CS 334
Dr. Ho
10/22/23

# Project Proposal

**Title:** Utilizing ML Models in order to detect and flag messages sent by users that may indicate cyberbullying or cyber harassment

**Group Members:** Andrew Chung, James Park

**Description of Problem:** Social media and chat messaging apps have become increasingly popular and a part of the daily lives of adults, teens, and children alike. Applications such as Discord, Messages, Whatsapp are a very popular and convenient way to chat and communicate with peers and colleagues. However, messaging and social media have also become a place where people can harass and bully people through demeaning comments, leading the victims to face anxiety, depression, and even suicidal thoughts. Our group is determined to find a way to filter out harmful comments before they reach the user in an efficient and accurate manner.

**Description of the Dataset:** The dataset "Cyberbullying Classification" from Kaggle is a dataset of 46,017 unique tweet texts scraped from twitter.com. The dataset originates from J. Wang, K. Fu, and C.T. Lu's paper "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection." There are two main categories, the first column being the tweet itself, and the second column being the classification of the tweet as not cyberbullying or the type of cyberbullying: age, religion, gender, ethnicity, other cyberbullying. The tweets were classified by humans. For this project, we will need to first preprocess the text data, as it is currently untokenized and some contain unnecessary parts such as the usernames or other outlying punctuation. We will also need to change the types to just be two types, cyberbullying and not cyberbullying, as we are just interested in classifying messages as either or.

**References/Work Done so Far:** Previously, researchers have used different types of classification and information retrieval algorithms, such as semantic analysis, the Genetic algorithm (a method for solving both constrained and unconstrained optimization problems that is based on natural selection), Fuzzy rule-based system (a mathematical tool for dealing with the uncertainty and the imprecision), and Fuzzy logic (an approach to computing based on "degrees of truth" rather than the usual "true or false" Boolean logic on which the modern computer is based). A specific example of this is "Online Social Network Bullying Detection Using Intelligence Techniques" by B. Sri Nandhini and J.I. Sheeba, where Nandhini and Sheeba proposed a framework of data preprocessing, feature extraction of certain word groups (like nouns, adjectives, pronouns), the FuzGen learning algorithm, and the Naïve classifier technique to detect the presence of cyberbullying activity in social networks, such as spring.me and myspace.com.

Furthermore, there have been additional studies looking deeper into this problem: NOM and NORM features, a count and normalization of bad words from nosewaring.com, were used to detect cyberbullying on social networks, but this only resulted in a 58.5% accuracy. This shows that there have been trials where researchers have looked into ML algorithms to detect cyberbullying, but this can always be improved. The applications and diction from these training datasets are outdated (as they use text from less modern websites and the older generation). A more modern approach to this problem can result in a more accurate, interesting, and presentable finding that can help decrease the presence of cyberbullying.

**Description of Tentative Plan:** The first objective is to preprocess the data in the form that is most beneficial to us. We will first replace the different types of cyberbullying to simply cyberbullying, making the column binary. We will also have to preprocess the text data in order to make it standardized to a certain degree. After the data is preprocessed, we will utilize three ML models, Recurrent Neural Network, Decision Tree Classifier, and the transformer neural network. With this we will also utilize various wrapper methods, filter methods, and embedding methods to find the most effective feature selection method for these different models.

After splitting the data into test and training as well as training our various models, we will compare accuracies of the predictions of each model using various methods and confidence intervals. The model's performance will be judged on the accuracy of their predictions put against each other.

**Citations for work:**

Raj, Mitushi, et al. "An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques." *SN Computer Science*, U.S. National Library of Medicine, 2022, www.ncbi.nlm.nih.gov/pmc/articles/PMC9321314/.

Desai, Aditya, et al. *Cyber Bullying Detection on Social Media Using Machine Learning*, www.itm-conferences.org/articles/itmconf/pdf/2021/05/itmconf_icacc2021_03038.pdf. Accessed 24 Oct. 2023.

B. Sri Nandhini a, et al. "Online Social Network Bullying Detection Using Intelligence Techniques." *Procedia Computer Science*, Elsevier, 25 Mar. 2015, www.sciencedirect.com/science/article/pii/S187705091500321X.

**https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification**