**Automated Detection of Cyberbullying Occurrences in Social Media Posts Through Text Classification Using Support Vector Machine (SVM) Algorithm**

**Submitted by:**

Samantha Mallari | Faith Ballesteros | Eva Samillano

(SCSPROJ – SS141)

**ABSTRACT**

Social media is defined as an electronic form of communication wherein people can create, share, and exchange information in the virtual community. Nowadays, it has significantly increased the communication platforms. Consequently, as the social networking domain expands in the cyberspace, it inevitably creates more opportunities for cyber bullies to oppress internet users. Cyberbullying is defined as an aggressive behavior in the cyberspace. It involves repeatedly making threats, sending provocative insults or racial slurs, bashing, and sending of spam messages. Recent studies indicate that cyberbullying has become a pervasive problem around the world and it is tremendously alarming. However, given the massive information on the Web, there is a need for intelligent systems to identify potential risks automatically. This paper presents the creation of a cyberbullying detection model. The initial step in creating the model is to gather data for the corpus by crawling a subset of YouTube, Facebook, and Twitter through the use of Import.Io. The team was able to extract 2000 statements from these sites. The data annotation process involves the creation of three datasets from the corpus: Cyberbullying, Non-cyberbullying, and Ambiguous Cyberbullying. To annotate the data from the pre-defined classes, the team distributed questionnaires for 100 respondents. The team members were also required to annotate the data as well. Pre-processing involves tokenization, stripping of whitespaces and removal of unnecessary characters. This was done through the use of Notepad++ and Excel. For the feature extraction, the team used Bag-of-Words (BoW) model, specifically a unigram model, to represent the vocabulary of words. The corpus will be further subdivided into two subsets: training set and testing set. 75% of the data will be used for the training set while 25% will be used for the testing set. The team employed Support Vector Machine (SVM) algorithm in WEKA as a classification tool. Lastly, the team was able to determine the accuracy of the cyberbullying detection model through the use of precision, recall and f-score.

Keywords:*Cyberbullying, Detection, Implications, Social Media*

**INTRODUCTION**

* 1. **Background of the Problem**

Long before men evolved into species of higher intellectual capabilities, bullying was believed to have been evident. Boehm (2012) stated in his book, Moral Origins, that primates, specifically monkeys and chimpanzees, frequently execute bullying-like deportment against members of their own kind. The said behavior would, in turn, provide them an edge in terms of social stature, acquired resources, and reproductive "opportunities" among the rest. Upon the rise of the Homo-sapiens (the genus into which humans of today are classified), the purpose of bullying was redefined from social dominance to a mere destructive act. Hogan Sherrow, an anthropologist, believes that "the ability of language to facilitate communications, coordinate behaviors, and express thoughts and gossip has completely altered the form and intensity of bullying". Fast-forward to the 21st century, likewise known as the era of widespread technological advancements, a new form of bullying emerges - cyberbullying. Cyberbullying is referred to as "modern-day bullying". For any ill-treatment to be considered as a form of cyberbullying, it should meet the following criteria: involuntary – the offensive action happened deliberately or intentionally; repetitive – the mistreatment has been reportedly known to be occurring recursively; harmful – the deed has brought upon negative feedback toward a particular person (or groups of people), and has utilized technology as his/her medium for accomplishing the said feat (e.g. through text messages, instant messages, emails, and the like). Altogether, they give meaning to the term cyberbullying as the “willful and repeated harm inflicted through the use of computers, cellphones, and other electronic devices”.

With the immense number of new gadgets being introduced into the market almost every year and the accessibility of acquiring a reliable internet connection, the probability of people engaging in different social media websites, forums, blogs or other forms of social communities online are not likely to decrease. Similar scenarios apply to the Philippines. A survey entitled, "Southeast Asia Digital in 2015" which was conducted by the people behind "We Are Social", a global agency dedicated to delivering world-class ideas with forward-thinking brands, indicated that the Philippines ranked 5th out of the 11 countries in Southeast Asia in terms of social media usage (based on the number of active Filipino social media users). Consequently, it leads to the formation of virtual “hang-outs” of some sort. And whenever groups of people are involved, specifically in areas where admin or moderator supervision is limited, the occurrence of cyberbullying becomes inevitable. The alarming fact about cyberbullying is that it can be done by anyone (including people whom the victim is not familiar with), in an instant, and may spread across different areas, harming a person without other people’s knowledge.

From being dubbed as the “Texting Capital of the World” to “Social Media Capital”, the Philippines had proven itself enough to be recognized as an overly social country. As of January 2016, the aforementioned global agency ("We Are Social") reported in their annual digital, social, and mobile statistics that the number of active social media users in the Philippines amounts to 48 million. While the existence of these particular types of media provided ample benefits with regard to improving former communication-related processes, such sites have likewise been considered as the launch-point of common cyberbullying assaults occurring within the country. According to a 2015 survey administered by a child-care nonprofit Stairway Foundation Inc, 80% of Filipinos have been cyberbullied through social media. Even celebrities were known to have been targets of cyberbullying attacks as well. Recently, a radio DJ, Karen Bordador has experienced extensive cyberbullying, following her arrest with her boyfriend in a drug-related buy bust operation.

In order to mitigate severe cases of cyberbullying in social media, the Republic Act 10627, also known as the Anti-Bullying Act of 2013, was introduced. It recognizes cyberbullying (as one of the types of bullying inclusive in the said law) as a major offense, specifically when elementary and secondary students are the people involved, and provides appropriate provisions on the consequences of their actions. This means that the law was mainly focused on school-related cyberbullying occurrences - those that took place between classmates regardless of whether it happened inside or outside the campus. However, Camarines Sur Rep. Rolando Andaya Jr. noticed that the scope of the said act (particularly with regard to cyberbullying) remained inefficient. Instead, he proposed a bill (known as House Bill 5718 - Anti-Cyberbullying Act of 2015) which hopes to extend the definition of cyberbullying (in terms of the people that will be affected) and its respective countermeasures. In spite of the fact that improvements in the mitigation of such incidents can possibly be presented by the bill, it was yet to be approved by country's lawmakers as an official law. Oddly enough, despite the dangers cyberbullying can inflict on an individual, only a small number of reports are continuously being submitted voluntarily to designated authorities. Dr. Ryan Guinaran, Ph.D. claimed that the latter was due to the fact that cyberbullying in the Philippines (in comparison to other countries) tends to be more on a conservative level. If Filipinos continue to practice this type of passive attitude regarding the matter at hand, then even with the efforts granted by the government and NGOs alike, cyberbullying will still persist. Thus, instead of waiting for the parties involved to voluntarily explain their side to the people concerned, the group had the thought of taking advantage of the same platform where the aforementioned event was known to have been rampant – technology – as a countermeasure to cyberbullying.

* 1. **Statement of the Problem**

How can the automation of detecting cyberbullying occurrences in public social media posts be made possible with the aid of the concept of text classification?

**1.3 Objectives**

**Main Objective**

This research aims to formulate a cyberbullying detection model.

**Specific Objectives**

* To acquire data for the corpus
* To apply text pre-processing methods on the dataset
* To extract features from the corpus
* To perform experiments in WEKA toolkit
* To evaluate the model’s accuracy

**1.4 Significance**

The creation of a cyberbullying detection model (which will be patterned according to selected cyberbullying statements found in social media posts bearing sensitive issues as perceived by the many) will greatly contribute to the improvement of social media monitoring. As of today’s time, online moderators have been utilizing the manual way of flagging offensive posts in social media sites (Van Hee, 2015). In 2001, the Children’s Internet Protection Act (CIPA) was enacted to address concerns on children’s access to visual offensive content over Internet. To comply with CIPA requirements, administrators of social media often manually review online contents to detect and delete offensive materials. In Japan, the Parent Teacher Association (PTA) performed website monitoring called “net-patrol”. In this method, once a harmful post was detected by a net-patrol member, he will immediately report it to the administrator. However, despite the efforts made by the authorities, it is unattainable to monitor activities occurring in the cyberspace from time to time. Most importantly, the previous methods were labor intensive, time consuming and not scalable in reality (Vinita, 2014). With the cyberbullying detection model that can be integrated to social media sites, the process of detecting harmful entries online will be automated. Furthermore, vicious posts will be swiftly and easily flagged and subjected for analysis (by the moderators) without rendering the moderators to keep an eye out for such statements in the site 24/7.

In 2016, a survey that was conducted by We are Social found that 47% of the Filipinos are active social media users. As the number of social media users tremendously increases, it consequently intensifies the cyberbullying problem (Chen et al., 2012). Another survey that was conducted by child-care non profit conclude that 80% of Filipino teenagers are victims of cyberbullying. Indeed, it has become rampant in the Philippines (Cheng, 2016). Social networking sites possess notable characteristics that make it an indulging tool for cyberbullies. These characteristics include the following: real time updating, wide spread dissemination of personal information, a rallying point for people, anonymity, instantaneousness, ability to reach large audience, cheap, connected to power dynamics, information posted in social media has a tendency go viral, it allows other social network users to generate comments, it builds up and can generate support from others, and it is very empowering (Gonzales, 2014). Nowadays, social media sites are beginning to adapt to an easier, user-friendly approach to reduce and possibly eliminate cyberbullying (White, 2012). Facebook provides tools and resources that will help their user to protect their account and report offensive content. In the privacy setting, the user can specify privacy for a specific message or post and limit how much information can be visible to others. Users can also report offensive posts and categorize which way the post is harmful to them to determine the issue and its magnitude. As for Facebook pages, administrators can set up a keyword moderation blocklist and enable a profanity blocklist that filters posts and comments by users into their page. Once the user includes a blacklisted keyword in their post, it will be automatically identified as spam. On the other hand, YouTube provides reporting tool wherein a user can flag inappropriate videos with hate content, nudity or graphic violence and report abusive comments. As a result to this, YouTube will verify if a certain video or post violates their terms of use before removing it permanently from their site. However, these methods imposed by social networking require users to take actions such as reporting posts or enabling plug in before they can execute appropriate action for it. Most people, typically Filipinos, are reluctant to admit to being victims of cyberbullying (Andrade, 2012). One of the possible reasons would be because they do not want to further instigate a conflict on the opposing party and that they thought that what the bully had done is not that much of a big deal (NCPC, 2007). Despite how much or how long the effect of the statement dwells on the person, at that point in time when the victim reads it, he or she will still get affected by it one way or another (Bersola-Babao, 2012). Therefore, even if there are only traces of cyberbullying occurrences present, it is still encouraged by the experts not to turn a blind eye over such statements. In order to address these issues, a cyberbullying detection model will be designed to detect even subtle posts implying cyberbullying attacks as much as possible. Thus, the model can be integrated by the system developers in order to be fully functional before it can automatically extract harmful information from the Web. And since the team based the model in the Philippine context, it can detect offensive posts written in Tagalog and English offensive posts.

One of the beneficiaries of this research are the system developers. They can implement the model to an automatic cyberbullying detection system, which can be further integrated to social networking sites to detect any events of cyberbullying in the cyberspace. Through the use of this model, the system can detect posts as long as it notices potential cyberbullying activity - even minor ones as much as possible since the model can generate an accuracy of at least 70-80% in terms of detecting cyberbullying occurrences present in social media posts.

The findings of this study will redound to the benefit of researchers who want to explore the field of both Cyberbullying and Natural Language Processing (NLP) techniques, specifically, in Text Classification. Since NLP is a broad field of study, the team will merely focus on the creation of cyberbullying model that will automate the process of detecting harmful entries in social networking sites through text classification using Support Vector Machine. As a result to this, the study can help researchers gain a better understanding on the processes of text classification and the incorporation of the model with Linear Support Vector Machine Algorithm. As for the researchers who want to explore the field of cyberbullying, this study can further enhance their knowledge on what cyberbullying is, the classification of cyberbullying and non-cyberbullying events, and the different categories of cyberbullying, based on sensitive issues in the Philippines.

The younger population tend to visit social media websites more frequently than the rest (Cheng, C. & Ng, L. 2016). A research that was conducted by National Crime Prevention Council (NCPC) concludes that teens ages 13 to 17 are an online population. Over 80 percent of teens use their phone regularly, making it the most popular form of technology and a common medium for cyber bullying (Cyberbullying Research Center, 2015). A survey that was conducted in the US, with a sample size of 935 teens with age ranging from 12 to 17 years old, found that 4 out of 10 teens are victims of cyberbullying (Lenhart, A. 2007). Adolescents have a tendency to deal with things impetuously due to their immaturity. They are most likely unable to identify the intensity of the damage that they had done until it finally occurred (Li, Q. 2006). As online platforms are increasingly used for cyberbullying, it poses a threat to teenager’s mental and physical well-being (Price, M. 2010). Thus, it can lead to depression, low self-esteem, poor academic performance, self-harm, and suicide (Hinduja, S. 2010). However, once the cyberbullying detection model is integrated into social networking sites, such incidents may be prevented before they get out-of-hand.

Although parents are vigilant about protecting their children from the content of sites and poses limits on the amount of time spent online, teens report shows that they are largely unsupervised by their parents online (NCPC, 2007). Recent survey shows that 73 percent of the parents keep the home computer in an open family area—either purposefully or inadvertently providing at least casual surveillance of the online activities of youth at home (Lenhart, Madden, & Hitlin, 2005). Other research has determined that 54 percent of parents use some type of Internet filter, 62 percent check up on the Web sites their children visit, and 64 percent have specified rules for the time their children spend online (Lenhart et al., 2005). However, despite these efforts made by the parents, teenagers can easily find a way to visit objectionable Web sites or participate in inappropriate online behavior (Corwin, 2008). In addition to this, an Internet Safety Coordinator from Illinois, Jace Galloway states that relying solely on parental control inside the house is insufficient because children can access the Internet from various locations. By integrating the cyberbullying detection model in social networking sites, it can help them monitor the different activities of their children in the cyberspace.

**1.5 Scope and Limitations**

This research aims to develop a cyberbullying detection model having an accuracy of at least 70-80% in terms of detecting cyberbullying occurrences in public social media posts expressed using the Filipino language (Tagalog and English), based on the context as to how they are typically comprehended and/or stated by Filipinos residing in Metro Manila.

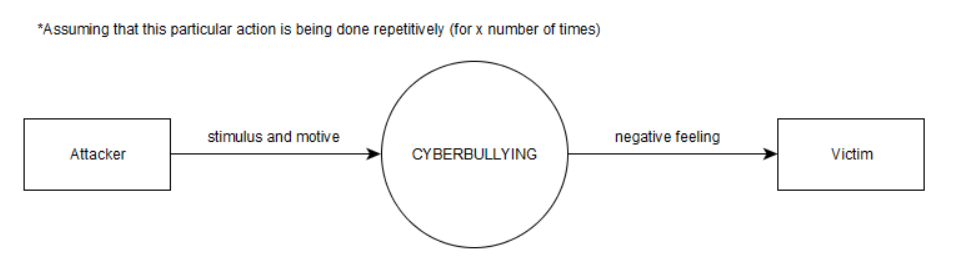
The model's initial capacity to distinguish cyberbullying statements from non-cyberbullying ones is based on the corpus (dataset) created by the members of the group themselves. It currently consists of 2000 statements which were obtained from either Facebook and Twitter posts or Youtube comments. The totality of these statements pertain to the major controversial issues in the Philippines (e.g. those involving the LGBT community, drugs, scandals and other major issues of famous people - celebrities, political entities, and the like)

Text preprocessing methods such as the cleaning of the dataset and annotation were implemented on the entirety of the dataset. The cleaning process began by removing all special characters, non-readable text (e.g. asdfghjkl), emoticons, links, and characters belonging to various foreign countries' writing systems. Text annotation was performed by initially identifying the 3 annotation schemes - cyberbullying, not cyberbullying, and ambiguous cyberbullying - to be used per statement in the dataset. Among the 2000 statements used, 1000 were annotated by each member of the group while the other 1000 were distributed among a sample of Metro Manila residents - 100 to be exact - with varying ages and occupations. About 90% of the selected respondents were college students while the rest of the 10% were comprised of immediate family members and other relatives. These sets of annotations will be combined and then utilized in the succeeding processes.

In this phase, the entire dataset was subdivided into 2 for the purpose of determining the constituents for both training and testing data. 75% of the corpus was used as training data (1500 statements) while the rest of the 25% was used as testing data (500 statements). In the feature extraction process, the Bag-of-Words model (unigram) was used for both data classifications. Features represented every unique instance of a particular word in all of the statements contained within the corpus. The number of occurrences for each feature in a given statement was likewise evaluated. These values would then be used as inputs for the WEKA toolkit to process.

The sole experiment that was performed for the purpose of determining the accuracy of the model involved the use of the Support Vector Machine (SVM) algorithm on the 2000 statements that had undergone various processes. Both the experimentation and subsequent evaluation of the model took place in the WEKA toolkit. The results that were procured by the two processes were based on the numerical data - obtained from the processing of the features in the preceding stages. The aforementioned data was subsequently fed unto the tool. An evaluation procedure (in WEKA) known as the Cross Validation Folds 10 was used in order for the team to determine the accuracy of the constructed cyberbullying detection model. It was read based on the quantitative values presented such as Precision, Recall, F-measure, and Kappa Statistic.

The members of the group had decided that the cyberbullying detection model will remain intact with WEKA toolkit and will be presented that way towards the panel for this particular term.

**1.6 Context Diagram**

**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 2014, Lam et al. conducted a research on classifying typhoon related tweets. In their study, they categorized typhoon related tweets as: resource coordination, urgent rescue needed, urgent rescue solution, damage reporting, and media storm coverage. For their dataset, they gathered 2,356 tweets. They used Bag of Words with TF-IDF weighting scheme for their data representation. Furthermore, these data were classified using Support Vector Machine and Naive Bayes classification. Ten-fold cross validation was used to evaluate the classifiers. Results show that the SVM classifier performed better with an F-score of 88.7% and a kappa statistic of 81.7% than the Naive Bayes classifier with 77.3% and 62.6% respectively.

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.

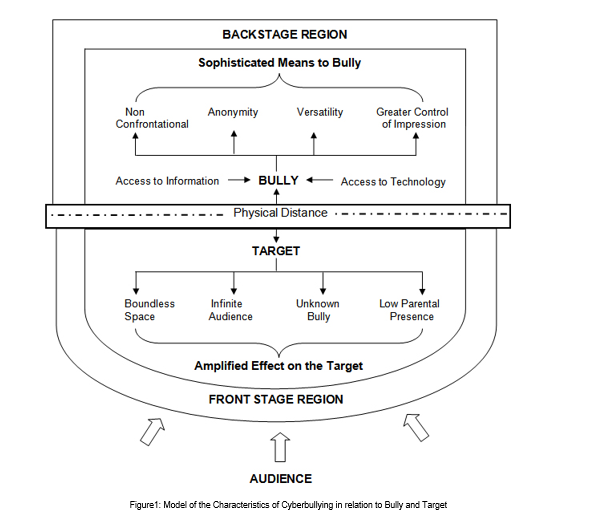
Sintaha et al. (2016) proposed a system for cyberbullying detection using sentiment analysis in social media. The data was gathered from Twitter through the use of Twitter API. For the data pre-processing, the team corrected the spelling mistakes in the tweets, converted the uppercase letters to lowercase 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.

**THEORETICAL FRAMEWORK**

In his book “The Presentation of Self in Everyday Life”, Ervin Goffman introduced the mechanisms of audience segregation. He describes how people play different roles in different situations. It is a mechanism wherein an individual perform roles, in order to create a favourable image of themselves and leave a good impression to others that is linked to the role they perform. The role that the individual performs is based on who their audience is.

Nowadays, more and more people are getting inclined to social networking sites because it provides an easier way for social interactions and communications. These sites allow users to share personal information about themselves through text, pictures, and other forms of media which in turn, creates an image for each user; however, the representation of oneself in the cyberspace is on a global scale in front of an audience which is possibly unknown and infinite. In social networking sites, the user’s privacy is threatened because a large audience might have access to his personal information. In order to handle privacy issues, there were few social media sites that offer limited options for making one’s profile visible for a specific set of individuals. As for some cases, audience segregation is used as a solution to protect user’s privacy; however, Goffman’s segregation of audiences is a lot harder in the era of the Internet. Difficulties begin when the audience is used to a certain type of performance from an individual or team but observes another performance which does not create the same impression which results to cyberbullying. The impression created on a social networking profile may not resemble an individual’s real life identity.

The nature of communicating in the cyberspace facilitates the potential for anonymous interactions. It was discovered that bullies who choose to use electronic means can easily hide their real identity and make themselves anonymous. Anonymity can be created through the use of temporary email addresses, fictitious names or unknown mobile number. The perception of anonymity in social media serves as a disinhibitor so that people are more likely to do and say things online that they would not do or say in a face to face situation. Another key characteristic of cyberbullying is the potential to reach a limitless audience. Due to the boundless nature of cyberspace, the audience is not confined to a single setting (such as school or office) but has the potential to be viewed by a global audience.



Goffman's framework offers not only a way of thinking about space in terms of performance but also a way of thinking about how people may act differently depending on the audience and setting which are relevant to an exploration of cyberbullying. Goffman defined three roles in this mechanism: performer, audience, and outsider. These roles can be paralleled to the roles of a target, bully, and bystander. By framing bullying as a performance, a framework is provided that enables us to consider the bystander group as an audience and how different settings may affect how young people act towards others. In order to set the scene for a performance, Goffman made a distinction between the two regions of social space where an individual interacts. The front region is defined as the public performance area. The backstage region is a place wherein the performer can privately prepare for the performance or where members of a group can openly construct the impression they are planning to give. By using Goffman’s framework of performance, cyberspace interactions can be executed by the bully in the backstage region which impacts on the target in the public front stage region. As the backstage region is a place that performers may privately prepare away from the audience, this provides time and space for the bully to plan the ways in which they wish to target others. The physical distance which cyberspace interactions facilitate may also result in the bully managing the impression ‘given off’, the ability for the bully to conceal their identity and the tone and meaning being open to wider interpretation.

**Natural Language Processing**

Natural Language Processing (NLP) is a field of study which focuses on discovering ways on how to bridge the gap between interactions involving humans and computers. It aims to provide a method for computers to analyze and comprehend natural languages (a.k.a. human languages) in an intelligent way, or by means of simulating the process of "understanding" - either through Symbolic approach, which utilizes a set of predefined rules, modelling a different language phenomenon, or Statistical approach, which makes use of machine learning algorithms to learn the language phenomena. Concepts in computer science, artificial intelligence (AI) and computational linguistics are what comprises NLP. After all, Natural Language Processing is said to be the main component of AI and that it relies on machine learning as well - in order to enable the system to derive patterns in a given dataset which would help improve its own understanding of speech. It differs from common word processor operations in such a way that NLP possesses the capability to analyze the word for its meaning rather than only for its structure (viewing the word in a symbolic approach).

A great number of current software applications have been incorporated with NLP tasks in order for them to function appropriately. Some of those tasks are as follows:

* Deep Analytics
* Machine Translation
* Named Entity Extraction
* Co-reference Resolution
* Automatic Summarization
* Sentiment Analysis
* Text Classification
* Conversational Agents

A system’s skill that could count as an example of a Natural Language Processing capability would be developing a decent conversation in pure human language. Likewise, computer systems that can convert human languages to computer languages and vice-versa are currently existent. Translation programs were also made possible by NLP. Additionally, grammar and spelling checkers were also programmed following the mechanism of implementing text processing techniques under Natural Language Processing. Lastly, a computer that can read human languages (in publications such as books) is also a product of NLP.

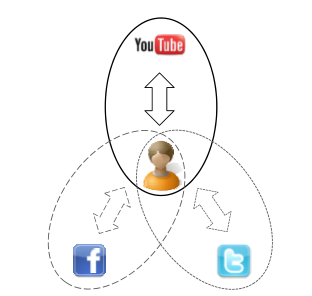
Despite the current capability of NLP in terms of Human-Computer Interaction, it still remained limited particularly in producing statements which involved 100% human reasoning and logic. NLP can only refer to a stricter subset of the human languages which means that it cannot allow anomalies which are often occurring in a particular human language.

**Text Classification**

The team used text classification for their research. It is the classification of documents into a fixed number of predefined categories. Through the use of machine learning, the main objective is to learn classifiers from examples which perform the category assignments automatically. Since categories may overlap, each category is treated as a separate binary classification problem. A Support Vector Machine (SVM) algorithm was used in the construction of the model. As a pre-requisite for SVM, an initial training data set must be classified among 3 categories accordingly: Cyberbullying, Non-cyberbullying, and Ambiguous cyberbullying.

**Corpus**

A corpus is a collection of machine-readable texts that have been produced in a natural communicative setting. They have been sampled to be representative and balanced with respect to particular factors. In order to detect occurrences of cyberbullying in social media, the team gathered a large amount of data as corpus in order for data mining to be done. In this research, they used Import.io, a web crawler to collect a corpus of text posts from social networking sites such as Twitter, Facebook, and Youtube. Lastly, they created a dataset of three classes: Cyberbullying, Non-cyberbullying, and Ambiguous Cyberbullying.



#### Bag of Words

One of the most crucial tasks in text classification is feature extraction. Through the use of machine learning algorithms for training the classifier, representation of text as a feature vector is also required. One of the most popular model used in Natural Language Processing is Bag of Words (BoW) model.

The primary stage of this model is the creation of vocabulary of words which indicates the collection of abusive words. In this model, each word is associated with a count of occurrences. This vocabulary can be understood as a set of non-redundant words where the order doesn’t matter. The BoW approach disregards grammar and detects offensive sentences by checking whether or not they contain offensive words.

The Bag of Words (BoW) model is created by using one of the following preprocessing techniques:

* Unigrams
* Bigrams
* Stemming
* Lemmatization
* Parts of Speech (POS) Tagging

In this research, the team used unigrams preprocessing technique. It is a model used in information retrieval can be treated as the combination of several one-state finite automata. In unigrams, words are treated independently.

**DESIGN AND METHODOLOGY**

In this section, the researchers illustrate their understanding on the process of creating a cyberbullying detection mechanism.

#### Data Collection

The process of creating a cyberbullying detection model begins with the collection of data for the corpus.

Social networking sites such as Youtube, Facebook and Twitter were used as sources of data for the corpus. The dataset from Youtube contains comments from videos focusing on controversial events in the Philippines such as cases of bashing against Filipino celebrities and video bloggers *(pabebe girls, AlDub bashers, Joga girl videos, and others)*, and scandals wherein politicians and celebrities are involved *(the accusation of Leila de Lima as a drug smuggler, the reaction of the Filipinos after Duterte was elected as the president, and others).*These topics are often a rich source for objectionable and rude comments. Most comments on YouTube are described as stand-alone, with users expressing opinions about the subject and content of the video. There were no clear patterns of dialogue in the corpus and it has no conversational features because some of the comments were constructed as responses to previously posted ones.

In Facebook, the team collected posts from the different universities' confession pages because these pages allow anyone to share personal secrets, rumors, gossips, and anything else they might want others to know about but are hesitant to post publicly or in a way that is tied to their identity. Thus, the anonymity of the person posting a confession makes these pages vulnerable to cyberbullying activities. In Twitter, posts from random Filipino people were collected. Twitter is also prone to cyberbullying occurrences since users can easily create fake accounts to launch their bullying cyber-attacks against people they don’t like or disagree with. In 2011, a study conducted by the University of Wisconsin-Madison found that 15,000 abusive tweets per hour, which equals 100,000 abusive tweets a week.

To obtain this data, the team crawled a subset of Youtube, Facebook, and Twitter using Import.io and extracted statements from these sites. The only field that was used in collecting their data was the textual content of the post while disregarding the other features such as the user information, links, and others. A total number of 2000 statements from these social networking sites were collected.

#### Data Annotation

Once the data has been obtained in the corpus, it must be organized in a way that the computer can easily find patterns and inferences by adding relevant metadata to a dataset. Any metadata tag used to mark up elements of the dataset is called an annotation over the input. Data annotation is the process of augmenting a corpus with higher-level information. The main purpose of adding this information to the corpus is it to allow the computer to find features that can make a defined task easier and more accurate.

The second part of the experimentation phase begins with the establishment of classes. After the team has successfully extracted information for their corpus, they formed a dataset of three classes: cyberbullying, non-cyberbullying and ambiguous cyberbullying. The previous data annotation process was accomplished by a single person therefore the researchers had decided to redo the whole process once again by conducting a questionnaire for 100 respondents. Participants will be given a sheet of paper containing 10 sentences taken from the corpus. They will be given the opportunity to annotate these sentences accordingly (whether they think the statement implies cyberbullying, does not imply cyberbullying, or they cannot identify at all). The project team members were also required to annotate the dataset as well.

#### Pre-Processing

After obtaining data for the corpus, the dataset must undergo pre-processing primarily because it contains unstructured text. The purpose of this step is to transform messages into a uniform format that can be understood by the learning algorithm. In pre-processing, the process pf tokenization, stripping of whitespaces and removal of unnecessary characters such as emoticons are performed through the use of Notepad++ and Excel.

**Feature Extraction**

One of the crucial steps in creating a cyberbullying detection model is feature extraction. By using machine learning algorithms for training the classifier, representation of text as a feature vector is required. For this process, the team used the Bag of Words(BoW) in a unigram technique model, which is one of the most commonly used representation in Natural Language Processing. The primary stage of this model is the creation of vocabulary of words which is in this approach indicates the collection of both abusive and non-abusive words. In BoW model, each word is associated with a count of occurrences. This vocabulary is defined as a set of non-redundant words wherein the order doesn’t matter. Each statement is represented as a feature vector composed of binary attributes for each word that occurs in that message Let {w1,…,wm} be a predefined set of m features (vocabulary of words) that can appear in a message. Let ni(d) be the number of times wi occurs in a message d. Then each messaged is represented by the message vector d:=(n1(d), n2(d),…,nm(d)). If a word present in the vocabulary appears in a given text message, its corresponding attribute is set i.e. 1, else it is set to 0.



*Bag of Words-Unigram*

#### Folding

Once the corpus has been adjudicated, it will be partitioned into two parts: the training set and the test set. The division of data into training and testing sets is a crucial part in evaluating the cyberbullying detection model. The main purpose of a training dataset is to train the model, by pairing the input with expected output. The test datase, on the other hand, is used to to estimate how well the model has been trained and to estimate model properties such as classiification errors for classifiers, precision and recall.

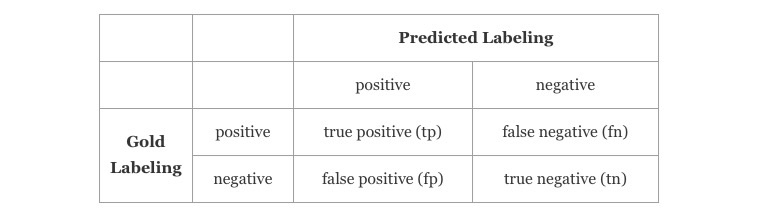
Oftentimes when separating the corpus into a training set and testing set, most of the data is used for training while a smaller portion of the data is used for testing. For this dataset, the team used 75% of the dataset for training and 25%% for testing purpose. From the 2000 statements that were extracted from social networking sites, 1500 statements were used for training the classifier and 500 statements were used for testing the statements against the classifier.

#### Analysis and Classification

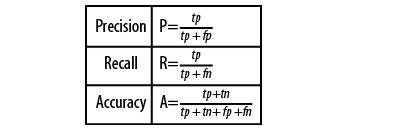
After the creation of Bag-of-Words (BoW) model, the statements were retrieved from Excel. As a pre-requisite step in data classification, the csv file was first converted into .arff (Attribute-Relation File Format) format since it is the one being used in WEKA. Classification is the task of identifying the labeling for a single entity from a set of data. in order to determine cyberbullying from not-cyberbullying data, an algorithm called a classifier is trained on a set of labeled data. Thus, these words are essentially treated as features that the classifier will use to model the positive instances of cyberbullying as compared to non-cyberbullying and ambiguous cyberbullying. In this study, the team used a supervised learning approach to train the classifier in creating a cyberbullying detection model. Moreover, they employed a Support Vector Machine (SVM) model in WEKA as a classification tool.

**Evaluation**

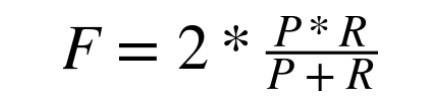
To evaluate the performance of SVM in the creation of a cyberbullying detection model, it is important to calculate how accurate it labels the dataset by measuring the fraction of the results from the dataset that are labeled correctly using a standard technique of "relevance judgment" called the Precision and Recall. For each label used to identify elements in the data, *(Cyberbullying, Non-cyberbullying, and Ambiguous Cyberbullying)*the dataset is partitioned into two subsets: one that is identified relevant to the label and one that is not relevant. Precision is a metric that is computed as the fraction of the correct instances from those that the algorithm labeled as being in the relevant subset. Recall is computed as the fraction of correct items among those that actually belong to the relevant subset. To illustrate how it works, a confusion matrix was provided below:



Given this matrix, the formula for precision, recall and the accuracy of the classifier can be defined as:



The values of both P and R are usually combined into a single metric called the F-measure, which is defined as harmonic mean of the two.



F-score is the average of both the precision and recall. A system with high precision but low recall means that most of its predicted labels are correct when compared to the training labels whereas a system with low precision but high recall means most of its predicted labels are incorrect when compared to the training labels.

**Recommendation and Conclusion**

As modern technology continues to evolve, it has manifested itself in a very serious social problem called cyberbullying. In this paper, the team focused on the process of detecting textual cyberbullying with a dataset from YouTube, Twitter and Facebook. Furthermore, three classes were created from the dataset: Cyberbullying, Non-cyberbullying, and Ambiguous Cyberbullying. The target languages for this experiment are Tagalog and English. In order to detect cyberbullying posts in social media, the team used a supervised learning approach to train a classifier; moreover, they deployed a Support Vector Machines (SVM) model in WEKA as a classification tool because it significantly outperforms the other classifiers in high dimensional feature spaces. The experimental results show that the proposed solution approach was appropriately able to identify textual cyberbullying occurrences with more than % accuracy.

This project is technically not finished yet. So far, the team had accomplished developing a working cyberbullying detection model. However, since they are planning to continue this project up to the two succeeding terms, they hope to add more processes in the experimentation phase for further verification of the model. Such additional processes will involve using TF-IDF scores to weigh both features and attributes, testing the model against other machine learning algorithms (Naive Bayes, Decision Trees, Convolutional Neural Network), utilizing other forms of the Bag-of-Words method (bigrams, trigrams), creating sub-categories for cyberbullying instances (e.g. determining the reason as to why a particular statement was classified as cyberbullying through interviews), and involving *Bekimon* and *Jejemon* words, and texting shortcuts with the aid of the normalization process. The team aims to conduct all the processes mentioned in the succeeding term. In addtion to that, they may also try creating a simple prototype by integrating Java with WEKA toolkit. This way, end-users will be able to type a particular sentence of their choice, submit it to the program, and view its respective results on a basic GUI.

One of the main shortcomings in this field of study was the limited size of dataset. Therefore, the team highly recommended the development of a larger and more diverse dataset for future researchers willing to work with a related project.

The team purposely limited the number of algorithms and methods involved in text classification to establish feasibility of this task. They chose not to use additional classifiers or more computationally heavy performance estimation designs such as nested cross validation to identify parameters. They also avoided comparing feature selection algorithms that may have improved some of the classifiers performance. The researchers focused on the feasibility of using machine learning text classification approach to identify cyberbullying occurrences by using SVM. A point of future work is to apply a more robust study design comparing classifiers as in their prior work.

Lastly, the cyberbullying detection model is implemented only on textual data. In the future, the team plans to extend the scope of their system by incorporating cross-media detection in the form of audio, video and images too.