Cyberbullying Detection Utilizing Artificial Intelligence and Machine Learning

Annaliese Watson (watsanna@kean.edu), **Dahana Moz Ruiz** (mozruizd@kean.edu), **Aysha Gardner** (gardneay@kean.edu),
Dr. Y. Kumar, J. Jenny Li (faculty advisors) **and Cody Glickman** (AI4ALL)
Department of Computer science and Technology, Kean University, Union, NJ

Introduction

Technology is a tool that can be used to gain knowledge and for advancements in areas like medicine, machinery, and everyday tasks. It can be used to connect with friends, work from home, and to improve quality of life. But some social media users can use it to hurt others. Cyberbullying is a major issue that has been steadily growing over the past few years. Cyberbullying has also steadily increased the rates of stress, anxiety, depression, violent behavior, low self-esteem and may cause suicide. Cyberbullying is an ongoing problem for social media users, and it is urgent that a solution is presented. The role of Artificial Intelligence (AI) in Cyberbullying detection (CD) is currently vital. Millions of messages, photos and other sources that include hurtful, abusive or threatening content are being transmitted over the Internet daily. We use Python programming language and Google BERT Transformer to classify the text message into two categories: cyberbullying (*cb*) vs not. Our dataset consists of nearly 5000 words and phrases of both categories. The Neural Network model currently catches the damaging content aka *cb* category before it is sent with 93% accuracy. The model can further be tuned, its accuracy improved.

Methods and Materials

We started our work in Fall 2022 during our initial AI4ALL training where our team was formed. We researched on the issue, collected our dataset and applied more straightforward methods of *cb* detection to touchbase on this topic and our dataset. Currently we are using Deep Learning, in particular Google's Transformer BERT for cyberbullying text classification. The model helps us to balance our dataset, that has more words with Cyberbullying rather than without it. We run our code on Colab PRO+ using our Google drive as a data and output storage. We applied Nadam optimizer to further improve our results. first train the model and then validate it with the test dataset. The importance of the topic in the current 'post-remote' environment is hard to unerestimate, therefore, we want to continue with the topic, present it varios conferences and raise the awareness of the issues and our work.

Results

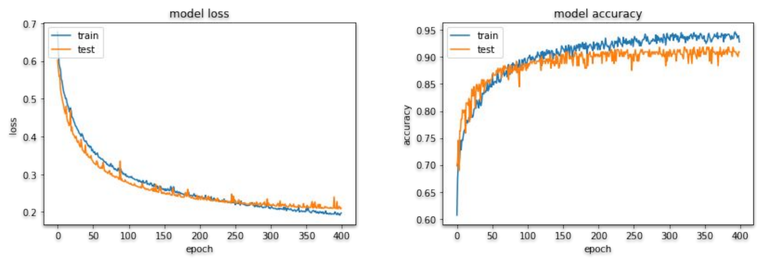

Fig. 1. The train loss (left) vs train accuracy (right) (15th epoch)

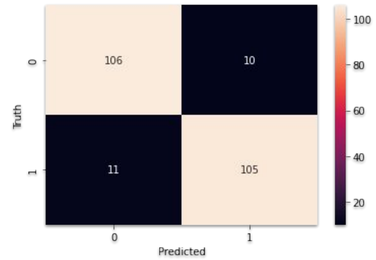

Fig. 2. The Confusion Matrix

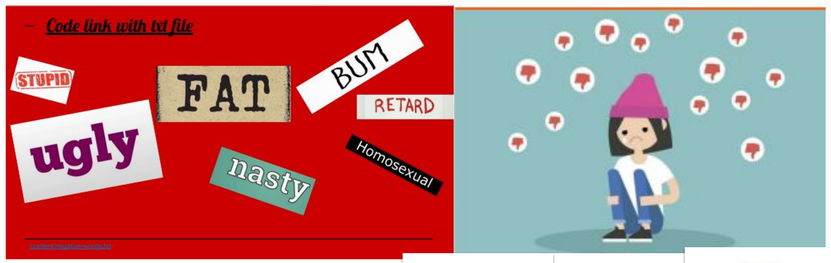

Fig. 3. Hugging Face, BERT and AI stop CyberBullying

Conclusions

In the future, we plan to develop a tool that will prevent social media users from sending harmful content. We will send a warning to the user attempting to do so and with the help of the embedded in the process AI recommender system we will suggest the sender on how to rephrase their text avoiding offensive language. Another plan we have includes implementing an unsupervised clustering technique to visualize the frequency of words used in a particular time period to help the researchers better analyze high risk vs low risk words and biases with the help of visualization. Fortunately, technologies, AI and Machine Learning (ML) has been so rapidly improving over several recent years, that there are many ways to lessen the effects of cyberbullying and online verbal abuse we can further explore. We will use these technologies to help block harmful words, insults, and derogatory phrases. In the same manner that we can type in google and find all images of a cat, we can use AI to help to identify trigger words and block them. We are aware of the currently popular ChatGPT, but its use for the matter has its advantages and disadvantages as machines do not have emotions and words may have many different interpretations. Further study is needed and ongoing.

Acknowledgement

This work was supported by funding from AI4ALL program.

References

[1] Bozyiğit, A., Utku, S., & Nasibov, E. (2021, April 30). ScienceDirect. Retrieved October 22, 2022, from https://www-sciencedirect-com.kean.idm.oclc.org/science/article/pii/S0957417421004425

[2] Muhammad, A. P., Zheng, J., Irfan, N. R., Mohammed, A. M., & Tehseen, Z. (2021, March 19). Abusive language detection from social media comments using conventional machine learning and deep learning approaches. EBSCO. https://link-springer-com.kean.idm.oclc.org/article/10.1007/s00530-021-00784-8

[3] Classify text with BERT https://www.tensorflow.org/text/tutorials/classify_text_with_bert

[4] Hugging Face Text classification https://huggingface.co/docs/transformers/tasks/sequence_classification