**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)

Table of Contents

[**ABSTRACT** 2](#_Toc468680891)

[**INTRODUCTION** 3](#_Toc468680892)

[1.1 Background of the Problem 3](#_Toc468680893)

[1.2 Statement of the Problem 5](#_Toc468680894)

[1.3 Objectives 5](#_Toc468680895)

[1.4 Significance 6](#_Toc468680896)

[1.5 Scope and Limitations 9](#_Toc468680897)

[1.6 Context Diagram 10](#_Toc468680898)

[**REVIEW OF RELATED LITERATURE** 10](#_Toc468680899)

[Related Studies 10](#_Toc468680900)

[**THEORETICAL FRAMEWORK** 13](#_Toc468680901)

[**DESIGN AND METHODOLOGY** 18](#_Toc468680902)

[**RESULTS AND DISCUSSIONS** 26](#_Toc468680903)

[**INITIAL FINDINGS AND PLANS FOR THE SUCCEEDING TERM** 28](#_Toc468680904)

# 

# **ABSTRACT**

As modern technology continuous to evolve, it has manifested itself in a serious social problem called cyberbullying. Cyberbullying is defined as an online aggressive behavior in the digital space. Recently, it has become rampant in many countries. As cases of cyberbullying continues to grow, it has deemed negative impact on both the psychological and emotional well-being of a person and in some extreme cases, it may lead to suicide. Therefore, this problem must be addressed immediately.

This paper presents the creation of a cyberbullying detection model through the use of text classification. The first step is to gather data for the corpus by crawling a subset of YouTube, Facebook, and Twitter. Furthermore, 2000 statements were gathered for the dataset (1000 from Youtube, 500 from Facebook, and 500 from Twitter). Data cleaning involves the removal of all special characters, non-readable text, emoticons, links, and characters belonging to various foreign countries' writing systems. The dataset will be divided per each word within a particular statement based on the whitespaces separating them. Furthermore, it was represented using Bag-of-Words model. The data was annotated manually by 100 respondents and the proponents themselves. The categories for the dataset are: Cyberbullying, Non-cyberbullying and Ambiguous Cyberbullying. The team deployed Support Vector Machine in WEKA to create their model.

For their preliminary experiments, 900 statements were utilized in Weka which yields an accuracy of 57.89%.

Keywords: *Cyberbullying, Detection, Implications, Social Media, SVM, Text Classification*

# **INTRODUCTION**

### 1.1 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.

### 1.2 Statement of the Problem

How can the detection of cyberbullying occurrences in public social media posts be automated?

### 1.3 Objectives

**Main Objective**

This research aims to develop a cyberbullying detection model based on the concept of text classification

**Specific Objectives**

* To acquire data for the corpus
* To perform the cleaning procedure for the dataset
* To extract features from the corpus
* To perform text annotations
* To perform the Bag-of-Words (BoW) approach
* To implement the text classification algorithm (SVM)
* To set-up a basic GUI for the system
* To perform experiments
* To evaluate the model’s accuracy

### 1.4 Significance

The main significance of this research project is aimed towards the field of Computer Science. Based on the 2013 Computer Science Curricula of the Association for Computing Machinery (ACM), concepts specifically those pertaining to algorithms, data structures, data abstractions, computer architecture, modeling and simulation, discrete mathematics, probability, automata theory, artificial intelligence, programming languages, etc. comprise the core of this particular field. The proponents of this study made certain that such concepts were included and utilized in their research (e.g. implementing the cyberbullying detection model with the use of a machine learning algorithm (SVM), utilizing discrete mathematics concepts in the annotation of the dataset, using concepts in automata theory, such as regular expressions, to aid in the cleaning of the dataset). Doing so will not only contribute to the said body of knowledge, but more importantly, it may inspire their fellow CS students to appreciate Computer Science concepts more, knowing that they are indeed beneficial (with this study as the proof).

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.

System developers are also likely to benefit from this research. They can implement the model to an automatic cyberbullying detection system, which can be further integrated with social networking sites to detect any events of cyberbullying occurring within the cyberspace. In the event that the said integration will be made possible in the future, the benefactors of this study will be extended to the following entities: Filipino social media users (in general), and Filipino adolescents together with their parents.

*For the Filipino social media users* - 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.

*For the adolescents* - 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.

*For the parents* - 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.

The current status of research and development in the Philippines is typically limited to presenting common findings. According to Professor Flor Lacanilao of the University of the Philippines, Diliman, this is due to the fact that the research practices being instilled in most Philippine universities and colleges are considered wrong, for they are done merely for thesis purposes and not for the industry. Such scenarios will, in turn, make research papers appear as "gray literature", specifically those relating to the field of Computer Science. This would prompt people to leave the job of performance evaluation to the scientists instead. Moreover, there appears to be a huge gap between universities and industries as well. Max V. de Leon of the Business Mirror stated that schools are not totally open in terms of handing their researches to private sectors. On the other hand, industries are not putting enough money for academic researches. The overall effect of the aforementioned scenarios lead to the production of lesser research materials that can benefit student researchers. As students of a college instilling the values of "Real Projects. Real Learning" and by being able to find aid in other NLP researchers, the proponents of this study have decided to take the first step in conquering research barriers in the Philippines with the hopes that their project may serve as a basis for the succeeding Computer Science batches in their college.

### 1.5 Scope and Limitations

With the aid of the concept of text classification, this research aims to develop a cyberbullying detection model having a target yield 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 with and/or stated by Filipinos residing in Metro Manila.

The corpus (dataset), created by the researchers themselves, consisted of 2000 statements which were obtained from either Facebook and Twitter posts or Youtube comments (with the consideration that they are all available to the public). The totality of these statements pertained to 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)

The cleaning of the dataset involved the removal of all special characters (excluding apostrophes and hyphens), non-readable text (e.g. asdfghjkl), emoticons, links, and characters belonging to various foreign countries' writing systems.

In the feature extraction process, the Bag-of-Words approach (unigram) was used. Features represented every unique instance of a particular word in all of the statements contained within the corpus. The number of occurrences (frequency) for each feature in a given statement was likewise evaluated through MS Excel with the aid of Excel formulas.

Text annotation was based on the 3 annotation schemes - cyberbullying, not cyberbullying, and ambiguous cyberbullying. Among the 2000 statements used, 1000 were annotated by the researchers while the other 1000 were distributed among a sample of Metro Manila residents - 100 to be exact - with varying ages and occupations, with the use of questionnaires containing 10 questions each. About 90% of the selected respondents were college students while the rest of the 10% were comprised of immediate family members and other relatives.

Currently, the process of flagging cyberbullying statements - by means of a text classification algorithm known as Support Vector Machine (SVM) - did not require a particular number of flagged statements to be met before considering one as a "cyberbullying" statement, which contradicts with the stated definition of cyberbullying in the preceding parts of the paper (that it has to be occurring repetitively before being considered as such). It may, however, be implemented in the future.

The sole experiment that was performed involved the use of the Support Vector Machine (SVM) algorithm on, supposedly, the 2000 statements. However, due to limited time, only 900 statements were utilized in the experimentation phase. 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 that was obtained from the processing of the features in the preceding stages. An evaluation procedure (in WEKA) known as the Cross Validation Folds 10 was used in order for the team to determine the accuracy, precision, recall, and F-measure of the constructed cyberbullying detection model.

The researchers 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

For the past years, the automatic classification of documents into predefined categories has been given an utmost attention by the researchers. Furthermore, the tasks involved in text classification including the machine learning approaches such as Naïve Bayes, Decision Tree, K-nearest neighbor (KNN), Support Vector Machines (SVM) and Convolutional Neural Network have been extensively studied.

In 2012, Wahbeh and Al-Khabi conducted a study on Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text. The aim of the researchers was to compare the performance of three most popular text classification techniques namely, Support Vector Machine (SVM), Naïve Bayes, and C4.5. The dataset were classified into four categories: sports, economics, politics, and prophet Mohammed sayings. The pre-processing steps for the dataset involve the removal of stop words, normalization of some characters, and the deletion of non Arabic text and symbols. After conducting the experiments the Naïve Bayes classifier achieves the highest accuracy followed by the SVM classifier, and C4.5 classifier respectively. The SVM requires the lowest amount of time to build the model needed to classify Arabic documents, followed by Naïve Bayes Classifier, and C4.5 classifier respectively.

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.

One of the crucial requirements in using Machine Learning algorithms is the representation of data as a set of feature vectors. Several studies were conducted to emphasize the significance of data representation in text classification. One of the studies on data representation was conducted in India by Soumya and Shibily (2014). In their study on Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature. In their research, they emphasized the importance of using a BoW model in text classification. For their dataset, they extracted 20 newsgroup data which is made available in the Web and contains a standard train and test split. The preprocessing step involves stemming using a snowball stemmer. The goal is to classify an article into four categories: ”alt.atheism”, ”comp.graphics”, ”rec.sport.hockey”, ”sci.electronics”. They used Naive Bayes classifier as their classification tool. The classification model yields 90% accuracy.

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

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

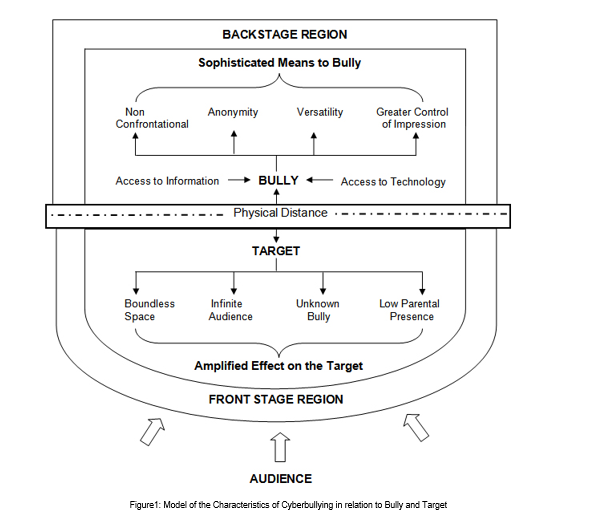
In 2016, Cheng and Ng conducted a research at De La Salle University. The research aimed towards detecting cyberbullying roles through textual context in Facebook and Twitter. First, the researchers identified six roles in a cyberbullying context: the bully, victim, assistants of the bully, reinforcers,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

# **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.

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:

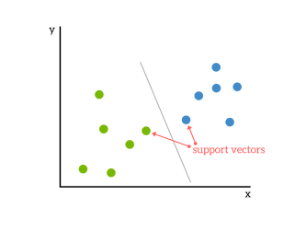
* Deep Analytics
* Machine Translation
* Named Entity Extraction
* Co-reference Resolution
* Automatic Summarization
* Sentiment Analysis
* Text Classification
* Conversational Agents

**Text Classification**

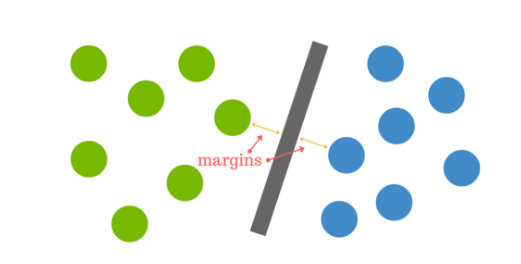
Among the aforementioned tasks that NLP can perform, the team used text classification for their research. It is the classification of documents into a fixed number of predefined categories based on their content. The classification task is also called text categorization. The goal is to create a classification model that is able to assign the correct class to a new document. The document can be classified as single label or multi-label. A single label document belongs to only one class while a multi-label document can belong to more than one class. Furthermore, text classification begins with the creation of the corpus which pertains to a large collection of texts.

**Support Vector Machine**

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in text classification. SVMs seeks to find a hyperplane that best divides a dataset into two classes. Support vectors are the data points nearest to the hyperplane. These elements are deemed critical in a dataset.



A hyperplane is a line that linearly separates and classifies a set of data. Whenever a data is added, the side of the hyperplane where it lands will determine the class that will be assigned to it. The margin, on the other hand, pertains to the distance between the hyperplane and the nearest data point from either set.



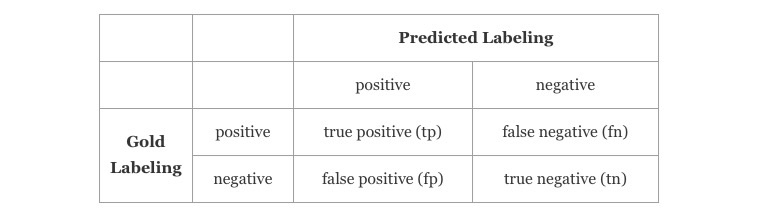
Some of the advantages of a Support Vector Machine (SVM) classifier are SVM yields a high accuracy and it works well with smaller cleaner dataset.

**Bag of Words**

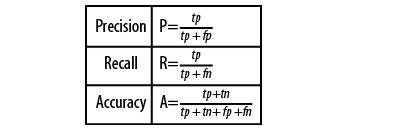
One of the crucial steps in text classification is creating features from a bag-of-words (BoW). The goal of BoW model is to create a numeric representation for the data. In this model, a text is represented as an unordered collection of its words, regardless of its grammar and even word order. A word in a document is assigned a weight according to its frequency in the document and frequency in between different documents. A BoW model is composed of words together with their respective weights.

**Performance Measures**

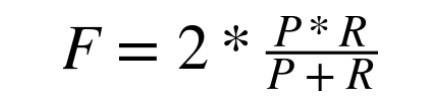
To evaluate the performance of a classifier, it is important to calculate how accurate it labels the dataset by measuring the fraction of the results from the dataset that are labeled correctly using a standard technique of "relevance judgment" called the Precision and Recall. For each label used to identify elements in the data, 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:



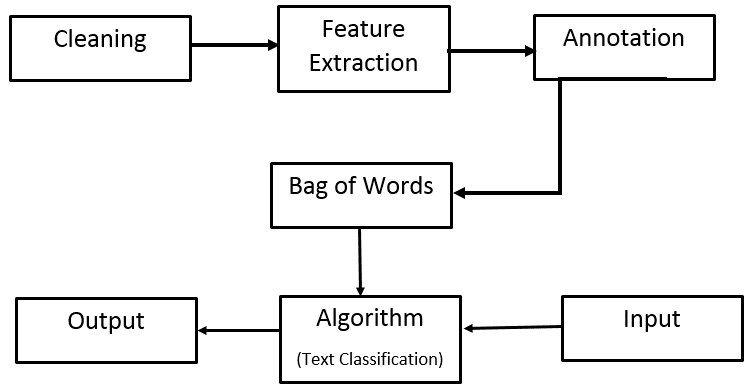
The values of both P and R are usually combined into a single metric called the F-measure, which is defined as harmonic mean of the two.



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.

# **DESIGN AND METHODOLOGY**

In this section, the researchers illustrate the processes involved in the creation of the cyberbullying detection model.



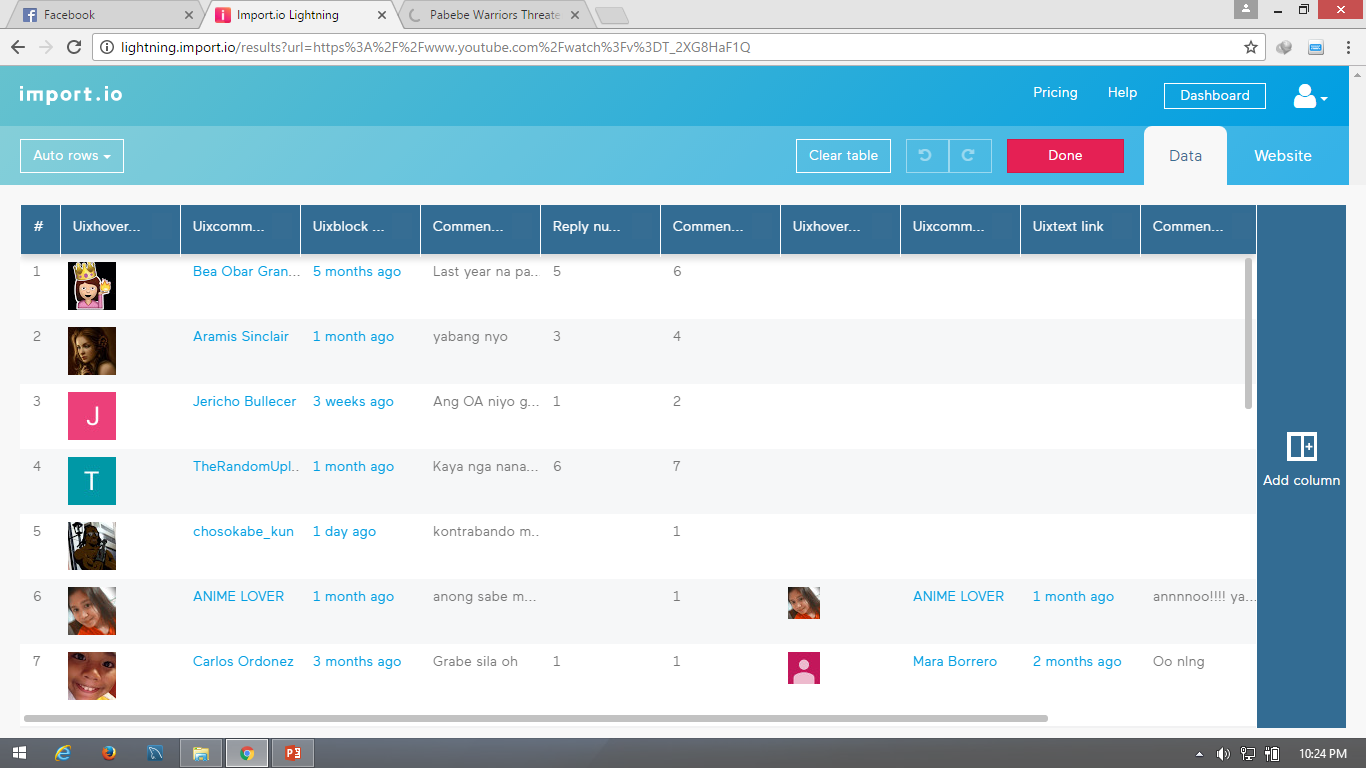
**Data Collection**

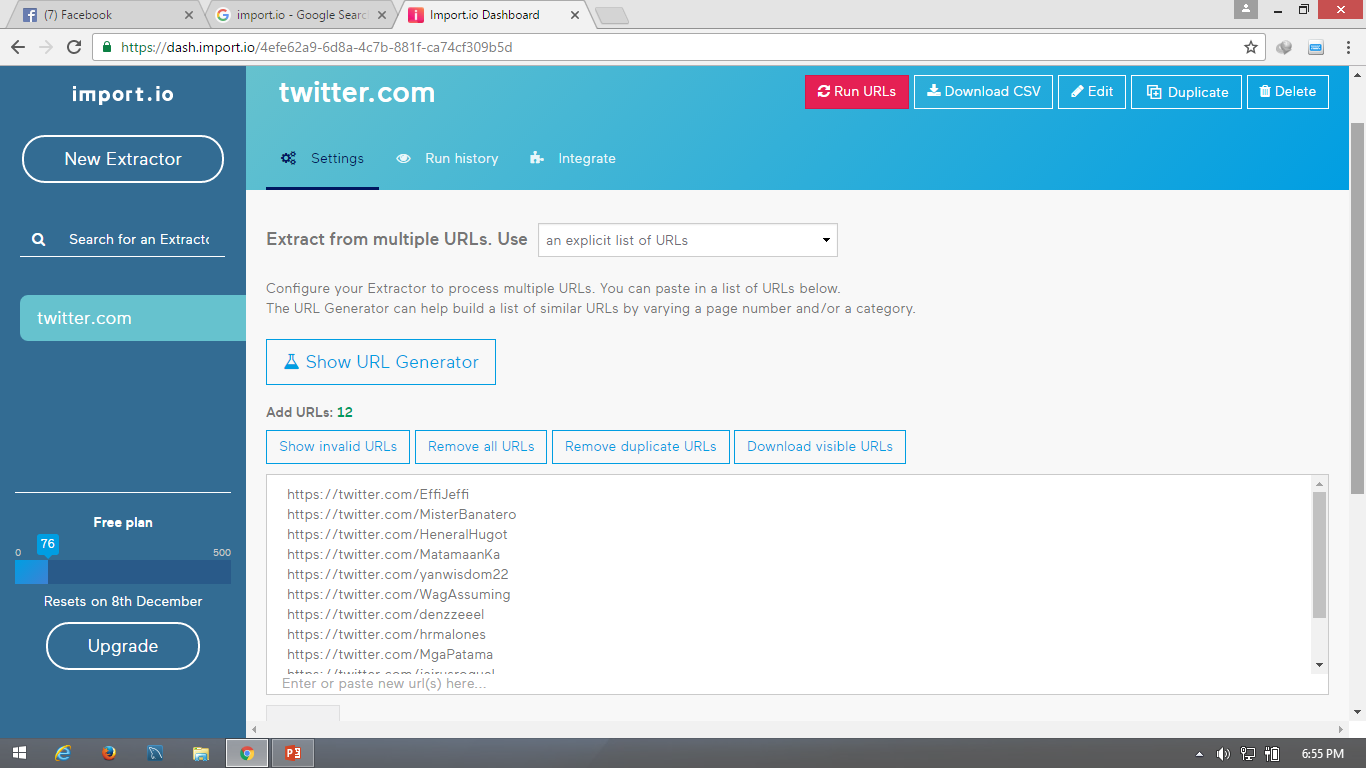
The process of creating a cyberbullying detection model begins with the collection of data for the corpus.

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 (all written in Filipino) from these social networking sites were collected.





*The use of Import.io in extracting data from the aforementioned social media sites*

**Cleaning of the dataset**

The cleaning procedure that was applied on the dataset involved the removal of all special characters (excluding apostrophes and hyphens), non-readable text (e.g. asdfghjkl), emoticons, links, and characters belonging to various foreign countries' writing systems. This was done in order to prevent complications from arising particularly during the experimental phase of the project. Such characters do not make any sense with regard to the detection of cyberbullying occurrences, therefore their appearance may contribute to a probable decrease in the accuracy rate of the model. Apostrophes and hyphens, on the other hand, were retained for they help join characters together in order to yield another word. Since the presence of distinct features were used as basis for the frequency of each word in every statement, it is important to include all words (that can be counted as one - not separated by spaces) preserved in forms understandable by Filipinos within the dataset. This was procedure was done using Notepad++

**Feature Extraction**

The feature extraction phase succeeds the cleaning of the dataset. In this phase, all of the statements that were cleaned will be divided per each word within a particular statement based on the whitespaces separating them. This will be done in MS Excel with the help of its Find tool (Ctrl + F) - which works by matching regular expressions. After that, an Excel function will once again be applied to the dataset. This function will help provide each distinct occurrence of all the words that were part of the statements stored within the corpus. Once this process had been accomplished, it is now time to determine the number of occurrences (frequency) of each feature as they occur in every statement. The acquired numerical values will then be used in the implementation of the Bag-of-Words (unigram) approach.

**Data Annotation**

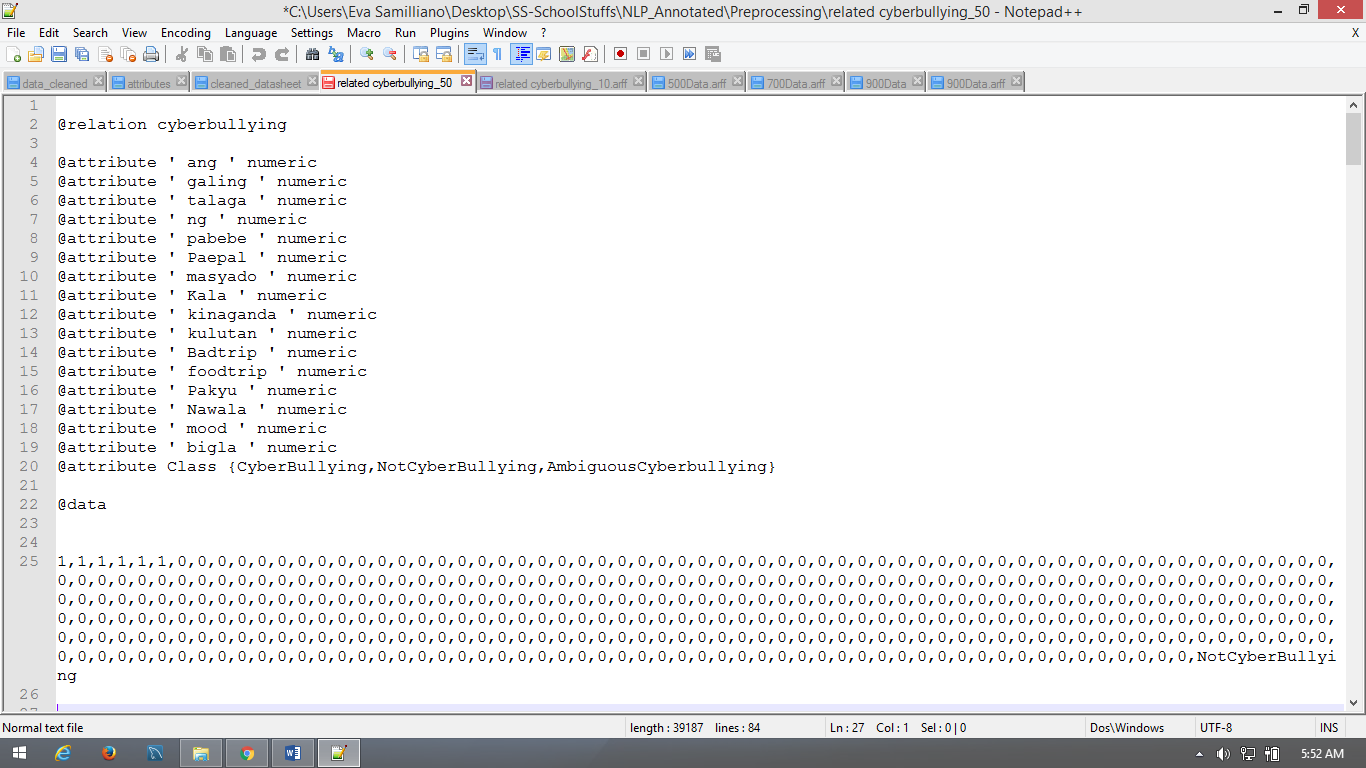
Once the data has been obtained in the corpus, it must be organized in a way that the computer can easily find patterns and inferences by adding relevant metadata to a dataset. Any metadata tag used to mark-up elements of the dataset is called an annotation over the input. Data annotation is the process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

After the team had 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 distributing questionnaires to each 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 (Cyberbullying), does not imply cyberbullying (Not cyberbullying), or they cannot identify at all (Ambiguous cyberbullying)). The project team members were also required to annotate the dataset as well. They worked with the remaining 1000 statements that were included in the corpus.

**Performing the unigram Bag-of-Words approach**

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 number 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.

After cleaning the dataset, the csv file was converted into .arff (Attribute-Relation File Format) format since it is the one being used in WEKA. In this format, the distinct features will be represented by the attributes, and the relation as the whole corpus itself. At the bottom part of the file, the number of occurrences (of each word in every statement) along with the annotations placed by both the researchers and their correspondents (in every statement), will be placed. Such data initially came from the .csv file containing the cleaned, parsed, and evaluated words comprising each of the 2000 statements.



*Bag of Words - Unigram*

**Implementing the Text Classification Algorithm**

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.

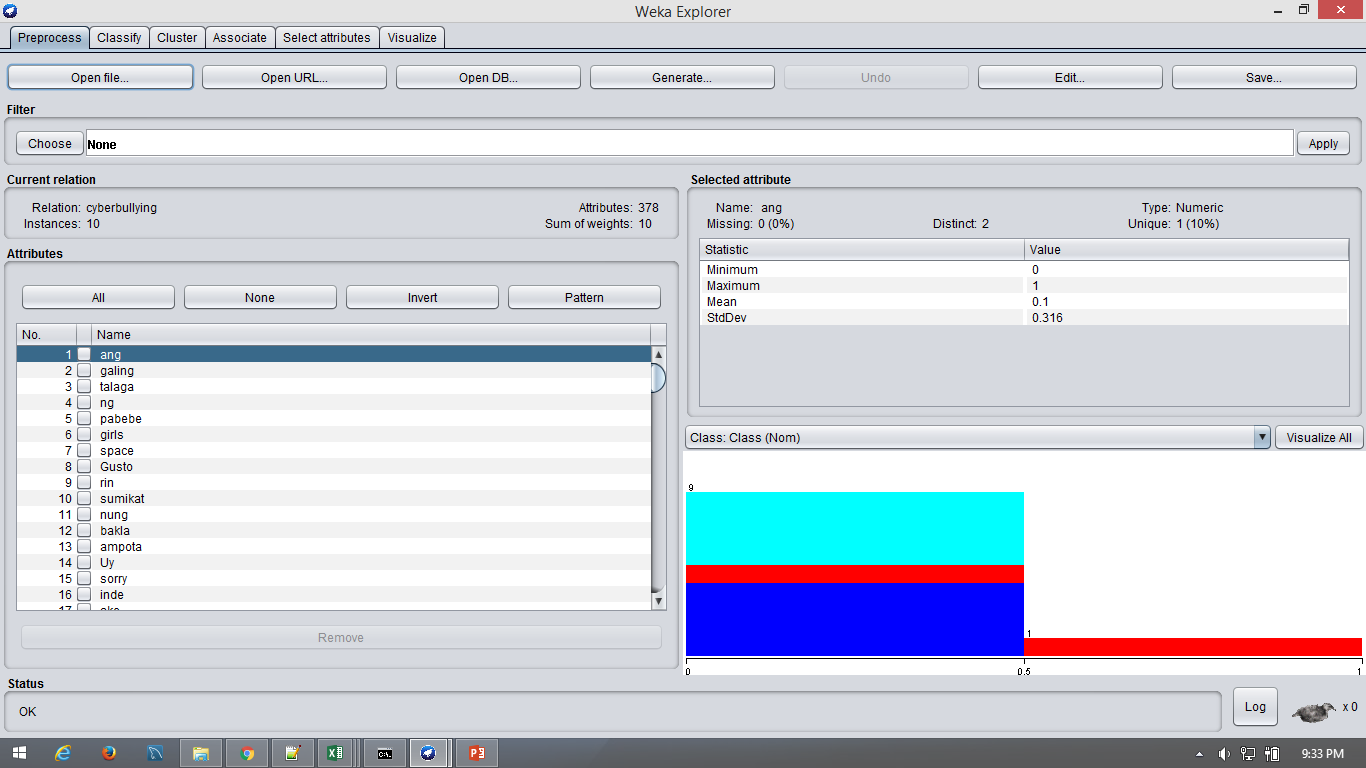
The Support Vector Machine algorithm was the only text classification algorithm that was used in the research project for this term. It was implemented in the WEKA toolkit – which is the software that the researchers used to test the accuracy of their model in terms of detecting and flagging statements containing implications of cyberbullying.

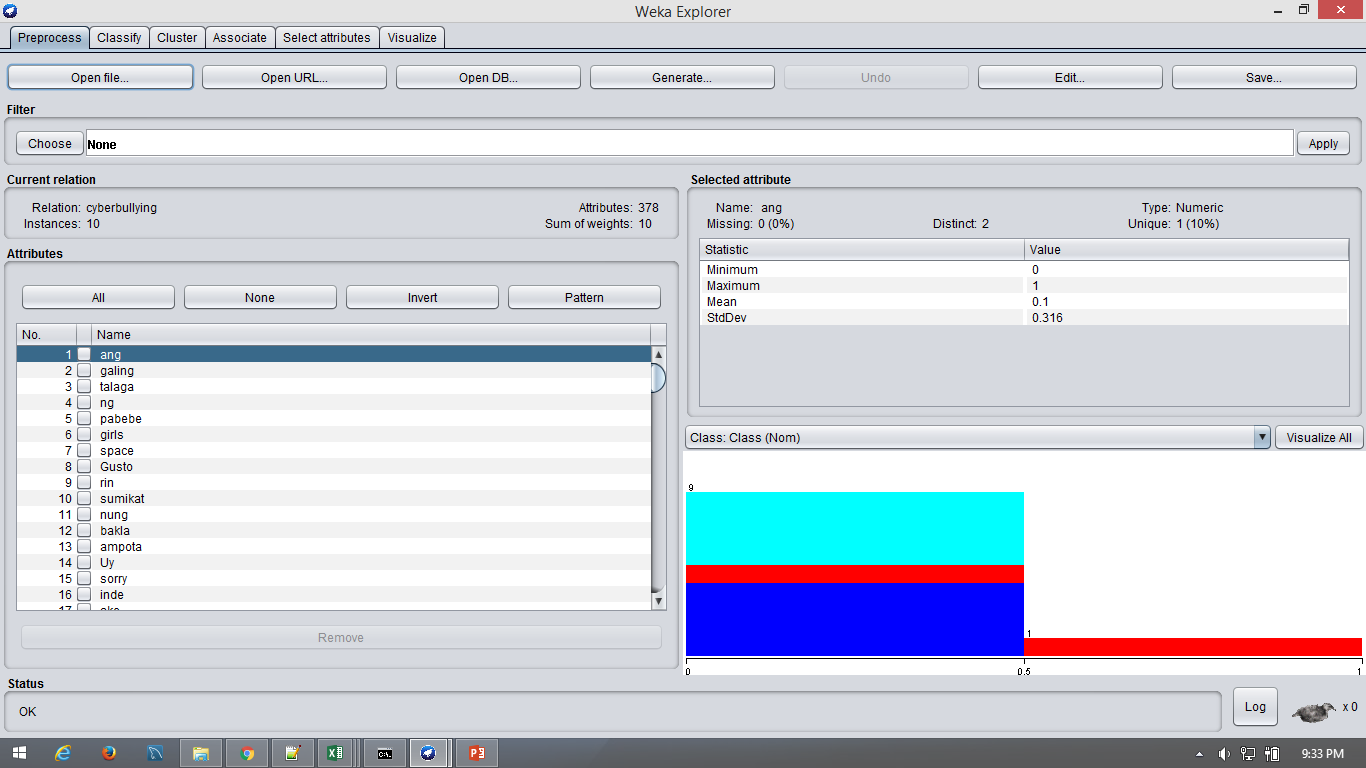
Following this method prevented the researchers to consider the repetition of the statements. As defined in the Background of the Problem, for cyberbullying to occur, it must also be repetitive. Instead, for this study, the only requirement for a cyberbullying statement to be considered as “cyberbullying” is the presence of the cyberbullying features in a particular statement. In the near future, however, the researchers may opt to consider implementing such functionality in the automation process.

**Performing experiments**

The sole experiment that was performed involved the use of the Support Vector Machine (SVM) algorithm on, supposedly, the 2000 statements. However, due to limited time, only 900 statements were utilized in the experimentation phase.

In this phase, the algorithm will be implemented together with the processed data with the help of a single button-click in WEKA. The flagging or detecting of the cyberbullying statements takes place in this phase, however it will not be explicitly demonstrated as WEKA handles the entirety of its processes in the background. There will be charts that the tool will present to indicate how it classified a particular statement.





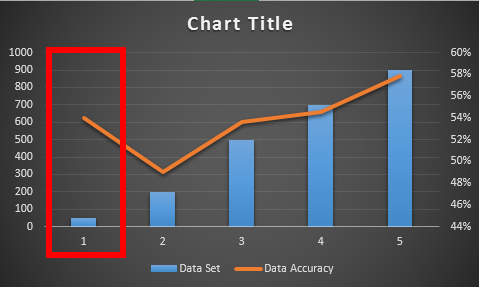
*The WEKA interface (before model implementation)*

# **RESULTS AND DISCUSSIONS**

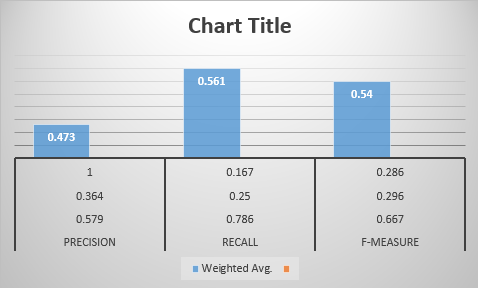
This section presents the results of the experiments conducted by the proponents.

For their first experiment, the proponents extracted 50 data from the corpus. The Bag of Words (BoW) model for these sample contains 378 attributes and were classified into 3 categories (Cyberbullying, Non-cyberbullying, and Ambiguous Cyberbullying). Furthermore, they were partitioned into testing and training data (10 statements were used for testing and 40 statements for training). The result of the first experiments show that the model yields and accuracy of 10% for the testing while 54% for the training.

For their second experiment, the proponents added 900 statements from the corpus. In this phase, they decided to use a different approach by using 10 fold cross-validation in Weka rather than partitioning the data into testing and training set. As the sample data became larger, the attributes in the BoW model has also increased. From the 900 statements that were used for this experiment, the BoW model generates a number of 7062 attributes. These attributes were classified into 3 categories as well. The purpose of conducting small experiments before inserting the whole data from the corpus is to illustrate how the accuracy of the model can change depending on the number of data. The model yields an accuracy of 57.89%



*The chart illustrates that as the number of data increases, there is a possibility that it will yield a higher accuracy*



*The chart shows the result of the second experiment based on the Precision, Recall, and F-Measure.*

# **INITIAL FINDINGS AND PLANS FOR THE SUCCEEDING TERM**

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 extending the number of statements fed unto WEKA from 900 to 2000, 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 addition 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. In the event that the researchers will be given ample time in the succeeding term, they may likewise opt to develop improvements on the prototype (e.g. allow the user to type his/her own set of statements to be put to test, and allow the automated expansion of the Bag-of-Words by adding more distinct features as attributes).

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.