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

The team constructed a corpus by collecting data from two popular social networking sites such as Facebook and Twitter. The team chose to extract information from these sites because they are more prone to cyberbullying trends since they allow users to post anything. Through web scraping, they were able to obtain the data for their corpus. Moreover, the team decided to use import.io as their primary tool for this process. 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, they were able to collect 4500 posts from Filipino users that were chosen randomly.

B. Data Annotation

There will be two levels of annotation. The first step is to identify whether the message contains indications of cyberbullying. When the message contains cyberbullying content, the annotator must indicate the harmfulness range of the post, which ranges from 0 to 2. Now, when the post scored 1 or 2, the annotator needs to indicate the role of the author of the post as well. 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 General Architecture for Text Engineering (GATE). GATE includes an [information extraction](https://en.wikipedia.org/wiki/Information_extraction) system called ANNIE (A Nearly-New Information Extraction System) which is a set of modules comprising a [tokenizer](https://en.wikipedia.org/wiki/Lexical_analysis), a [gazetteer](https://en.wikipedia.org/wiki/Gazetteer), a [sentence splitter](https://en.wikipedia.org/wiki/Sentence_boundary_disambiguation), a [part of speech tagger](https://en.wikipedia.org/wiki/Part-of-speech_tagging), a [named entities](https://en.wikipedia.org/wiki/Named_entity_recognition) transducer and a [co reference](https://en.wikipedia.org/wiki/Coreference) tagger. [1]

For each post, annotators define whether the post contains indications of cyberbullying and whether these 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. Threats or blackmails are expressions which contain intentions to inflict pain, injury, damage, or other hostile action on someone in retribution for something done or not done. Insult contains abusive and offensive language that are intended to degrade someone. Defense contains expression in support of the victim and it can be expressed by the victim himself or the bystander defender. Lastly, Curse involves offensive phrases used to express anger or annoyance towards someone.

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.

**Sources:**

[*https://en.wikipedia.org/wiki/General\_Architecture\_for\_Text\_Engineering*](https://en.wikipedia.org/wiki/General_Architecture_for_Text_Engineering)

*https://en.wikipedia.org/wiki/Import.io*