Cyberbullying Detection using Support Vector Machine (SVM) Algorithm

# Problem Opportunity

Social media allows people to create and share information through the use of virtual communities. Through these sites, people can express their feelings, opinions and emotions. However, as people have fully embraced these media for socializing and communicating, it has become an avenue for cyber bullying, which is described as a deliberate act that is conducted through the use of digital media to hurt someone in terms of abusive text messages or images. For victims, such behavior can lead to depression and other severe life-threatening problems. As the scope of cyberbullying becomes larger, it has become a major concern in the educational field as well. In 2015, a child-care nonprofit Stairway Foundation conducted a survey among 1,268 school children aged 7 to 12 and 1,143 aged 13 to 16 in the National Capital Region; Silang, Cavite; Zamboanga Sibugay; Bayawan City, and Bacolod City, Negros; Cebu City; and Tiu, Batangas. The study found that eighty percent of teenagers aged 13 to 16 have been cyberbullied through social media. Students who are victims of cyberbullying face various academic and social problems. The impact of cyberbullying ranges from withdrawal from school activities, school absence, and school failure, to eating disorders, substance abuse, depression, and even suicide. Aside from students, teachers are also vulnerable to cyberbullying attacks. In UK, the National Association of Schoolmasters Union of Women Teachers (NASUWT) conducted a survey among school teachers. The study found that 42% said they had been a victim of cyberbullying. Among the participants, 61.2% said they had been subjected to a pupil writing an insulting comment about them on a social network or internet site, 38.1% said a student had made comments about their competence or performance as a teacher, and 9.1% said they had faced allegations that they behaved inappropriately with pupils. Qualitative evidence gathered through a survey of teachers has demonstrated that cyberbullying affects the working lives of staff and impacts severely on staff motivation, job satisfaction and teaching practice.

Thus there is an extreme need to monitor and detect harmful online activities. However, given the massive amount of information available online, it is impossible to monitor all the activities in the cyberspace. To address this issue, several studies have focused on the application of Natural Language Processing to automate the process of detecting cyberbullying.

Natural Language Processing bridges a gap between human and computer interaction by providing a way for computers to analyze, understand, and derive meaning from human language. Through the use of NLP, system developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, text classification, sentiment analysis, speech recognition, and topic segmentation. Given the set of tasks NLP can perform, most studies focused on text classification process in the detection of cyberbullying occurrences. Text classification (also known as text categorization) is the task of assigning predefined categories to text documents. It involves the use of machine learning methods to learn automatic classification rules based on human-labeled training documents. For the past years, it has been given an utmost attention by the researchers. Furthermore, the tasks involved in text classification including the machine learning approaches such as Naïve Bayes, Decision Tree, K-nearest neighbor (KNN), Support Vector Machines (SVM) and Convolutional Neural Network have been extensively studied.

Several studies tackled the area of cyberbullying using text-classification. In 2012, Dadvar, Jong, Ordeiman, and Trieschnigg conducted a study on Improved Cyberbullying Detection using Gender Information. The team believes that developing gender-specific features would lead to more accurate classification of harmful contents. In their study, they used a supervised learning approach to detect occurrences of cyberbullying; moreover, they created a Support Vector Machine classifier using Weka. As for their dataset, they gathered posts from MySpace then compared the most frequently used foul words by each gender through the use of Wilcoxon signed rank test. For their baseline, the researchers used four types of features: profane words, second person pronouns, other pronouns, and the TFIDF value of all the words in each post.

Van Hee et al. (2015) conducted a research on Automatic Detection and Prevention of Cyberbullying. The team presented the construction and annotation of a corpus of Dutch social media posts annotated with fine-grained text categories, such as insults, threats, sexual talk, defamation, defense, and curse. The participants in a cyberbullying context were also identified in order to enhance the analysis of human interactions involving cyberbullying. Initially, the researchers had decided to use this particular research paper as their main basis for creating the project; however, the process of manually annotating the statements within the dataset, according to the aforementioned fine-grained text categories, proved to be difficult as some of the categories were closely related to each other. Additionally, the succeeding methods after the data annotation process proved to be difficult to comprehend given the current knowledge the researchers possess under the NLP field.

In 2016, Cheng and Ng conducted a research at De La Salle University. The research aimed towards detecting cyberbullying roles through textual context in Facebook and Twitter. First, the researchers identified six roles in a cyberbullying context: the bully, victim, assistants of the bully, reinforcers, outsiders, and defenders. Among the three algorithms used by the researchers such as Naïve Bayes classifiers, decision trees and Support Vector Machine (SVM), the SVM had the highest accuracy. The optimal model produced an accuracy of 59.7% in detecting the bullying roles; while detecting the bully role produced an accuracy of 80.9%. The researchers are currently using this study as their basis in the creation of their proposed cyberbullying detection model because unlike the other study (as mentioned before), they found this paper easier to comprehend. It gave them a clear picture of what they should do in order to achieve their desired output. Additionally, since SVM has been proven to be the most accurate model, the researchers were also planning to use SVM in automating the detection of cyberbullying occurrences.

Thus, the previous studies served as a baseline for the proponents to conduct a new cyberbullying detection system that is based on Philippine context.

## Proposed Solution

The ultimate goal of this project to create a cyberbullying detection model using Support Vector Machine (SVM) algorithm and integrate it with a web application that will serve as a small social networking site wherein users can post anything on their timeline. Moreover, the said web application will be developed by the proponents for testing purposes only. The system will prove beneficial to the school as it aims to eliminate the current manual method of monitoring cyberbullying in social media posts among students and school faculties. Moreover, the model can detect harmful posts written in English and Tagalog. Once the system detects a cyberbullying scenario, it will generate a report to the school administrator regarding the information about the post and the bully. Thus, the sanction that will be given to the harasser will be based on the student’s handbook of Asia Pacific College.