**DESIGN AND METHODOLOGY**

In this section, the researchers illustrate their understanding on the process of creating a cyberbullying detection mechanism.

**The Creation of the Corpus**

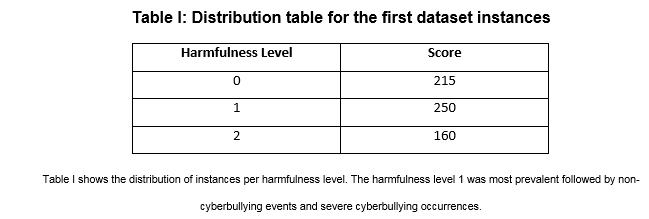
**Data Collection and Annotation**

The team acquired data from Youtube. A total of 600 comments (both cyberbullying and non-cyberbullying) were obtained and evaluated according to their corresponding level of harmfulness (0 - for comments containing no cyberbullying occurrences, 1 - for comments expressing indirect cyberbullying content, 2 - for comments explicitly demonstrating severe cyberbullying events). Those statements which were labeled 1 or 2 were further classified into four categories - Physical Appearance, Race and Culture, Intelligence, and Social Rejection - which were deemed related towards controversial issues in the Philippines.

Moreover, the team decided to use Import.io as their primary tool for the annotation 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. The YouTube dataset contains comments from videos on controversial events in the Philippines that were considered controversial as these topics would most likely to encourage viewer to post their opinions about it. Most of the cyberbullying statements that the researchers obtained came from YouTube. In Facebook, they collected posts from different university secret pages, wherein a user can post his/her story about controversial school experiences. In Twitter, posts from random Filipino people were collected.

The dataset was cleaned in order to remove unnecessary symbols present in text. These symbols were removed as so they will not be extracted as features later on. Furthermore, the dataset underwent normalization using Special Text Replacement function in Excel.

One of the researchers was in charge of annotating the dataset. The annotator simply tagged each instance based on how she perceived it. All annotations are performed using General Architecture for Text Engineering (GATE) and Brat Rapid Animation Tool (BRAT). GATE is a tool wherein a user can easily input an annotation schema for entities. Brat, on the other hand, is an intuitive web-based tool for text annotation supported by Natural Language. These tools are open source software.



At the second level of annotation, the following guidelines were given:

**Sexuality** – instances that contains both vulgar and filthy words.

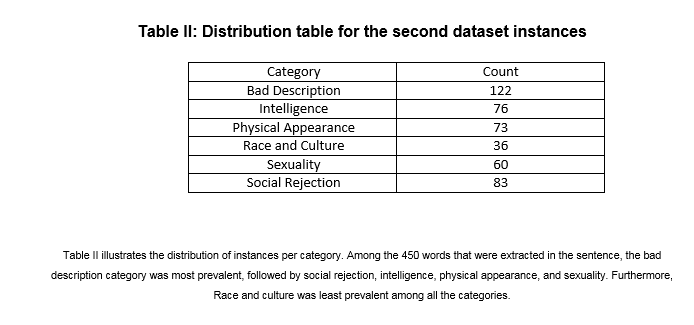
**Physical Appearance** – instances that contains cyberbullying statements with regards to the physical characteristics of a person.

**Intelligence** – instances that attacks the mental capacity of a person.

**Race and Culture** – instances that contains both racial and cultural discrimination.

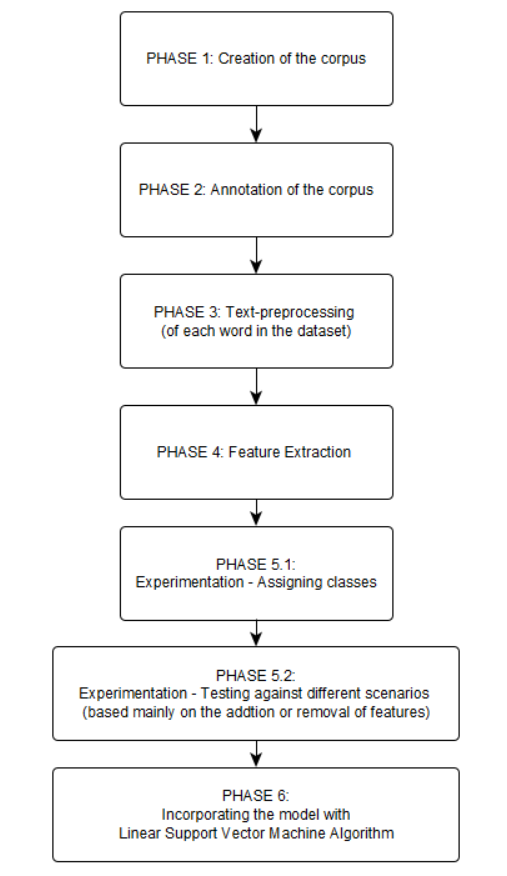
**Social Rejection** – instances that contains cyberbullying statements which isolates an individual from the society or it degrades a person.

**Bad Description** – instances that contains cyberbullying statements that pertains to the attitudes and behavior of an individual. However, this categorization can be misused with the Physical Appearance.



In the succeeding phases of the project, the researchers are planning to continue with the text pre-processing tasks that are needed in order to lessen the difficulty of extracting linguistic features from the text (which is the phase of the project that will come after the pre-processing phase). Feature selection will be implemented using the following methods: Bag of words – which will yield bigrams and trigrams, and application of TF-IDF scores – which will aid in the process of determining the most relevant words in each cyberbullying category, as these scores will be used to evaluate the bigrams and trigrams of the instances resulting from the Bag of words approach (the higher the score, the more relevant the term is). Profane words and words written in uppercase letters are going to be considered offensive by default.

The initial part of the experimentation phase begins with the establishment of classes involving the predefined classifications of cyberbullying expressions based on sensitive issues. This is to verify the accuracy of the model in terms of functioning side by side with features that are closely related to each other. This process will also determine which words can fall in more than one class so they can be given appropriate weights (**feature weighting**) – e.g. a particular word is more appropriate to be classified in this class compared to the other despite having the characteristics if being included in the latter class as well. The second part of the experimentation phase will take into consideration the final features that will be selected (or the most determining words per each cyberbullying classifier based on offensive issues). The group may either remove or add more features depending on the results the experiments will yield regarding the cyberbullying detection model’s accuracy.



Proposed plan of activities for the accomplishment of the project