Spring 4-2024

Reducing Bias in Cyberbullying Detection with Advanced LLMs and Transformer Models

Dahana Moz Ruiz

Annaliese Watson

Anjana Manikandan

Zachary Gordon

Reducing Bias in Cyberbullying Detection with Advanced LLMs and Transformer Models

Dahana Moz Ruiz, Annaliese Watson, Anjana Manikandan, Zachary Gordon
Dr. Yulia Kumar, Dr. J.Jenny Li, Patricia Morreale
Department of Computer Science and Technology, Kean University, Union, NJ 07083

Introduction

This paper delved into a comprehensive exploration of the inherent biases present in Large Language Models (LLMs) and various Transformer models, with a focus on their role in identifying and addressing instances of cyberbullying. The objective was to refine and enhance the accuracy and fairness of these models by mitigating the biases deeply ingrained in their structures. This was crucial because language models could inadvertently perpetuate and amplify existing biases present in the data they were trained on.

Methods and Materials

The foundation of this study rested on empirical data meticulously collected from 'X' (formerly Twitter), where cyberbullying instances were systematically classified into several categories including Age, Ethnicity, Gender, Religion, and Others, alongside instances where no cyberbullying was observed. This comprehensive dataset allowed for thorough analysis. A sophisticated cyberbullying detection application was developed, leveraging the advanced capabilities of the OpenAI API as its backbone. This application served as a filtering mechanism, issuing alerts for content identified as inappropriate, thus fostering a safer online environment.

Future Work

Future plans involve integrating multimodal data, such as images, videos, and audio, into cyberbullying detection models to capture nuanced forms of online harassment and enhance detection accuracy. Additionally, there are plans to conduct user studies and feedback sessions to understand user preferences and requirements for cyberbullying detection tools, with the goal of developing more user-friendly and effective solutions.

Results

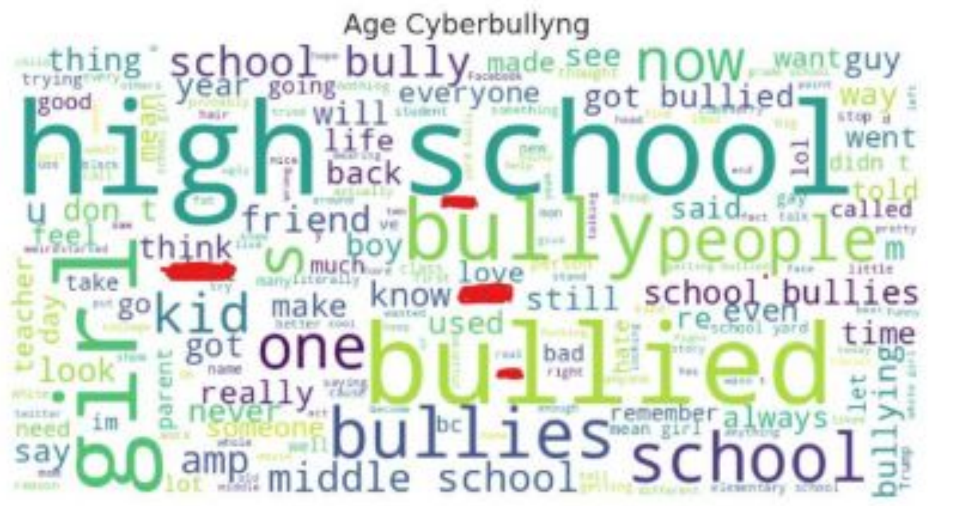


Figure 1: Word cloud of the Age Cyberbullying Dataset

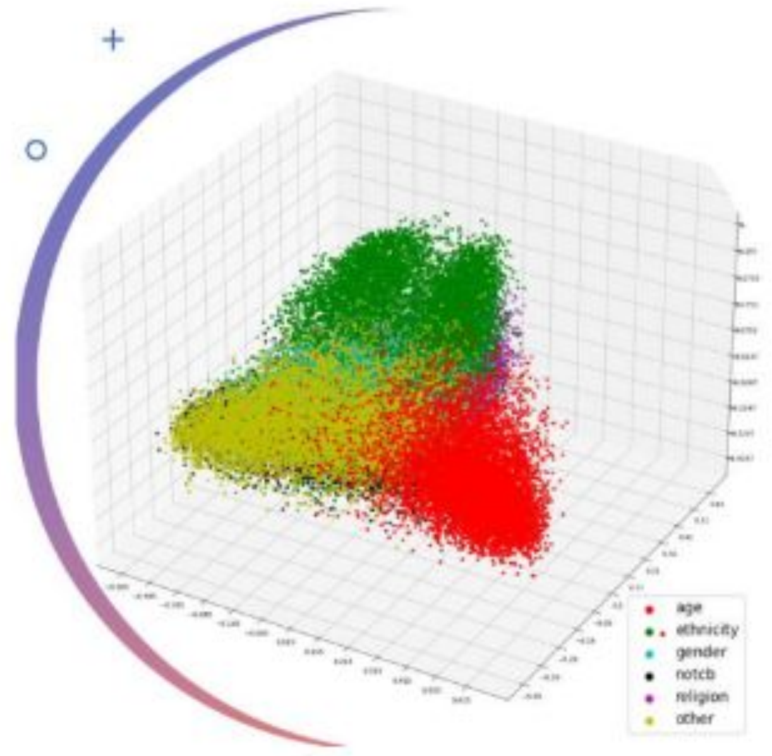


Figure 2: Data clustering by the Sentence transformer.

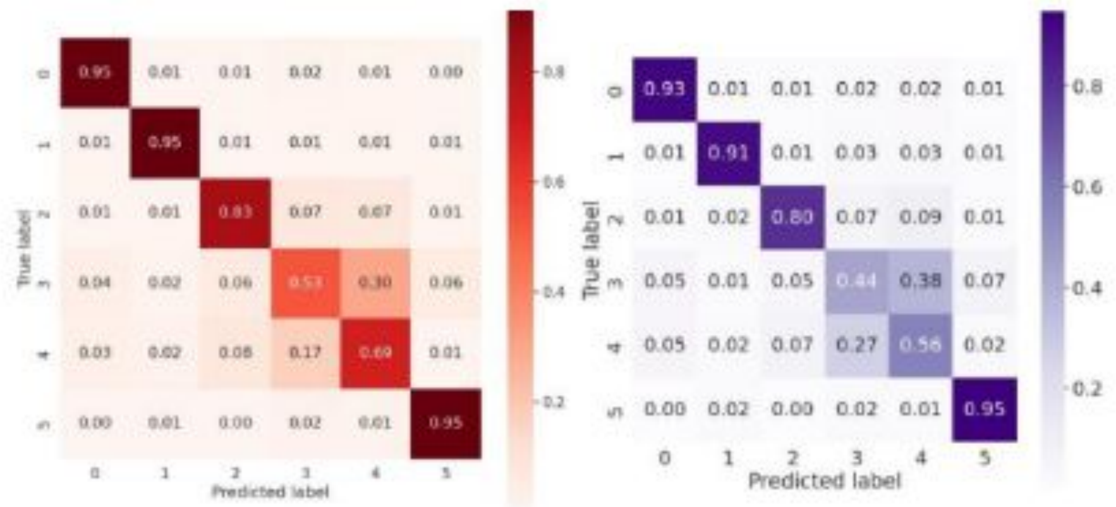


Figure 3: Results of Logistic Regression (left) vs Random Forest classification (right)

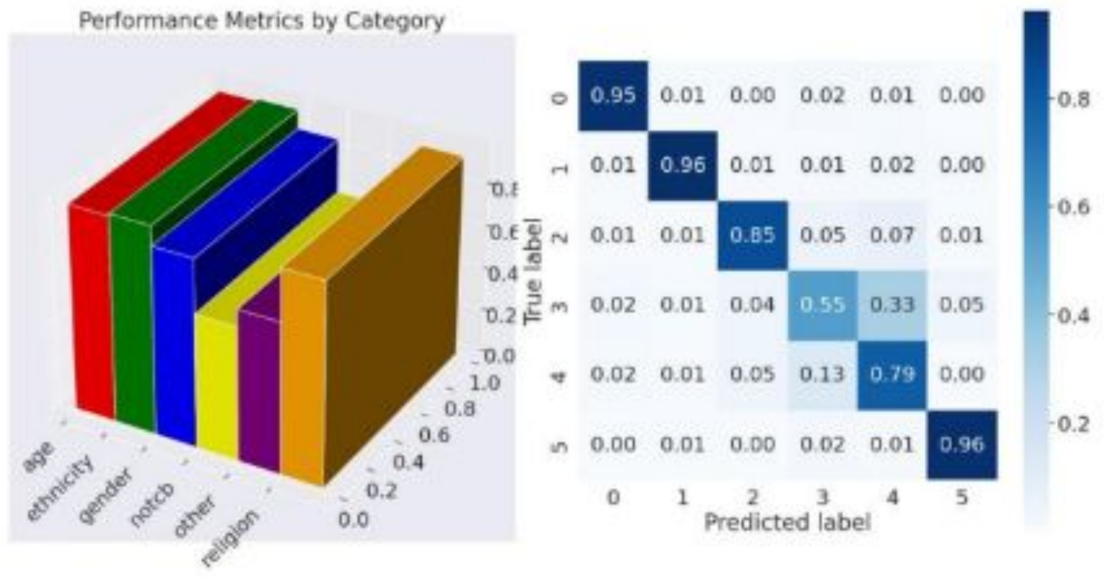
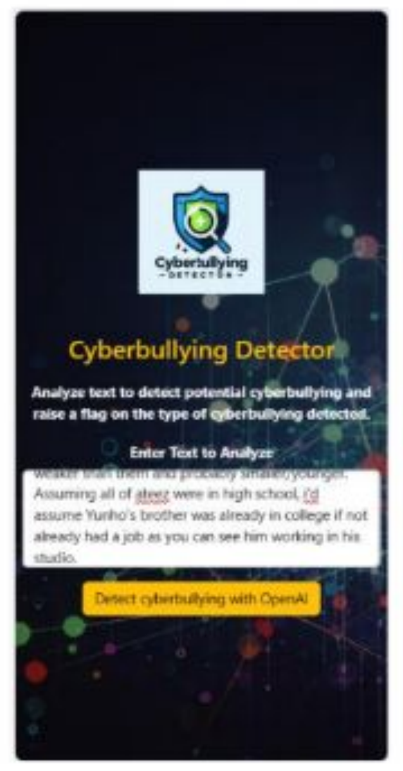


Figure 4: Results of Support Vector Machine: By Category (left) vs Confusion Matrix (right)



Cyberbullying Detector

Analyze text to detect potential cyberbullying and raise a flag on the type of cyberbullying detected.

Enter Text to Analyze

weaker than them and probably smaller/younger. Assuming all of ateez were in high school, i'd assume Yunho's brother was already in college if not already had a job as you can see him working in his studio.

Detect cyberbullying with OpenAI

Figure 5: GUI of the cyberbullying Detection App. [Images generated by DALL- E 3].

Conclusions

The imperative to address the well-documented dangers of cyberbullying necessitates effective and unbiased detection systems. This study illuminates the critical importance of mitigating inherent biases in Large Language Models (LLMs) and Transformer models utilized for cyberbullying detection. By delving into the inner workings of these models, we uncover how they inadvertently perpetuate and potentially amplify biases ingrained in their training data. Such biases can result in the unjust flagging of content or the overlooking of genuine cyberbullying incidents. Our cyberbullying detection application, constructed on the robust foundation of the OpenAI API, serves as a testament to the progress achievable in this arena. Not only does it identify inappropriate content, but it also cultivates a more respectful online environment through emoji-driven feedback mechanisms. While our research sheds light on these issues, it represents just one step in the ongoing journey toward crafting a safer digital landscape. The findings from this study and our ongoing efforts can provide a solid groundwork for others in the field. As technology continues to evolve, so too does the landscape of online interactions, underscoring the importance of remaining proactive in ensuring that our digital communities remain respectful and inclusive.

Acknowledgement

This work was supported by funding from AI4ALL & L.E.A.P

References