**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 networking sites are primarily used for communicating and connecting with other people; however, some people use these technologies to harm others emotionally. Cyberbullying refers to the use of technology to humiliate and slander behaviour towards other people with the use of computers or mobile phones. It takes form of hate messages sent through social media and e-mails. Truly, cyberbullying has become rampant in many countries. In order to address this issue, the team proposed a method to detect cyberbullying entries on social media through the use of Natural Language Processing (NLP). The data that was used by the researchers was collected from Youtube, Twitter, and Facebook. Each statement that was extracted from these social networking sites were ranked according to their harmfulness level (wherein 0, indicates no indication of cyberbullying event, 1 indicates mild cyberbullying occurrences and 2 indicates severe cyberbullying occurrences). Furthermore, this statements were annotated into six cyberbullying categories: bad description, intelligence, physical appearance, race and culture, sexuality, and social rejection. Among the 450 words that were extracted from 625 statements, bad description was most prevalent with a frequency of 27%.

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 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, 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 2013 Report" which was conducted by the people behind ComScore indicated that out of the 6 countries in Southeast Asia bearing an online population of at least 62 million, the Philippines ranked fourth. 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 the year 2015, there are about 47 000 000 Filipino users who are active on Facebook. This particular norm of owning at least one personal account in any of the renowned social media sites continues to expand day by day. 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 by child-care nonprofit Stairway Foundation Inc, eighty percent 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 10175 (also known as Cybercrime Act of 2012) was introduced. It recognizes cyberbullying as a kind of cybercrime and provides provisions on the consequences for cyberbullying. However, a lot of people deemed its provisions inefficient (as it requires thorough monitoring of an assigned personnel). 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 Support Vector Machine (SVM) algorithm be utilized in the process of automating the method of detecting cyberbullying occurrences in public social media posts through text classification?

**1.3 Objectives**

**Main Objective**

This research aims to formulate a cyberbullying detection model which will yield at least 70-80% accuracy in terms of detecting cyberbullying occurrences present in public social media posts.

**Specific Objectives**

* To acquire ample data for the corpus
* To apply text pre-processing to the statements included in the dataset
* To extract significant features from the corpus
* To evaluate the extracted features
* To devise appropriate experiments which will test the cyberbullying detection 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 here in the Philippines. As of today's time, Filipino moderators have been utilizing the manual way of flagging offensive posts in social media sites. In other cases, they allow the users themselves to report a particular post in those websites privately so they can execute appropriate action for it. There are also mechanisms which deals with automatically censoring profane words that were typed in a chat box; however, such scenarios were typically present in gaming sites and not social media websites. With the automation of cyberbullying detection, more 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.

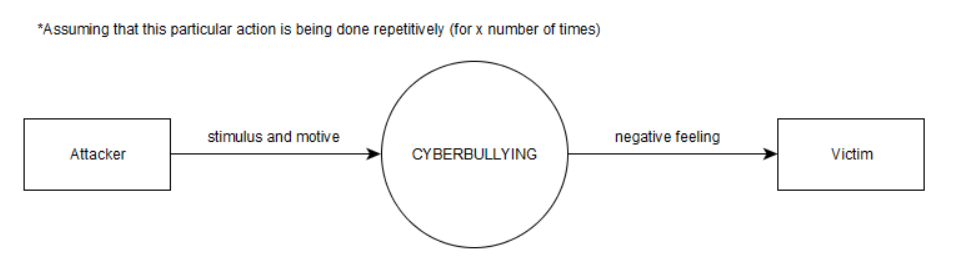
The younger population tend to visit social media websites more frequently than the rest. That fact alone is upsetting. Adolescents have a tendency to deal with things impetuously due to their immaturity. They are most likely unable to identify the intensity of the damage that they had done until it finally occurred. With the automated cyberbullying detection model, such incidents may be prevented before they get out-of-hand. The model will detect posts as long as it notices potential cyberbullying activity - even minor ones as much as possible. The results of the said detection can also give parents and other parties concerned an insight on the behavior or status of an individual who has an account on that particular social media site (granted that they request for it to the moderators).

Most people, typically Filipinos, are reluctant to admit to being victims of cyberbullying (unless it becomes fairly obvious). Severe bullying scenarios rarely happen in the Philippines; therefore, Filipinos tend to become indifferent towards such scenarios. Another possible reason 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. However, 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. 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. As mentioned above, the cyberbullying detection model will be designed to detect even subtle posts implying cyberbullying attacks as much as possible.

**1.5 Scope and Limitations**

This research will be centered towards the concepts and processes involved in the development of the cyberbullying detection model dedicated to aiding a system in comprehending human languages in order for it to acquire the ability to discern and classify cyberbullying from non-cyberbullying statements, given that the aforementioned statements were written in Filipino (which follows the Manillan context), and that they were made public by the user. (creation of the corpus, annotation of the corpus, text pre-processing, feature extraction, and testing)

\*selected methods under Natural Language Processing and Machine Learning algorithms (due to their inter-dependency towards each other) dedicated to aiding a system in comprehending human languages in order for it to acquire the ability to discern and classify cyberbullying from non-cyberbullying statements, given that the aforementioned statements were written in either Tagalog or English, or possibly, both (Taglish), and that they were made public by the user. This paper will likewise introduce concepts which bear utmost significance in terms of performing processing tasks over a large collection of text such as the formation and the subsequent annotation and pre-processing of the textual corpus (dataset), the process of extracting features from the dataset, and the creation of classes which will revolve around the predefined cyberbullying categories (refer to IV. Design and Methodology). Additionally, experiments will be conducted in order to validate the efficacy of the proposed cyberbullying detection model. These experiments will involve the simulation of various instances that may affect the model's classification accuracy. The said accuracy will be measured numerically using bases such as Precision, Recall, F-measures and Kappa statistics. The final part of the experiment involves testing the model incorporated with a specific type of Machine Learning algorithm - Linear Support Vector Machine algorithm.

**1.6 Context Diagram**

**REVIEW OF RELATED LITERATURE**

**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 - the extraction of information from large or multi-source datasets with the help of advanced data processing techniques. It is most useful in dealing with precisely targeted or highly complex queries to be done in unstructured or semi-structured data.
* Machine Translation - translating one human language into another.
* Named Entity Extraction - a data mining process which is concerned with the identification of one item in a set from other items having similar attributes based on a named entity definition - a word or phrase that can clearly distinguish an object from the rest.
* Co-reference Resolution - determining which words refer to the same object.
* Automatic Summarization - creating a summarized version of a large chunk of text.
* Sentiment Analysis - identifying subjective information (e.g. judgement, opinion, emotional state, etc.) in text.
* Text Classification - assigning predefined categories to a text.
* Conversational Agents - systems having given the capability to coherently converse with humans.

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.

**Textual Corpus**

Or simply **corpus**, is a collection of large sets of text specifically chosen by the researcher (or linguist, in terms of linguistics) to deduct his/her own linguistic analysis – may either be statistical analysis and hypothesis testing, checking occurrences or validating specific rules - in the text of interest. In the field of Natural Language Processing, the analysis conducted can be used as a basis or test bed for constructing NLP systems. They are typically processed and stored electronically (with the use of databases, etc.) in today’s time.

**Corpus Annotation**

The method of adding linguistic information to the data included in the corpus is known as corpus annotation. The most common type of annotation being done to the dataset is by putting labels or tags which indicates the class to which a particular word in the body of text belongs to.

**Text Pre-processing Tools**

The performance of the terminology extraction system highly depends on the quality of the linguistic preprocessing; therefore, a high-performance pre-processing toolkit is crucial in achieving the pre-processing steps.

**LeTs Preprocess Toolkit** is a suite of robust pre-processing modules. It includes Part-of-Speech Taggers, Lemmatizers and Named Entity Recognizers for four different languages: Dutch, English, French and German. By using this toolkit, each token will be assigned as a class tag, PoS, named entity, or a lemma tag. It also supports both monolingual (single language) and parallel corpus (two languages). As an initial step, the data in the corpus is separately pre-processed, by splitting it into sentences, tokenized, Part-of-Speech tagged, lemmatized, and finally named entities are extracted.

**The Natural Language Toolkit (NLTK)** is a platform used for creating Python programs to work with human language data in the application of natural language processing (NLP). It contains a suite of text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. The NLTK’s basic functionalities include: sentence tokenization, word tokenization, wordnet, synsets, and synonyms, stemming words and lemmas.

**Methods in Pre-processing Text**

**Chunking.** It is used for recovering phrases constructed by the part-of-speech tags, finding noun phrases, finding verb groups, and others. It is commonly used for information extraction, keywords extraction, entity recognition, and even relation extraction. Like tokenization, it also omits whitespaces; however, chunking usually selects a subset of the tokens.

**Lemmatization.** It is the process of grouping together the different inflected forms of a word in order to determine the lemma in the given context. Lemma is defined as the canonical form, dictionary form, or citation form of a set of words. The lemmatization process involves understanding the context and determining the part of speech of a word in a sentence.

**Stemming.** It involves the process of reducing inflected words to their word stem, base or root form by removing its affixes. The stem does not need to be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

**Part-of-Speech Tagging.** It is the process of classifying words into their parts of speech and labeling them accordingly. Parts of speech are also known as lexical categories; moreover, the collection of tags used for this process is known as a tagset.

**Tokenization.** Naturally, text needs to be segmented into linguistic units such as words, punctuation, numbers, and alpha-numeric characters before text processing is to be done. This process is called tokenization. It is the process of breaking a set of text into words, phrases, symbols or tokens. At the same time, it omits certain characters such as punctuations and white spaces. Tokens are used as inputs in parsing or text mining.

**Related Studies**

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.

Dinakar et al. conducted a research at Massachusetts Institute of Technology. The research was aimed towards detecting cyberbullying content in Youtube comments. The first level of classification involves determining if the comment is in a range of sensitive cyberbullying topics such as sexuality, race, intelligence and physical attributes. The second level is to determine what topic. The experiment achieved 66.7% accuracy for detecting instances of cyberbullying. This project also used a support vector machine learner. The researchers based their pre-defined classification of cyberbullying types on the sensitive topics enumerated above. Hinduja and Patchin included harmfulness as one of the constituents of cyberbullying. Similarly, in terms of sensitive topics, there will always be that particular group of people which would have been offended by it (which is why it was deemed controversial); therefore, the researchers thought that by recalling well-known controversial issues in the Philippines, they will be able to determine the borderline between cyberbullying and non-cyberbullying statements, making it easier to tell them apart and classify them.

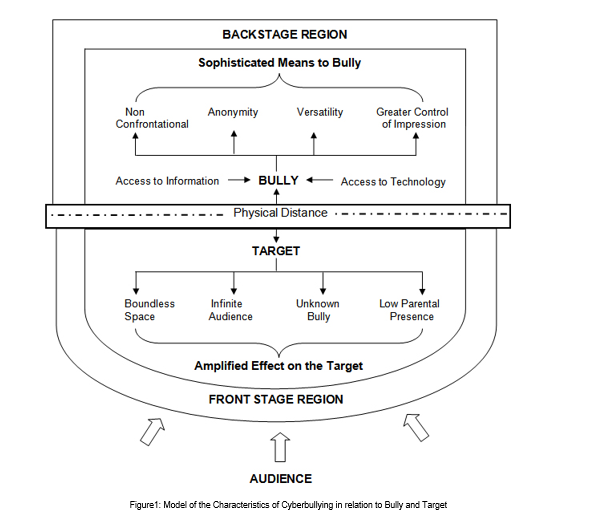
In 2016, Cheng and Ng conducted a research at De La Salle University. The research aimed towards detecting cyberbullying roles through textual context in Facebook and Twitter. First, the researchers identified six roles in a cyberbullying context: the bully, victim, assistants of the bully, reinforcers, 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.

**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.

**DESIGN AND METHODOLOGY**

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

**The Creation of the Corpus**

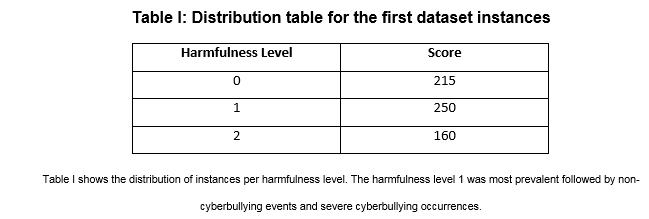
**Data Collection and Annotation**

The team acquired data from Youtube. A total of 625 comments (both cyberbullying and non-cyberbullying) were obtained and evaluated according to their corresponding level of harmfulness (0 - for comments containing no cyberbullying occurrences, 1 - for comments expressing indirect cyberbullying content, 2 - for comments explicitly demonstrating severe cyberbullying events). Those statements which were labeled 1 or 2 were further classified into four categories - *Physical Appearance*, *Race and Culture*, *Intelligence*, and *Social Rejection* - which were deemed related towards controversial issues in the Philippines.

Moreover, the team decided to use Import.io as their primary tool for the annotation process. It is a powerful and easy-to-use tool for data extraction that has the aim of getting data from any website in a structured way. The YouTube dataset contains comments from videos on controversial events in the Philippines that were considered controversial as these topics would most likely encourage viewer to post their opinions about it. Most of the cyberbullying statements that the researchers obtained came from YouTube. In Facebook, they collected posts from different universities secret pages, wherein a user can post his/her story about controversial school experiences. In Twitter, posts from random Filipino people were collected.

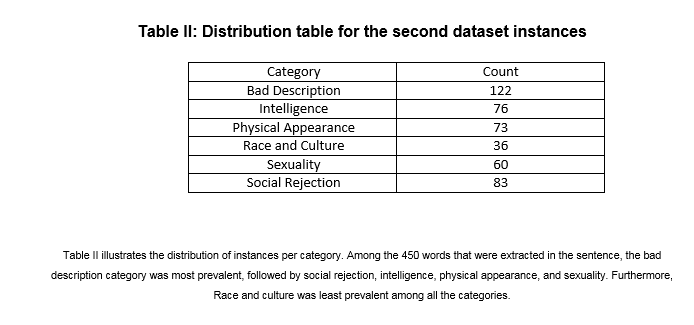
The dataset was cleaned in order to remove unnecessary symbols present in text. These symbols were removed as so they will not be extracted as features later on; furthermore, the dataset underwent normalization using Special Text Replacement function in Excel.

One of the researchers was in charge of annotating the dataset. The annotator simply tagged each instance based on how she perceived it. All annotations are performed using General Architecture for Text Engineering (GATE) and Brat Rapid Animation Tool (BRAT). GATE is a tool wherein a user can easily input an annotation schema for entities. Brat, on the other hand, is an intuitive web-based tool for text annotation supported by Natural Language. These tools are open source software.



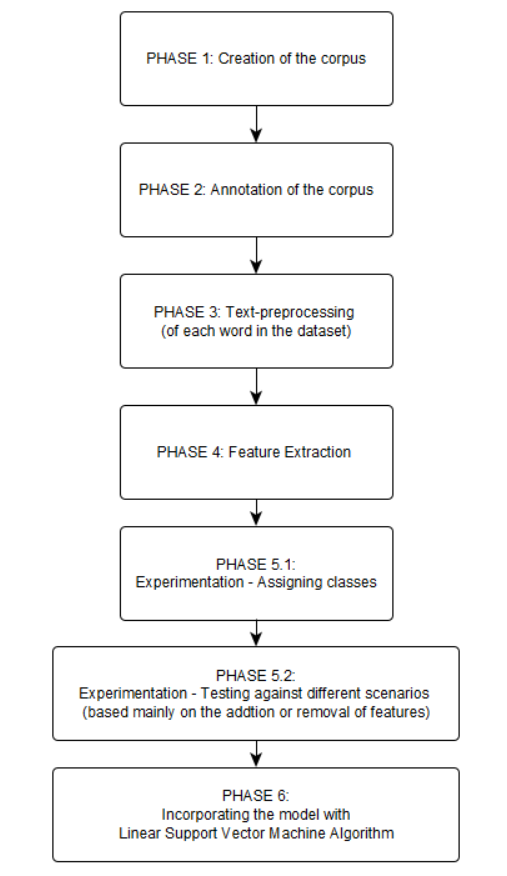
At the second level of annotation, the following guidelines were given:

* Sexuality – instances that contains both vulgar and filthy words.
* Physical Appearance – instances that contains cyberbullying statements with regards to the physical characteristics of a person.
* Intelligence – instances that attacks the mental capacity of a person.
* Race and Culture – instances that contains both racial and cultural discrimination.
* Social Rejection – instances that contains cyberbullying statements which isolates an individual from the society or degrades a person.
* Bad Description – instances that contains cyberbullying statements that pertains to the attitudes and behaviour of an individual; however, this categorization can be misused with the physical appearance.



In the succeeding phases of the project, the researchers are planning to continue with the text pre-processing tasks that are needed in order to lessen the difficulty of extracting linguistic features from the text (which is the phase of the project that will come after the pre-processing phase). Feature selection will be implemented using the following methods: Bag of Words – which will yield bigrams and trigrams, and application of TF-IDF scores – which will aid in the process of determining the most relevant words in each cyberbullying category, as these scores will be used to evaluate the bigrams and trigrams of the instances resulting from the Bag of Words approach (the higher the score, the more relevant the term is). Profane words and words written in uppercase letters are going to be considered offensive by default.

The initial part of the experimentation phase begins with the establishment of classes involving the predefined classifications of cyberbullying expressions based on sensitive issues. This is to verify the accuracy of the model in terms of functioning side by side with features that are closely related to each other. This process will also determine which words can fall in more than one class so they can be given appropriate weights (feature weighting) – e.g. a particular word is more appropriate to be classified in this class compared to the other despite having the characteristics if being included in the latter class as well. The second part of the experimentation phase will take into consideration the final features that will be selected (or the most determining words for each cyberbullying classifier based on offensive issues). The group may either remove or add more features depending on the results the experiments will yield regarding the cyberbullying detection model’s accuracy.



Proposed plan of activities for the accomplishment of the project

**RESULTS AND DISCUSSIONS**

This section presents the outcome of the proponent's preliminary experiments.

The data was acquired from social networking sites such as Facebook, Twitter and YouTube; moreover, two classification tasks were carried out: cyberbullying event detection and the classification of text categories related to cyberbullying. Using the Special Text Replacement function in Excel, the dataset was normalized. The researchers get the frequency of each instance of the word under each category. Out of 625 statements that were extracted from the social media sites, the harmfulness level 1 was most prevalent with a frequency of 33%. It was followed by non-cyberbullying events, with a frequency of 41%. The occurrences of severe cyberbullying events were the least prevalent amongst the three classifications since it obtained a frequency of 26%.

At the second level of annotation, the bad description was the most prevalent with a frequency of 27%, it was followed by Social Rejection (18%), Intelligence (16%), Sexuality (13%), Physical Appearance (10%) and the category Race and Culture was the least prevalent among the six categories, it has a frequency of 8%.