Course Project: Cyberbullying Classification

Ensemble Learning

Caterina Conz, Yago Alessandro Bardi Vale, Nevina Dalal, Vivian Koutroumani

January 17, 2022

Abstract

As social media usage becomes increasingly prevalent in every age group, a vast majority of citizens rely on this essential medium for day-to-day communication. Social media's ubiquity means that cyberbullying can effectively impact anyone at any time or anywhere, and the relative anonymity of the internet makes such personal attacks more difficult to stop than traditional bullying.

On April 15th, 2020, UNICEF issued a warning in response to the increased risk of cyberbullying during the COVID-19 pandemic due to widespread school closures, increased screen time, and decreased face-to-face social interaction. The statistics of cyberbullying are outright alarming: 36.5% of middle and high school students have felt cyberbullied and 87% have observed cyberbullying, with effects ranging from decreased academic performance to depression to suicidal thoughts.

In light of all of this, this dataset contains more than 47000 tweets labelled according to the class of cyberbullying:

- Age;
- Ethnicity;
- Gender;
- Religion;
- Other type of cyberbullying;
- Not cyberbullying

The data has been balanced in order to contain 8000 of each class.

1 The dataset

1.1 How to use the dataset

We have two main objectives set for this project:

1. Create a multi-classification model to predict cyberbullying type.
2. Create a binary classification model to flag potentially harmful tweet.
3. Explore words and patterns associated with each type of cyberbullying.

1.2 How to obtain the dataset

You have to register yourself in `this link` to get access to the data.

1.3 Data preprocessing tasks before starting the project

For the first classification task (multi-class classification)
We have only two columns: the tweets and the classes.
1. We will have to perform specific text preprocessing tasks, like removing stopwrods and punctuations, stemming and lemmatization. We will also have to build word clouds or other visual techniques to highlight key words in each type of cyberbullying.
2. After cleaning the data we will have to perform feature engineering and extract features that would help us in the classification part of the project. This includes extracting features like length of the tweet, vowels and the TF IDF matrix which is used to find the most relevant words that associate with each class.

For the binary classification:

1. We use the data pre processing steps 1 and 2 to find the features.

2. Based on these features and our own metrics we will create a binary variable which will be set to 1 if the tweet is harmful, and 0 otherwise.

1.4 Tasks for the project

After this preprocessing step, we now use the features to build state-of-the-art ensemble methods. The project guidelines are:

1. Apply all approaches taught in the course and practiced in lab sessions (Decision Trees, Bagging, Random forests, Boosting, Gradient Boosted Trees, AdaBoost, etc.) on this data set. The goal is to predict the target variable **cyberbullying_type**.

2. Compare performances of all these models.

3. Conclude about the best model for this classification task