**REPORT TIPS**

* Concise and clear rather than discursive style
* Parts:
  + Problem statement and hypothesis
  + Description of your data set and how it was obtained
  + Description of any pre-processing steps you took
  + What you learned from exploring the data, including visualizations
  + How you chose which features to use in your analysis
  + Details of your modeling process, including how you selected your models and validated them
  + Your challenges and successes what didn’t work and why
  + Possible extensions or business applications of your project
  + Conclusions and key findings
  + Appendices: data + code
  + References
* Specify the roles and the tasks performed of every member of the group

1. PROBLEM STATEMENT AND HYPOTHESIS

We have decided to take part to a competition launched by the [kaggle](https://www.kaggle.com/) platform, regarding the classification of comments posted online in blogs or discussions according to different levels of offense, in order to address the problem of free expression of personal opinions online and promote civil and effective conversations on the web as well as stop people from leaving online discussions because of lack of respect. The aim of the competition is to study negative online behaviours by looking for a supervised machine learning model able to accurately assign comments to different categories of rudeness or toxicity, such as threats, insults, obscenity.

This is a natural language processing (NLP) problem, since it requires to understand texts, its meaning and its structure, and to come up with new insights about huge amount of textual data. Our idea has been to perform some exploratory analysis on the dataset, taken from Wikipedia’s talk page edits, and pre-process the dataset in order to prepare it for the model training and test, mainly focus on the transformation of text comments to a matrix of numeric vectors representing each of the comments. Once the input for our models is defined, we will show the training and test of some machine learning models and their evaluation to understand the performances.