**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, 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: voluntary – 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, may spread across different areas, and may harm 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. Camarines Sur Rep. Rolando Andaya Jr., a lawmaker, deemed its provisions inefficient (as he later on decided to propose House Bill 5718 which further defines and penalizes the act of committing the crime of cyberbullying). 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. The results that will be generated by the model can be of great help to the moderators (of that particular social media site where it is being used) in order for them to prevent and in turn, mitigate, cyberbullying occurrences as soon as its traces become visible (in the post).

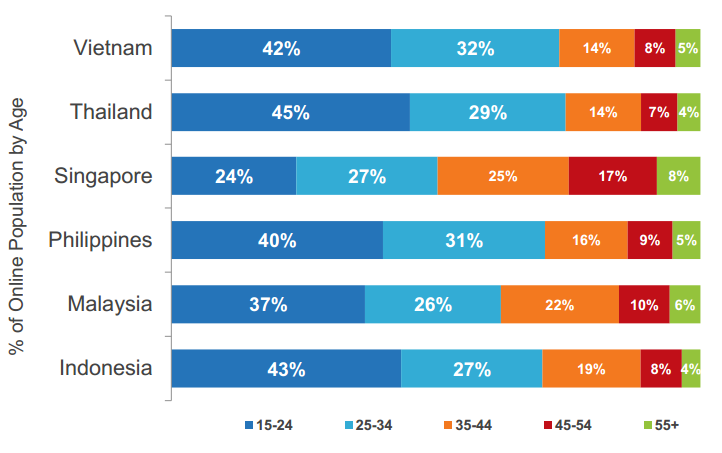


Figure 1: Southeast Asia Digital 2013 Report Survey

* 1. **Statement of the Problem**

How can Natural Language Processing (NLP) techniques be applied in the process of automating the method of detecting cyberbullying occurrences in public social media posts?

**1.3 Objectives**

**Main Objective**

This research aims to formulate a cyberbullying detection model which will yield an accuracy of least 70-80% in terms of detecting cyberbullying occurrences present in public social media posts. Results that will be derived from the model will aid social media moderators in their task of preventing the growth of cyberbullying events.

**Specific Objectives**

* To acquire ample data for the corpus
* To apply text pre-processing to the statements included in the dataset
* To extract and evaluate significant features from the corpus
* To develop classes based from the predefined categories of cyberbullying (refer to IV. Design and Methodology)
* To devise appropriate experiments which will test the cyberbullying detection model's accuracy

**1.4 Significance**

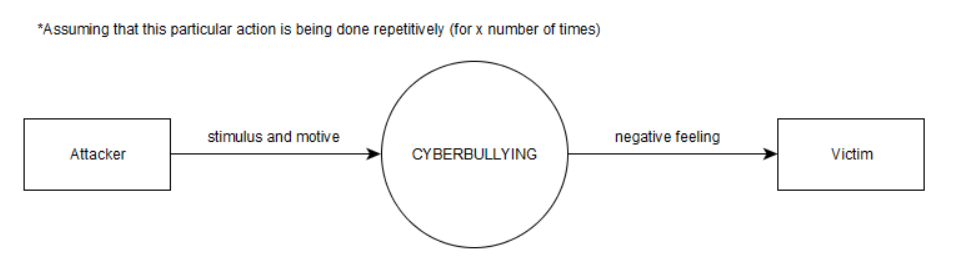
The creation of a cyberbullying detection model (which will be patterned according to selected cyberbullying statements found in social media posts bearing sensitive issues as perceived by the many) will greatly contribute to the improvement of social media monitoring here in the Philippines. As of today's 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).

As mentioned before, most people, typically Filipinos, are reluctant to admit to being victims of cyberbullying (unless it becomes fairly obvious). The current solution that social media sites are implementing will not be able to help them unless they do it (the reporting) voluntarily. Automated cyberbullying detection can greatly contribute in these scenarios (as they detect and flag such statements automatically.

**1.5 Scope and Limitations**

This research will discuss selected methods under Natural Language Processing and Machine Learning (ML) algorithms (at least 5 for NLP and 2 for ML) 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 – as how it is used in Manila - 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 two specific types of Machine Learning algorithm.

**1.6 Context Diagram**