II. Review of Related Literature

**Related Studies**

Dinakar et al. (2011) created a model for the detection of textual cyberbullying. The dataset for their study was obtained from YouTube for comments posted on videos through the use of YouTube PHP API. The data were grouped into clusters of sexuality, race and culture, and intelligence. The datasets for each cluser were divided into 50% training, 30% validation and 20% test data. Each dataset was subjected to three operations: removal of stop-words, stemming and removal of unimportant sequence of characters. Then they select and populate feature space for three supervised learning methods along with a Naive Bayes classifier: JRip, J48, and SVM. At the first part of their experimentation phase, binary classifiers were trained on each three datasets for each of the labels, namely, sexuality, intelligence, and race and culture to predict if a given instance is classified into its respective label. At the second phase, the three datasets were combined to form a new dataset for the purpose of training a multiclass classifier. Lastly, the trained models were evaluated through the use of a kappa statistic. Although JRip yield the highest accuracy, its kappa values were lesser compared to SVM. Thus, SVM's high kappa values implies better reliability for all labels.

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.

In 2015, Kansara and Shekokar proposed a framework for detecting negative online interactions in terms of abusive contents carried through posts or comments as well as images. They believe that the combination of text and image analysis techniques can yield an efficient result for detecting potential risks of cyberbullying. The framework aims to detect abusive image or text and block it immediately before it can be disseminated to the cyberspace. The designed framework has two modules: abusive image detection and abusive text detection. The process of abusive image detection begins from feature extraction wherein the Local Binary Pattern (LBP) will be used to detect and describe interest points of an image. The extracted features using LBP will be mapped to the existing visual word in vocabulary. The event of certain visual words provides powerful hints for the presence of offensive content in an image. Finally, SVM is used for classification thereafter. Given a set of training images to the classifier, each image marked as belonging to abusive class if given image contain abusive or pornographic contents. SVM training algorithm builds a model that assigns new image into abusive categories by applying the learned rules to identify abusive images. As for the process of abusive text detection, the system will perform pre-processing of the text messages and bag of words is applied for extracting the features of the message. After the feature extraction, the matrix is generated which is used by the Naive Bayes model to categorize abusive text messages. At the final stage, the Boolean system is used to categorize cyberbullying or non-cyberbullying event by analyzing the result obtained by both the image and text classification.

Marathe and Shirsat (2015) proposed a mechanism that can automatically identify videos and users promoting cyberbullying, using a set of discriminatory features and classification algorithms. The proposed solution is a multi-step process primarily consists of three phases: training and testing profiles collection, dynamic model building, and an implementation based on Naive Bayes algorithms. In the first phase, the researchers collected positive training dataset (which contains occurrences of cyberbullying). Through the use of YouTube API, they were able to download the availble meta-data of several relevant videos. Furthermore, the meta-data will be extracted to build the training set. In the second phase, they use character n-gram based approach to build a dynamic model from these training profiles. In phase 3, they built a system based on Naive Bayes algorithm. It is based on Bayes rule for text classifications. It takes one video as an input, finds an extent of textual similarity between this video metadata and the training data. Based on the probability score, a video can be classified as relevant cyberbullying promoting or irrelevant.

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.

Sugandhi, Pande, Agrawal, and Bhagat (2016) proposed a system for automatic monitoring and prevention of cyberbullying through the use of machine learning. The data was collected from Twitter through the use of Twitter API while the labeled training data was gathered from ChatCoder. The collected data is then preprocessed and passed on to the classifier. The team tested the accuracies of various classification algorithms (Naïve Bayes, Support Vector Machine, and KNN) in the detection of cyberbullying on their training data. Among the three algorithms used, SVM was said to be the most consistent and yields the highest accuracy. The sentiment of the statement is calculated in parallel with the SVM classification; moreover, the sentiment analysis system employs a method in which it assigns polarity values to each statement based on a certain formula. The multiclass SVM takes the bullying data and classifies it into three different classes namely low, medium, and high depending on its harmfulness level. Lastly, once the post is put into its respective class, a response grading system implemented by the researchers is executed. The system will give response based on the class in which the post will be categorized: low level post will result in a popup in the form of a reflective user interface while a high level post will result into a temporary ban.

Sintaha et al. (2016) proposed a system for cyberbullying detection using sentiment analysis in social media. The data was gathered from Twitter through the use of Twitter API. For the data pre-processing, the team corrected the spelling mistakes in the tweets, converted the uppercase letters to lowercase order, remove usernames, URLs, and unnecessary white spaces, and emoticons are replaced with the corresponding word that defines the emoticon to find out polarity of the tweets. In their dataset, they use 80% of the dataset for training and 20% for testing. The team also used an automated training set classifier wherein they can collect thousands of tweets and run an algorithm through those tweets to classify whether the word is positive or negative. For their baseline, they used four types of features: stop words, repeating letters, punctuation, and words with alphabets at the beginning. Furthermore, the dataset were retrieved from database, classified and run through three machine learning classifying techniques (SVM, Naïve Bayes Classifier, and Convolutional Neural Network) to compare the performance of those algorithms. After the classifier has been trained, the test tweets were run through the classifier to detect the polarity and after it has been detected, it was used to compare the accuracy of the classifiers. Among the different approaches, SVM generate the highest accuracy followed by Naïve Bayes.