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

Author: Ángel M. Sánchez García
Advisor: Nelliud Torres Batista
Electrical & Computer Engineering and Computer Science Department

Abstract

Cyberbullying is an increasingly common problem in the digital space, where offenders often use sarcastic language to disguise insulting messages and mock the automated moderation system. This study examines how sarcasm detection models can improve cyberbullying detection, which can identify abusive texts identified as sarcastic. The study suggests thoroughly preprocessing the data, addressing the limitations reported in the literature review, such as text cleaning and the use of advanced models. Techniques such as replacing abbreviations, removing emojis and emoticons, and filtering multilingual and meaningless texts are applied. In addition, RoBERTa, a Transformer-based model, is employed to identify cyberbullying and sarcasm due to its superior contextual comprehension. The objectives are to enhance the identification of hostile material hiding inside sarcasm, minimize false negatives, and strengthen content moderation systems.

Introduction

Communication is the way in which 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 wide spectrum of information, encompassing emotions and intentions [2]. Social media platforms have become one of the main channels of long-distance communication, facilitating the exchange of messages almost instantly. Despite their ability to facilitate communication, they are also susceptible to being used as mechanisms of aggression and harassment [2]. Cyberbullying, conceptualized as the use of technological tools to intimidate, harass, or humiliate one or more people, represents an emerging problem that poses challenges on social media platforms [3].

Background

Cyberbullying represents an emerging problem on digital platforms that manifests itself in multiple ways. These can range from explicit insults to subtle attacks that can be disguised as 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 [3].

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]. This communicative ambiguity results in NLP models being unable to adequately identify patterns of cyberbullying [3].

A textual analysis on the Reddit platform revealed that user groups use sarcastic language to mask harassment, thus facilitating evasion of detection systems and human moderators [2]. The inability of moderation systems to identify this type of content could cause perpetrators to persist in harassing their victims under the pretext of irony [3].

The literature review regarding cyberbullying and sarcasm identification reveals fundamental methodological discrepancies [3]. Current approaches lack meticulous text analysis and comprehensive data cleaning methods, which are limited to removing punctuation marks or special characters without considering other noise-generating components in NLP models [3]. Additionally, while models like BERT have proven effective, RoBERTa's potential for capturing complex contextual relationships remains unexplored in this specific task [4].

Problem

Cyberbullying is a growing problem in digital environments that affects the mental health of users around the world [3]. Attackers often use sarcasm to disguise offensive messages, evading automated moderation systems [1], [2], [3]. Current cyberbullying detection models have difficulty identifying sarcastic offensive content [1], [3], and researchers have not explored advanced models such as RoBERTa for this specific task [4]. This research addresses how the detection of cyberbullying expressed through sarcasm can be improved through comprehensive data preprocessing and advanced models such as RoBERTa.

Methodology

This research methodology 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. Two types of datasets are utilized: a sarcasm dataset with binary labels (sarcasm/non-sarcasm) and cyberbullying datasets containing multiple classification categories, including race/ethnicity, gender/sexuality, and religion [5].

For data preprocessing, the following steps are implemented:

- Combining cyberbullying datasets and converting to binary classification
- Cleaning and normalization by removing special characters, URLs, mentions, hashtags, emojis, and emoticons
- Normalizing character repetitions to avoid bias
- Filtering data by language and eliminating texts with high numerical content
- Removing null and duplicate entries

The model training process follows this sequence:

- Tokenization of text using RoBERTa tokenizer
- Division of datasets between training and validation (80%/20%)
- Loading of RoBERTa with dropout regularization
- Implementation of AdamW optimization with learning-rate adjustment
- Training with loss and accuracy monitoring
- Application of early stopping techniques
- Selection of the best model based on lowest validation loss

For the cyberbullying detection model specifically, the data is balanced using RandomUnderSampler before training RoBERTa with a similar configuration to the sarcasm model. The evaluation includes verification of accuracy and loss metrics, comparison of performance before and after integrating sarcasm detection with the cyberbullying model, and qualitative analyses of how cyberbullying classification improves when considering sarcasm detection.

Results and Discussion

The results demonstrate the effectiveness of combining a sarcasm detection model in a cyberbullying classification system, fulfilling the study's objective to improve detection of aggressive expressions hidden under sarcasm.

The RoBERTa model trained for sarcasm detection achieved 80% accuracy, with balanced performance between non-sarcasm (f1-score: 0.80) and sarcasm (f1-score: 0.79). The recall for the sarcastic class was 81%, showing high sensitivity in detecting sarcastic messages. Meanwhile, the cyberbullying detection model achieved 94% accuracy with an f1-score of 0.94 for both cyberbullying and non-cyberbullying classes.

The relationship between sarcasm and cyberbullying detection (Figure 1) reveals a negative trend, confirming that traditional cyberbullying models often overlook hostile messages using sarcasm. The scatter analysis of 30 texts classified as aggressive sarcasm shows that higher probability of sarcasm correlates with lower probability of being classified as cyberbullying by traditional models.

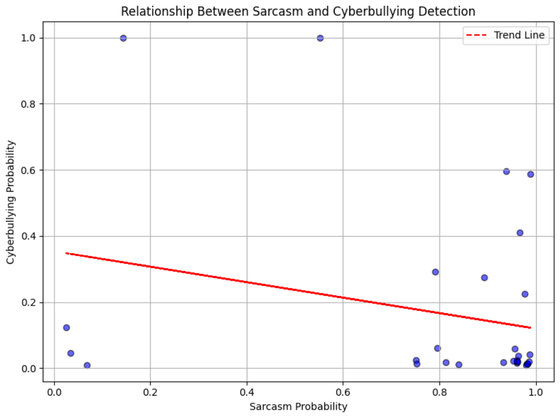

Figure 1
Relationship Between Sarcasm and Cyberbullying Detection

The distribution analysis (Figure 3) of these texts demonstrates how both models complement each other:

- 63.3% were classified as sarcasm-only
- 6.7% were classified as cyberbullying-only
- 6.7% were classified as both sarcasm and cyberbullying
- 23.3% were not recognized by either model

When combined, both models identified 76.7% of texts with abusive sarcasm, significantly improving content moderation capabilities. This cross-analysis reinforces the importance of using both models together to detect insulting expressions that might otherwise go unnoticed.

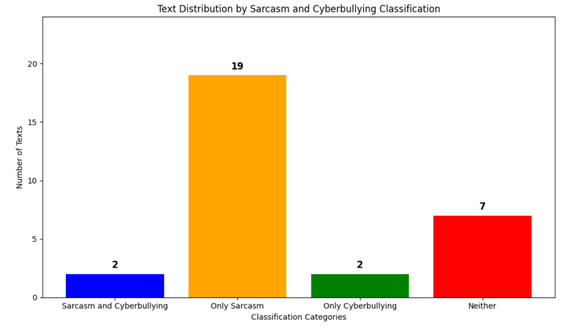

Figure 2
Text Distribution by Sarcasm and Cyberbullying Detection

Conclusions

Integrating a specialized sarcasm detection model significantly improves the effectiveness of cyberbullying classification systems. The combined approach reduced false negatives in cases where abusive language was masked by sarcasm and demonstrated outstanding results: RoBERTa achieved 94% accuracy in identifying cyberbullying and 80% in recognizing sarcasm. The analysis revealed that 76.7% of the hostile sarcastic messages were correctly identified, evidencing the complementary value of both models. While the isolated cyberbullying model often misses highly sarcastic texts, our RoBERTa-based approach, powered by sophisticated preprocessing, represents a significant step toward more accurate content moderation systems and establishes a solid foundation for future developments in this specialized field.

Future Work

Despite promising results, the main limitation is the lack of specific datasets with examples of aggressive sarcasm. This limits the models' ability to capture patterns of insulting sarcasm. Future work will focus on creating a specialized database by collecting and manually classifying examples processed by the sarcasm model, allowing continuous improvement through learning from mistakes. Additionally, integrating functionality to highlight decisive words or expressions in classification would enhance system transparency and facilitate human oversight when justification of automated decisions is required.

Acknowledgements

I sincerely thank Dr. Nelliud Torres Batista for his guidance throughout this research. I also thank the Polytechnic University of Puerto Rico and the Department of Electrical Engineering and Computer Science for their support. Special thanks to my partner, family, friends, and colleagues for their encouragement, and to the contributors of public datasets and NLP tools used in this study.

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. https://riunet.upv.es/handle/10251/197947 2022.

[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. https://doi.org/10.1016/j.eswa.2021.116398

[3] 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. https://aclanthology.org/2024.trac-1.7/ 52–59, 2024.

[4] 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.

[5] 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. https://doi.org/10.1109/BigData50022.2020.9378065