NLP Project: Paperwork

**DESIGN AND METHODOLOGY**

**Data Collection and Annotation**

The availability of a suitable data plays a crucial role in research on cyberbullying. Moreover, building a suitable dataset is needed for building representative models for cyberbullying detection. This section describes the construction of a corpus of social media posts which contains both cyberbullying and non-cyberbullying content.

A. Data Collection

We constructed a corpus by collecting data from different social networking sites such as Facebook and Twitter. Through web scraping, we were able to obtain data for our dataset. The web scraping tool that we used for obtaining our data was [**Import.io**](https://import.io/). It is a powerful and easy-to-use tool for data extraction that has the aim of getting data from any website in a structured way. [1] In total, we were able to collect 1000 posts from Filipinos in their own social networking sites account.

B. Data Annotation

There will be two levels of annotation. The first is the level of the message or the post itself. By taking the entire message into consideration, annotators will define as a first step whether it contains cyberbullying content or not by indicating the harmfulness of the message. When the message is considered harmful and contains indications of cyberbullying, annotators should indicate the role of the author of the post. At the second step, text spans with relevant information to the use case of cyberbullying will be identified and categorized. All annotations are performed using brat, a web-based tool for text annotation. It is designed specifically for structured annotation, where the notes are not freeform text but have a fixed form that can be automatically processed and interpreted by a computer. [2]

For each post, annotators define whether the post contains indications of cyberbullying and whether this indications are severe through assigning a harmfulness score to the post on a three point scale. At level 0, the post does not contain indications of cyberbullying. At level 1, the post contains indications of cyberbullying however it is not severe. Lastly, at level 2, the post contains severe indications of cyberbullying.

The annotators also identified the role of the participants in a cyberbullying scenario. Van Hee et al distinguished four different roles in a cyberbullying context in their research which we adopted for our data annotations as well. These roles are: harasser, victim, bystander-defender, and bystander-assistant. These roles must be identified when the harmfulness score is equal to 1 and 2. The harasser is the person who initiates the cyberbullying attack. The victim is the one being attacked. The bystander-defender helps the victim and discourages the harasser from initiating an attack. Lastly, the bystander-assistant serves as an ally of the harasser.

At the second level of annotation, data will be classified into one of the five categories: Threat, Insult, Defense, Sexual talk, and Curse.

C. Experimental Corpus

For our preliminary experiments, we focused on the dataset that we gathered from Facebook because large majority of posts coming from it were considered as occurrences of cyberbullying. With regards to the occurrences of the fine-grained categories, we observe that Curse is the most frequent type of cyberbullying activity in our corpus, followed by Insults, Threats and Sexual talk. Defense is the least represented in the category.

For each category, the number of instances marked with a harmfulness score of 0, 1, and 2 is given. 280 curses were identified in a non-cyberbullying context. If we consider the different roles in the annotated bullying events, we figured out that the role of bully features in more than half of the annotated instances, followed by the victim role in about 20% of the instances. The bystander role in its two different sub roles yields 15% of the experimental corpus.