NLP Project: Paperwork

**REVIEW OF RELATED LITERATURE**

**Specific Definition of Cyberbullying**

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).

**Common Scenarios of Cyberbullying in the Philippines**

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).

**Efforts Contributed by the Filipino Government to Stop the Occurrence of Cyberbullying**

**Anti-Bullying Act of 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).

**Anti-Cyberbullying Act of 2015**

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).

**Natural Language Processing: The Basics**

Natural Language Processing (NLP), in its simplest sense, refers to the utilization of a computer’s capability to process sentences stated in natural human languages (such as English, Japanese, Korean, etc.) instead of specialized computing languages (Java, C++, C#, etc.). However, NLP still holds significant differences between natural human languages. The term “natural language” actually refers to a stricter subset of the said human languages. They do not function the same way as how people carry-on with their everyday conversations. In fact, a writer once said that “human languages allow anomalies that natural languages cannot allow” (Steedman, p. 229). The phrase “natural language processing” can cover a wide array of topics involving speech recognition, context reference issues, discourse planning and generation, and syntactic and semantic analysis and processing. The term “processing” in the similar phrase is not restricted to “interpretation” only. It pertains to a combination of both “interpretation” and “generation”, which is why systems possessing the said capability (NLP) are not limited to getting a grasp of the natural languages but are also given the ability to produce their own. 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. Additionally, computer systems that can convert human languages to computer languages and vice-versa are already existing. They provide assistance to people with disabilities such as blindness in terms of using personal computers. Translation programs were also made possible by NLP. There are applications in the market having the capability to translate one human language to another. Likewise, grammar and spelling checkers are also programmed following the NLP algorithm. Lastly, a computer that can read human languages (in publications such as books) is also a product of natural language processing. With respect to the interpretation side, NLP involves the following subtopics: signal processing, syntactic analysis, semantic analysis, and pragmatics. Signal processing is basically acquiring spoken words and turning them into text. Syntactic analysis, on the other hand, focuses on the grammar composition of the sentences. Semantic analysis deals with how the words and sentences refer to their corresponding worldly elements, or rather their own meaning in a particular spoken human language. Finally, the term pragmatics concerns how a specific word functions in an individual’s day-to-day conversations with one another. (Introduction to Natural Language Processing, 2006).

Public information dissemination across the Internet triggered the emergence of research areas under the field of NLP such as Information Extraction, Information Retrieval, and Question-Answering. (Mendes & Antunes).

**Related Studies**

**Automatic Detection and Prevention of Cyberbullying - An Application of Natural Language Processing**

**Overview**

This study, formulated by the members of the Language and Translation Technology Team (LT3) and Computational Linguistics Group (CLiP) of Belgium, pertains to the creation of a system which would make the automatic detection of cyberbullying trends in every social media content (in the form of text) possible by means of a fine-grained annotation of the corpus - a large collection of texts. In order for the system to appropriately distinguish non-cyberbullying from cyberbullying patterns, the researchers initially devised their own dataset. This would serve as their basis for examining the characteristics of a particular content. Fine-grained categories related to cyberbullying (such as threats and insults) are the ones that are being considered during the system's annotation process. Through this, the type of cyberbullying scenario that took place (e.g. threat or blackmail, general insult, curse inclusion, etc.) at that current time and its corresponding severity according to how the statement was used in the context will be determined. Additionally, the roles of the people involved in the similar cyberbullying event (e.g. bully, victim, and bystander) will be identified. According to the researchers, the said procedure should also be considered due to the fact that there is a significant difference between the victim being helped by a bystander and a victim being forced to face the situation alone. Furthermore, messages that will undergo the process of detection in this study is written in Dutch. The system, when implemented, is expected to fulfill the purpose of being the "first filter" in social networking sites, with the hope of reducing the amount of messages that will be directed to the human moderators.

**Data Collection Process**

Ask.fm ([http://ask.fm](http://ask.fm/)), a social networking site, was used by the researchers as their platform for collecting ample data (which would then be used to construct the corpus) by means of simulating cyberbullying scenarios with the help of a few adolescents (who volunteered to participate). The participants were likewise asked to provide several textual evidences of cyberbullying scenarios that they can find across various social networking sites. GNU Wget software (<https://www.gnu.org/software/wget>) retrieved the data from Ask.fm by crawling seed sites. After the execution of the former process, the devised corpus is expected to be written in Dutch (all of it). However by this time, non-cyberbullying and cyberbullying trends remain combined.

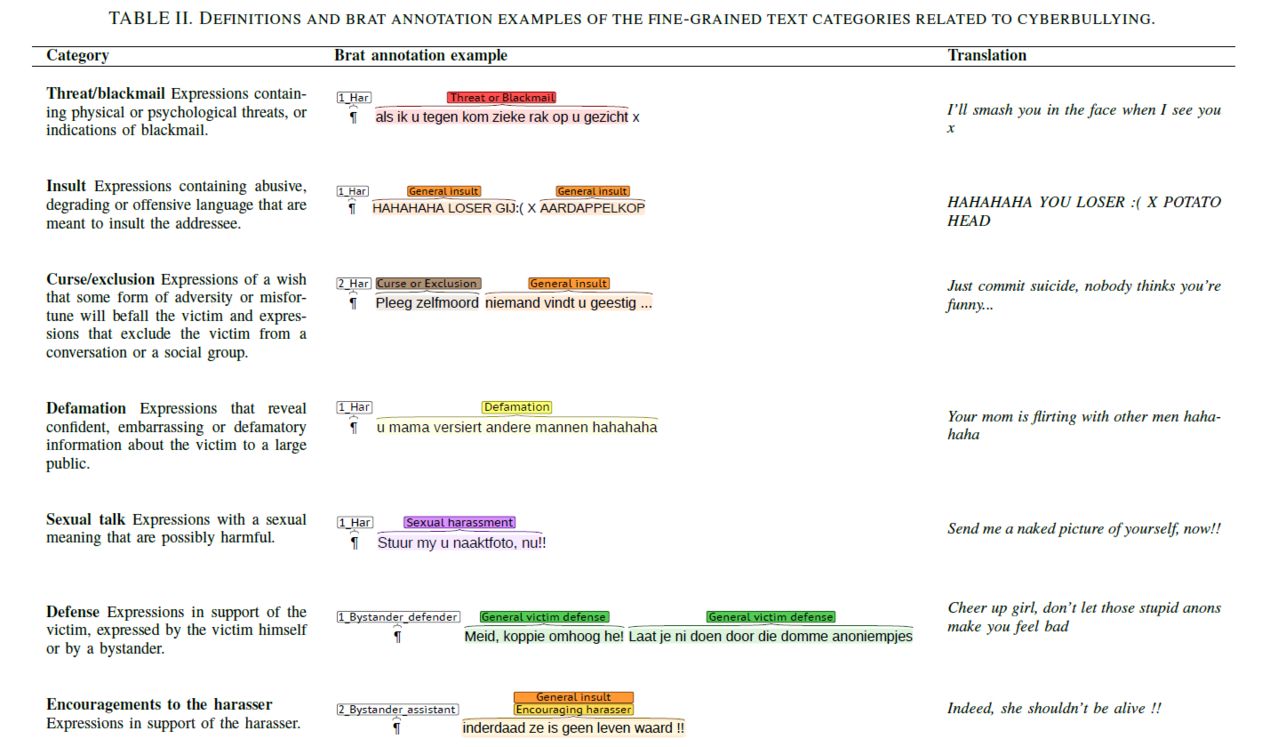
**Data Annotation**

In this phase, cyberbullying patterns will be differentiated from non-cyberbullying ones with the help of a fine-grained annotation scheme. It consists of two levels - with identifying whether a post suggests a cyberbullying event and evaluating its intensity with the help of a three-point scale (with 0 signifying that the post contains no traces of cyberbullying trends, 1 indicating the presence of mild cyberbullying, and 2 representing serious cases of the said type of bullying being present) as the first one, and distinguishing the role of the person who posted that particular cyberbullying related statement (whether he/she is the bully, victim, or bystander) as the second. Likewise, the posts that would fit the defined category were annotated by two annotators. Kappa was used to calculate the scores. The numerical results will then be used to classify the detected cyberbullying trend into categories such as *Threat*, *Insult*, *Defense*, *Sexual Talk* for scores that are found between 0.52 - 0.66 and *Defamation*, *Encouragements*, and *Curse*, on the other hand, for scores lower than 0.52.

**Experimental Setup**

The researchers made use of an algorithm known as the Support Vector Machines (SVM) for classifying statements under the non-cyberbullying or cyberbullying category. This particular algorithm works well for highly-skewed textual classifications similar to the ones being investigated. Additionally, they also utilized linear kernels (into which the value of *c* is equivalent to 1). All of the experiments done in this study were carried out by Pattern. In addition to that, pre-processing steps such as tokenization, PoS tagging, and lematization were applied by means of the LeTs Preprocess Toolkit. The system being described in this research has to be "learning", therefore it will take a set of default instances and follow a particular model (based on those instances) in order to accommodate future errors or additional words. To be able to make the model construction possible, instances were represented as vector of features which contained the useful information for distinguishing potential cyberbullying content from non-cyberbullying ones. Two types of lexical features - Bag-of-Words and polarity - were implemented by the researchers during the actual experimentation. This was made possible by the following: word unigram and bigram Bag-of-Words - which indicates the presence of a single word or sequence of two words, character trigram Bag-of-Words - a sequence of three characters, and sentiment lexicon features - includes the number of positive, negative, and neutral lexicon words in the text (averaged over text length), and overall post polarity (the sum of the values of identified sentiment words, averaged over text length).

**Sample Result**



**Experts and Machines Against Bullies: A Hybrid Approach to Detect Cyberbullies**

**Overview**

The main concern of this study is the identification of potential cyberbullies based on the analyses that will be done on the statements an individual posted over the internet. Such analyses may include consideration of writing style and possible network activities of the person of interest. Basically, the prevention of cyberbullying will be executed by means of discovering the attacker's profile by assigning scores to all social media users with regard to their level of "bulliness" or possible misbehavior based on the history of his/her activities.

**Data Collection Process**

The researchers conducting this study constructed their dataset with Youtube as their corpus. The acquisition of data occurred specifically on the comment section of Youtube videos - which experts claimed to have been one of the major platforms cyberbullies make use of in harming their victims.

**Using Machine Learning to Detect Cyberbullying**

**Overview**

Identifying the presence of cyberbullying through machine learning is the main purpose of this study. The aforementioned process will make the analyzation of the bully's language patterns possible. Datasets, which contained data that were taken from a webcrawl of Formspring.me - a popular website among teenagers and college students in Pennsylvania, USA, were also utilized by the researchers of this particular study in order to provide assistance to the acquisition of precise results. Lexical features (extracted from Formspring.me), along with a few data mining algorithms (which were made available by using a tool called Weka toolkit) were used in the development of the model for this cyberbullying detection mechanism. Likewise, the system willingly accepted false positives in order to increase the number of true positives that could be accepted by the same system.

**Data Collection Process**

Formspring.me is somewhat similar to Ask.fm specifically in terms of the type of SNS to which they both belong to (a question-and-answer based website), and the admittance of anonymity among all of their users - making the two sites more vulnerable to cyberbullying (the main reason as to why these sites were chosen). In obtaining the required information, a particular number of users were chosen at random. These users are the ones who were "crawled" to obtain the 10 files needed for the training set and another 10 for the testing set. With the use of Amazon's Mechanical Turk service - a website where requestors are free to post any type of task which the workers can accomplish for a fee, labels were determined from both of the constructed truth sets.

**Inputs of Interest**

The researchers avoided the use of the Bag-of-Words (BoW) approach in the extraction process of the distinguished cyberbullying statements due to reasons concerning their own convenience. Instead, they identified a list of "bad" words (mostly made up of swears and insults) posted on www.noswearing.com. The latter were able to download a total of 296 terms - each having their own intensity levels (with regard to cyberbullying) defined by the same people. Since the number and the density of the amount of these "bad words" are the sole focus of the said research, researchers further obtained two more training sets - the first one containing the count information, and the second one holding the normalized information. The normalization of the words took place after they divided the number of words (at each severity level) by the total number of words in the post, and then multiplying the answer by 100 to get an integer value. A feature that the researchers used in this study is called the SUM feature. It can be calculated by means of taking a weighted average of the "bad" words. These features would then be included in NUM and NORM versions of the company's datasets.

**Products of NLP**

**Software Applications**

The following applications were developed based on NLP: Thinknowlogy – a grammar-based software, making use of the Natural Laws of Intelligence in grammar to be able to develop a certain intelligence through natural language in software, Bitextor – an application bearing the objective of generating translation memories with the help of multilingual Web sites as a source, TagAligner – a software used to generate translation memories from two XHTML tagged files, RelEx - is an English-language semantic dependency relationship extractor that can identify subject, object, indirect object, etc. between a group of words in a sentence, and Wintermute – a framework of applications using neural networking to learn about its host. (13 Projects Tagged "Natural Language Processing", 2015).

**Application Programming Interface (API)**

An API, as defined by Webopedia.com, involves routines, protocols, and tools for creating software applications, specifying how these software components should communicate. The following are some examples of NLP APIs:

* **Text Razor** - an API that helps a user extract the Who, What, Why, and How from textual documents such as research papers, survey verbatims, legal documents, and the like.
* **Stremmor Automated Summary and Abstract Generator** - formulate summaries through the process of extraction while at the same time, maintaining readability by keeping related words in a sentence intact.
* **Skyttle 2.0** - an API that extracts topical keywords and sentiments from a text.
* **Textuality** - has the capability to find relevant pieces of information on webpages.