Project Proposal: Text Classification of Cyberbullying on tweets

The project I propose focuses on trying to classify a tweet into several type of cyberbullying or to classify the tweet as not containing any form of cyberbullying.

**To obtain the data**

You have to create an account on [www.kaggle.com](http://www.kaggle.com), and then you can download the dataset at [this link](https://www.kaggle.com/andrewmvd/cyberbullying-classification). The tweets will need to be cleaned and preprocessed first.

**Handling**

The data appears in CSV format but isn’t very well represented. Each line of the file is a data points, where everything before the last comma is the tweet and the classification is the last word that immediately follow the last comma. Once properly separated, we can handle the data with something like the Pandas library. As mentioned, the data is composed of two part:

1. The tweet (up to 280 character)
2. The class (Not cyberbullying, Age, Ethnicity, Gender, Religion and Other)

The aim of this project is to use the first to predict the second.

**Data preprocessing**

We will preprocess the tweets to extract some statistics about the data using NLTK (Natural Language Toolkit) and extract TF-IDF features with sklearn.

1. Clean all mentions, remove URL, Emojis and Smileys.
2. Remove stopwords (“The”, “will”, …)
3. Collect all hashtags, extract some statistics (tweet length, …)
4. Finally, use sklearn to extract TF-IDF matrix (or do it manually…)

**Task**

The main focuses of the project will be to implement from scratch an efficient Random Forest Classifier. We will compare the performance of our implementation with the one from scikit learn, in terms of speed and options available. The report will mention the accuracy obtained with the implementation and the time it took to train the model, depending on multiple parameters. The code for the implementation will be provided alongside the report.

Killian Susini