though RoBERTa has outperformed models like BERT and its variants, HateBERT and CyberBERT, in classification tasks [8], [9].

The lack of detection of potentially offensive texts is exacerbated when considering the findings of studies that demonstrate how groups or

collectives use coordinated patterns of sarcastic language to conceal systematic harassment [2], [5]. This harassment is scarily directed at vulnerable communities and generates adverse psychological effects on their victims [1], [4]. This lack of processing techniques and effectiveness in detecting sarcastic texts produces a high rate of false negatives, allowing aggressive comments disguised as sarcasm to go unnoticed by cyber content moderation systems [5], [7].

Therefore, the central question guiding this research is: How can the automatic detection of cyberbullying expressed through sarcasm be improved through comprehensive data preprocessing and advanced models like RoBERTa?

Answering the question proves fundamental for developing more effective protection systems in digital environments, reducing occurrences, and mitigating the negative impact on victims or communities on social media platforms [4], [7].

OBJECTIVE

The general objective of this project is to develop a better cyberbullying detection system where cyberbullying texts hidden under sarcasm can be presented. This is by designing a thorough pre-processing of data until the training of the advanced model, RoBERTa. To this end, work will be carried out in a pre-processing stage that will apply different techniques, elimination of punctuation marks, URLs, emojis, emoticons, symbols, texts in another language, nonsense words, as well as normalization through the substitution of abbreviations, to reduce noise without losing relevant context for the detection of sarcasm. Subsequently, RoBERTa's performance in tasks of classification of ironic texts and cyberbullying will be evaluated, its ability to detect cases of cyberbullying "disguised" with sarcasm, combining both approaches. Finally, very precise metrics will be used and validated to quantify the efficiency of the models with a special focus on the reduction of false negatives to improve the detection rate of cyberbullying hidden in sarcastic expressions.

METHODOLOGY

The methodology of this research involves developing a model for detecting sarcasm based on the RoBERTa algorithm and evaluating its integration with the cyberbullying detection model to improve its effectiveness. To this end, a pipeline of preprocessing, training, and evaluation of the models is developed.

Since in the data acquisition phase, two types of datasets will have to be handled. On the one hand, a dataset with data aimed at detecting sarcasm in a series of short social media texts, labeled as sarcasm (1) or non-sarcasm (0), with separate files for training and testing [13]. As for cyberbullying datasets, the first is a balanced set with about 100,000 examples and classified as non-cyberbullying (50,000) and cyberbullying related to race/ethnicity (17,000), gender/sexuality (17,000), and religion (16,000) [14]. In this way, it will be possible to train a RoBERTa model to detect cyberbullying created by users of different social networks accurately. The second cyberbullying dataset is based on the one by Wang et al. (2020) entitled "SOSNet: Fine-Grained Cyberbullying Detection using Graph Convolutional Networks", which is composed of social media posts with categories by type of bullying (gender, race, religion) and neutral cases [15]. This dataset will bolster the model's training with more examples, making it more capable of generalizability and robustness.

In this project, two different cyberbullying detection datasets will be used; both input datasets have multiple classification categories, such as race/ethnicity, gender/sexuality, and religion. These datasets will be joined to transform their tags into a binary classification: all cyberbullying categories will be unified as positive (cyberbullying), and neutral instances will be unified as negative (non-cyberbullying). In this way, a RoBERTa-based model for binary cyberbullying detection can be

trained, considering different examples and obtaining a good dataset diversity. This process of unification processing and conversion to binary applies to cyberbullying datasets, not to the sarcasm dataset since it has a binary classification structure (sarcasm and non-sarcasm) and is treated separately. Then, cleaning and normalization are done by removing special characters, URLs, mentions, hashtags, emojis, and emoticons. A normalization of character repetitions is also performed to avoid bias in training. An exploratory data analysis is executed such as class distribution, text length, elimination of null and duplicate data, filtering of data by language and elimination of texts with high numerical rate.

The training of both sarcasm and cyberbullying models is carried out in the following order: tokenization of text using RoBERTa tokenizer, division of the dataset between training and validation (80%, 20%), loading of the RoBERTa model with dropout regularization, AdamW optimization, and learning-rate adjustment. The models are trained with loss and accuracy monitoring, using early stopping and selecting the best model with the lowest validation loss.

For cyberbullying detection, the data is balanced with RandomUnderSampler, and RoBERTa is trained with a configuration like the sarcasm model. The evaluation of results includes verifying the accuracy and loss of the models, comparing metrics before and after adding sarcasm detection with the cyberbullying model, and qualitative analyses on how cyberbullying classification improves considering sarcasm detection.

Cyberbullying is an increasingly common problem in digital ecosystems, where aggressors use sarcasm to hide insults and avoid content moderation systems. This study analyzes how sarcasm detection improves cyberbullying detection in cases where malicious texts are perceived as sarcastic. A comprehensive preprocessing is proposed to address limitations mentioned in the literature, such as textual cleanliness, including the replacement of abbreviations, removal of emojis and emoticons, and filtering of multilingual and nonsensical texts. RoBERTa, a model based on Transformers, is used due to its effectiveness in contextual understanding for detecting both sarcasm and cyberbullying. The results aim to improve the detection of aggressive content hidden under sarcasm and reduce false negatives, contributing to a safer digital environment and mitigating the psychological impact on victims of harassment on social networks.

LIMITATIONS OF THE MODEL

One of the main limitations of this study is the dataset used to detect sarcasm. There is currently no specific public dataset for the detection of aggressive sarcasm, which prevented the model from being trained with several typical examples of this speech.

Instead, a general sarcasm dataset was constructed of texts that do not exclusively target an unpleasant or aggressive intent. As a result, the model tends to label harmless sarcastic texts as those with aggressive intent as sarcasm. This generalization can lead to misclassifications when the primary goal is to classify cyberbullying, as not all instances of sarcasm contribute to abusive language.

However, this limitation does not negate the benefit of having a sarcasm detection model on a cyberbullying classifier. On the contrary, the findings indicate that if a specifically aggressive sarcasm dataset is accessible, the model's performance could significantly increase in its training, focusing its detection capabilities on the most relevant cases for this context. In this way, the need for future research that builds and labels more specific data to improve the training of models for detecting aggressions disguised as sarcasm is reinforced.

In addition to the restriction due to the absence of a specific dataset to identify aggressive sarcasm, another important constraint relates to the nature of the texts in public cyberbullying datasets. Many of the instances belonging to these public datasets

mostly contain explicitly offensive language such as insults, obscenities, and extremely discriminatory phrases.

The graph titled "Most Common Words in Cyberbullying Texts" (see Figure 1) evidences this phenomenon, showing the most common words in texts labeled as cyberbullying. Among the most common insults are related to gender, religion, sexual orientation, race, etc. This excessive presence of explicit vocabulary creates inherent biases in the data, as the model may tend to learn that cyberbullying manifests itself only through insults or extreme words.

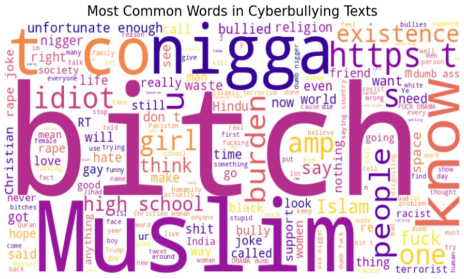

Figure 1
Most Common Words in Cyberbullying Texts

However, cyberbullying is not limited to clearly aggressive comments in real situations of cyberbullying. This can also be expressed in a passive-aggressive way, with sarcasm or implicit contempt without insulting directly. As public datasets lack instances with aggressive passive texts, it was necessary to remove texts with explicit insults to balance classes and reduce model bias, increasing greater generalizability in diverse scenarios.

This restriction demonstrates the need to develop future datasets containing cyberbullying without explicit language to train models that are more sensitive to the various forms of social media aggression.

Relationship between Sarcasm Detection and Cyberbullying

The graph titled "Relationship Between Sarcasm and Cyberbullying Detection" (see Figure 2) shows a scatter analysis between the probabilities that both models assign to 30 texts classified as aggressive sarcasm. The X-axis shows the probability that a text will be classified as sarcastic, and the Y-axis shows the probability that the exact text will be classified as harassment. The red line indicates the correlation of both variables.

The results show a slight negative trend, indicating that the higher the probability of sarcasm, the lower the probability of classification by the cyberbullying model. This trend confirms one of the main hypotheses of this study: that traditional models of cyberbullying often overlook hostile messages that use sarcasm.

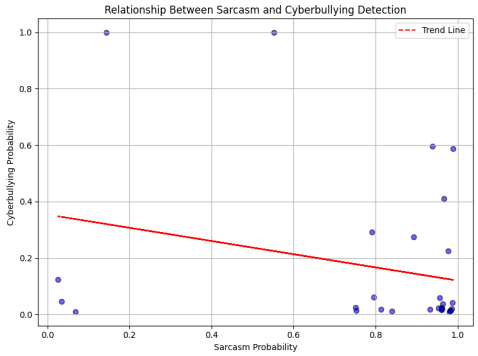

Figure 2
Relationship Between Sarcasm and Cyberbullying Detection

An important observation is the concentration of dots in the bottom right corner of the graph, where the texts have a high probability of being sarcasm but a low probability of cyberbullying. This pattern suggests that sarcasm, although it has an aggressive tendency, is mainly classified as cyberbullying if this model acts independently.

On the other hand, certain isolated points show high probabilities of cyberbullying without sarcasm, which confirms that there are also cases of more direct and explicit aggression. This points to the need to apply both models in a complementary way to contribute to expanding coverage and improving the detection of different forms of cyber abuse.

This graph shows the correlation between sarcasm and cyberbullying detected by isolated models and the importance of a joint approach for a more effective and contextualized cyberbullying

detection, especially when hidden or disguised under sarcasm.

Analysis of Results

The results demonstrate the effectiveness of combining a sarcasm detection model in a cyberbullying classification system, fulfilling the corresponding objectives of this study: to improve the ability of the models to detect aggressive expressions hidden under Sarcasm.

The RoBERTa model trained for detecting Sarcasm obtained an overall accuracy of 80%, with a balanced performance between Non-Sarcasm (f1-score: 0.80) and Sarcasm (f1-score: 0.79). These values demonstrate a good trade-off between the model's ability to correctly discover contrary sarcastic messages without misclassifying non-sarcastic messengers. The recall for the sarcastic class was 81%, which shows a high sensitivity of the model to detect these messages, reducing the number of sarcastic examples passed off as unnoticed by the system. This functionality is essential in content moderation environments, as irony can be used to hide texts with aggressive content.

On the other hand, the RoBERTa model trained for the detection of cyberbullying achieves an overall accuracy of 94%, with an f1-score of 0.94 for the Cyberbullying and Non-Cyberbullying classes. This indicator shows that the model not only correctly diagnoses bullying situations but does so consistently and with little margin for error. The good quality can be attributed to the rigorous process of cleaning, normalizing, and balancing the dataset, which allowed us to generalize the model and significantly reduce bias.

Cross-analysis between Sarcasm and Cyberbullying

The graph "Text Distribution by Sarcasm and Cyberbullying Classification" (see Figure 3) shows the distribution of 30 texts with aggressive sarcastic language according to their joint classification of both models. This cross-analysis makes it easier to identify how both models complement each other when classifying aggressive language under Sarcasm.

- Nineteen texts (63.3%) were classified as Sarcasm-only, showing a high presence of sarcastic texts that were not detected as aggressive.
- Two texts (6.7%) were classified as Cyberbullying Only, meaning that the models detected those texts as abusive, but not sarcastic.
- Two texts (6.7%) were classified as Sarcasm and cyberbullying, which says that both models classified these models as abusive and sarcastic.
- Seven texts (23.3%) were labeled as Neither (23.3%) and could be recognized as Sarcasm or cyberbullying; these texts may reflect high ambiguity or limitations of the models to identify these texts.

When added together, the texts classified as Only Sarcasm (19), Only Cyberbullying (2), and Sarcasm and Cyberbullying (2) give a total of 23 texts identified under the sarcasm and cyberbullying detection models, which represent 76.7% of the total sample. This means that if both models are complementary, they can effectively detect three-quarters of texts with abusive Sarcasm, significantly improving content moderation systems.

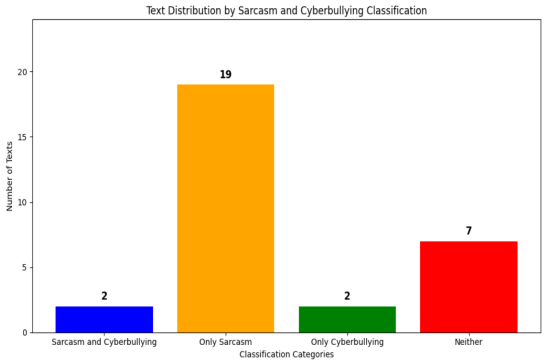

Figure 3
Text Distribution by Sarcasm and Cyberbullying Classification

CONCLUSION

The analysis of results strongly confirms the main hypothesis of this research: the use of a specialized model in the detection of sarcasm considerably increases the effectiveness of cyberbullying classification systems. Hybrid architecture models, based on the combination of models for sarcasm and cyberbullying, effectively demonstrated a reduction in false negatives, especially in cases where offensive texts are hidden within sarcasm.

A key factor in these results was the use of the RoBERTa model, which matched 94% accuracy in the cyberbullying classification task and 80% in the detection of sarcasm. These metrics demonstrate the ability to understand the linguistic context and capture passive-aggressive or indirect nuances in offensive forms of communication. Given its Transformer-based design and understanding of context, RoBERTa proved relevant to traditional approaches and models, making it a truly powerful instrument for sensitive tasks such as moderation automation.

In addition, the cross-analysis between both models, shown in the text distribution graph, shows that 76.7% of a sample of 30 texts with aggressive sarcasm were correctly classified by at least one of the models. This figure reinforces the importance of using both models together since their combination facilitates the more accurate detection of insulting expressions that could go undetected by online content moderation systems.

The results also showed that texts with a high presence of sarcasm were not identified as aggressive by the cyberbullying model, confirming the need to combine both approaches to address the hostile language complex on digital platforms.

Overall, the solution proposed in this study based on RoBERTa and reinforced by a strict preprocessing pipeline is a great step in developing fairer and more precise moderation systems. This project not only increases the accuracy and sensitivity in the detection of cyberbullying but will also serve as a methodological precedent for future research aimed at generating safer and more resilient digital environments against forms of verbal violence.

REFERENCES

[1] S. Frenda, Sarcasm and Implicitness in Abusive Language Detection: A Multilingual Perspective, Ph.D. dissertation, PRHLT Research Center, Universitat Politècnica de València & Department of Computer Science, University of Turin, Italy, Jun. 2022. Available: https://riunet.upv.es/handle/10251/197947.

[2] S. Frenda, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, and P. Rosso, "The Unbearable Hurtfulness of Sarcasm," Expert Systems with Applications, vol. 193, 2022. DOI: https://doi.org/10.1016/j.eswa.2021.116398.

[3] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, "Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques," Electronics, vol. 10, no. 22, 2021. DOI: https://doi.org/10.3390/electronics10222810.

[4] A. Muneer and S. M. Fati, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter," Future Internet, vol. 12, no. 11, 2020. DOI: https://doi.org/10.3390/fi12110187.

[5] A. Perera and P. Fernando, "Accurate Cyberbullying Detection and Prevention on Social Media," Procedia Computer Science, vol. 181, pp. 605–611, 2021. DOI: https://doi.org/10.1016/j.procs.2021.01.207.

[6] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT," Information, vol. 14, no. 8, 2023. DOI: https://doi.org/10.3390/info14080467.

[7] X. Guo and S. Gauch, "Using Sarcasm to Improve Cyberbullying Detection," in Proceedings of the TRAC 2024 Workshop on Trolling, Aggression and Cyberbullying, ELRA, pp. 52–59, 2024. Available: https://aclanthology.org/2024.trac-1.7/.

[8] D. Šandor and M. Bagić Babac, "Sarcasm Detection in Online Comments Using Machine Learning," in *Information Discovery and Delivery*, vol. 52, no. 2, pp. 213–226, 2024. DOI: https://doi.org/10.1108/IDD-01-2023-0002.

[9] V. S. S. Settipalli and N. M. K. Dasireddy, *Reducing Unintended Bias in Text Classification Using Multitask Learning*, Master's thesis, Blekinge Institute of Technology, Karlskrona, Sweden, Jan. 2021.

[10] T. Dhamija, Anjum, and R. Katarya, "Comparative Analysis of Machine Learning and Deep Algorithms for Detection of Online Hate Speech," in *Proceedings of the*

International Conference on Machine Learning and Data Science (ICMLDS), New Delhi, India, 2021.

[11] GeeksforGeeks. (2023). *Overview of RoBERTa Model* [Online] Available: https://www.geeksforgeeks.org/overview-of-roberta-model/.

[12] Hugging Face. (n. d.). *RoBERTa* [Online] Available: https://huggingface.co/docs/transformers/model_doc/roberta.

[13] A. Muhammad. (n. d.). *Sarcasm Detection Dataset* [Online] Available: https://github.com/muhammadadyl/SarcasmDetection/tree/master/data.

[14] M. Ahmadinejad, N. Shahriar, and L. Fan. (2024). *Cyberbullying Detection Dataset* [Online] Available: https://github.com/muhammadadyl/Cyberbully-Detection-Dataset.

[15] J. Feng and L. C. Yu, "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection," in *Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020)*, IEEE, 2020, pp. 1968–1977. DOI: https://doi.org/10.1109/BigData50022.2020.9378065.