Project Members: Samantha Mallari | Faith Ballesteros | Eva Samillano | Paulo Capuz | Jason Juarez

Section: SS141

Project Title: Automated Detection of Cyberbullying Occurrences in Social Media Posts Using Support Vector Machine (SVM) Algorithm

1. Background/ Project Rationale Literatures

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. Since NLP is a broad field of study, the team merely focused on the concept of text classification. It is the classification of documents into a fixed number of predefined categories based on their content. 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. For the text classification task, the team will implement a Support Vector Machine (SVM) which is a supervised machine learning algorithm that can be employed for both classification and regression purposes. It is often 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.

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.

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.

1. Current Situation

Since social media sites are vulnerable to cyberbullying attacks, some of them implemented techniques in order to address the issue of cyberbullying. To illustrate, Youtube’s safety mode, once activated, can hide all comments that contains harmful content from users. And on Facebook, users can add comma separated keywords to the “Moderation Blacklist” so when people include blacklisted keywords in their post or comment, it will be automatically identified as spam. Most social networking sites use lexicon-based approach to filter vicious posts. Furthermore, they rely solely on users to report offensive contents to take actions. These systems are proven to have low accuracy and may generate many false positive alerts due to their use of simple lexicon-based automatic filtering approach to block harmful posts and comments.

1. Description of the Problem/ Opportunity

As social media become widespread it has also brought a societal problem that deserves utmost attention and must be addressed immediately – cyber bullying. It is a form of harassment that occurs via Internet which involves vicious forum posts, name calling in chat rooms, creating fake profiles, and mean email messages. As the number of social media users tremendously increases, it consequently intensifies the cyber bullying problem. Legal and ethical issues are now confronting the abuse of using the social media to bully other internet users. This is why the development of automated systems is essential to help monitor activities occurring in the cyberspace. Successful detection of cyberbullying occurrences in the Internet is therefore of key importance to identify and classify possible threatening situations online and prevent them from escalating.

As cyberbullying becomes rampant: authorities and researchers proposed and implemented different solutions to mitigate this problem. These methods involved manual review tasks of identifying offensive contents, integrating lexicon-based approach to filter offensive contents, communication model to track online predators, lexical features with machine learning classifiers to determine the victim and a bully in each instance of cyberbullying, profiling features to detect aggressive discussions and others. However, some of these methods were proven to be inefficient because they rely merely on the user to report cyberbullying incident before they take actions. It is also not possible for the moderator to detect all cyberbullying occurrences in social networking sites. To address this limitations, the team decided to create a cyberbullying detection model that can generate up to 70-80% accuracy in detecting vicious online contents. Moreover, the target market of the present study are researchers, Filipino social media users, adolescents, parents, and students of APC.

1. Proposed Solutions

Cyberbullying has become a pervasive problem around the world and it is tremendously alarming. However, given the massive information on the Web, there is a need for intelligent systems to identify potential risks automatically. Thus this leads to the creation of a cyberbullying detection system by deploying Support Vector Machine (SVM) as a text classification tool. The system will be able to detect cyberbullying statements written in both English and Tagalog. The system aims to yield an accuracy of 70-80% in terms of detecting correct occurrences of cyberbullying. After the detection process, the system will procure a report which contains the user and violation details. Sanctions may vary depending on its use. To illustrate, in the school a student who will be caught doing the act of cyberbullying may be apprehended as per what is stated on the student’s handbook, etc.

1. Significance of the project

The main significance of this research project is aimed towards the field of Computer Science. The proponents of this study made certain that some important concepts such as text classification and machine learning were included and utilized in their research. 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 Text Classification. As a result, 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.

*For the Filipino social media users* - Most people, typically Filipinos, are reluctant to admit to being victims of cyberbullying (Andrade, 2012). In order to address these issue, a cyberbullying detection model will be designed to detect even subtle posts implying cyberbullying attacks as much as possible. 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). 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). 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. 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.