NLP Project: Paperwork

**REVIEW OF RELATED LITERATURE**

Many researchers have given diverse definitions for the term cyberbullying. In this study, however, 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 (to the point where a specific pattern of attack had become easily recognizable by many), harmful – the deed has brought upon a 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, online games, 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”. (Hinduja & Patchin, October 2014).

Sonnie Santos, a cyberbullying expert, classified cyberbullying cases in the Philippines under two types: the cyber mob and the day-to-day under-reported cyberbullying cases. A cyber mob bears the similar characteristic of attack which Jamie Paula Salvosa, famously known as the “Amalayer” girl, whose video of berating an LRT security guard (due to a circumstance involving security checks) went viral over renowned social media sites, experienced. According to Urban Dictionary, a cybermob involves groups of people (at least two may do) holding another person accountable for a misdeed or social blunder (that he/she had committed) within the cyber environment. On the other hand, day-to-day cyberbullying cases that are not formally reported include those that are coming from people whom the victim maintains connections or interactions with. Likewise, this particular bullying scenario occurs through social networking sites. (What Can You Do When You're Cyberbullied?, 10 October 2013).

The Anti-Bullying Act of 2013 (R.A. 10627) covers bullying in totality (physical, social, verbal, and cyber) for elementary and secondary schools only. It requires schools to come up with specific policies defining and prohibiting bullying inside or outside school premises, and retaliation against people reporting bullying incidents. They must also think of admin procedures and disciplinary actions (with regard to the intensity of the attack), rehabilitation procedures for the bullies, and strategies and procedures for recording and reporting the incident, with the counseling of the victim and the educating of his/her parents regarding the situation inclusive. Only the principal of the school (or any individual with comparable role or assigned by the latter to represent him/her) can implement the said law and handle the bullying cases. Six months after the implementation of this law, every school is expected to have made the new policies public by means of a written document and reported the said rules to their respective division superintendent. They are required to do the said report annually. (Republic Act No. 10627, 13 September 2012).

Camarines Sur Rep. Rolando G. Andaya, Jr., a Bicol lawmaker who proposed the “Anti-Cyberbullying Act of 2015” said that, “By penalizing acts of cyber-bullying, people are encouraged to become responsible netizens and make them accountable for their cyber-actions”. He clearly stated in his proposal the forms of cyberbullying that will be reprimanded under HB 5178 such as repetitively sending rude messages towards the victim, disclosing derogatory information about the victim, posting or sending offensive photos of the victim, breaking into the victim’s personal accounts and using it to commit actions that may harm others or the victim himself/herself, and repeatedly sending messages that threatens the personal security of an individual. The doers of the aforementioned actions will be fined with 50 000 – 100 000 pesos and charged with 6 months up to 6 six years of imprisonment as penalty. (Anti Cyber-Bullying Act of 2015 Pushed, 27 May 2015).

In Japan, Niita, Masui, Ptaszynski, Kimura, Rzekpka, and Araki proposed a system on Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization. The researchers proposed three methods in conducting their system: Phase Extraction, Categorization and harmful word detection together with harmfulness polarity determination and Relevance maximization. Phrase extraction involves defining phrases which contains words that gain harmful meaning when used in a specific context, or in a combination with other words. Through phrase extraction, words that are used in dependency relation that contains harmful meaning are also extracted. In the process of harmful word detection and categorization, they detect words that conveys harmful meaning and categorized them into three categories: obscene, violent and abusive. In the method of maximization of relevance score, they calculate harmfulness polarity score of phrases with each seed word for all three categories. All entries are sorted beginning with the one with the highest harmfulness score.

Reynolds, Kontostathis, and Edwards conducted a research on Using Machine Learning to Detect Cyberbullying. Machine learning is the process of training a computer to predict a label using a set of attributes and a truth set. First, the researchers obtained their data through a web crawler in a subset of the Formspring.me site and extracted information from the sites of 18,544 random users. They used Amazon’s Mechanical Turk service to identify whether each post can be considered as cyberbullying. The researchers also identified a list of insult and swear words, posted on the website www.noswearing.com. The list contains 296 terms and each word was given a severity level by their team. The levels were 100, 200, 300, 400, and 500. They also extracted two different training sets, the count information and the one which contains the normalized information. The process of count information involves counting the totality of each word in a sentence while the normalization involves dividing the number of words at each severity level by the total number of words in the post and then multiplying by 100 to get an integer value. The team used 4 algorithms for their project: J48, JRIP, IBK, and SMO. The J48 option is used to generate a decision tree model from the attributes provided. JRIP is a rule based algorithm that creates a broad rule set then repeatedly reduces the rule set until it has created the smallest rule set that retains the same success rate. The instance-based (IBK) algorithm was used with the IBK method with k=1 and k=3. Lastly, the SMO algorithm is a support vector machine that the team used in order to test their research.

Chen, Zhu, Zhou, and Xu conducted a research on Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In their research, the team proposed the Lexical Syntactical Feature (LSF) approach to identify offensive contents in social media and to predict a user’s potentiality to send out offensive contents. First, the team conceptualized the notion of online offensive contents and distinguish the contribution of profanities and obscenities in determining offensive contents, and introduce hand authoring syntactic rules in identifying name-calling harassment. Second, the team also incorporated style features structure features and context-specific features to predict a user’s potentiality to send out offensive content in social media.

Van Hee et al, conducted a research on Automatic Detection and Prevention of Cyberbullying. First, the researchers constructed a corpus by collecting data from the website www.ask.fm. They used GNU Wget, a web crawler software in order to gather their data. The team also developed an annotation scheme for the analysis of textual cyberbullying. The annotation scheme describes two levels of annotation by assigning a harmfulness score to the post on a three-point scale, with 0 signifying that the post does not contain indications of cyberbullying, 1 that the post contains indications of cyberbullying although they are not severe and 2 that the post contains serious indications of cyberbullying. The team also defined two types of bystanders in the annotation scheme: the bystander defenders, who help the victim and discourage the harasser from continuing his actions and bystander assistant, who indirectly take part in the action of the harasser. The team also categorized each words as Threat, Insult, Defense, Sexual Talk and Threat range from moderate to substantial. The team also developed bag-of-word which represents their corpus as a set of word or character sequences.

Dinakar, Jones, Havasi, Lieberman, and Picard (2012, September) conducted a research on Common Sense Reasoning for Detection, Prevention and Mitigation of Cyberbullying. The team established support-vector machines as one of their methods in text categorization. The team used two datasets: Youtube and Formspring. The dataset from Youtube was obtained through a web scraper in different comment sections. The team categorized the comments in three labels: Sexuality, which involves attacks on sexual minorities and sexist attacks on women. Race and culture, which involves attacks on racial minorities, cultures and mocking of cultural traditions. And intelligence, which attacks the mental capacities of a person. The Formspring dataset contained instances that were flagged as inappropriate by other users and topics pertaining to sexuality. Removing of stop-words and tokenizing the text to separate words from punctuation marks were performed through the use of NLP. The team also selected four canonical concepts, the affective valences positive and negative, as well as gender (male and female). The plan is to compare each extracted concept for similarity with each of the canonical concept to get an overall measure of how similar the given comment is to each of the canonical concept. Each comment is subjected to the process wherein it yields similarity scores for the canonical concepts of good, bad, boy, and girl.