# Hate-speech and offensive language on Twitter

### M2 Individual Assignment 2 - Working with Natural Language

Daniel Hain, Roman Jurowetzki

Deadline: 28/10/2020 - 23:55

This assignment is less structured than previous individual assignments.

You are given a collection of approximately 25k tweets that have been manually (human) annotated. `class` denotes: 0 - hate speech, 1 - offensive language, 2 - neither

`https://sds-aau.github.io/SDS-master/M2/assignments/data/twitter_hate.zip`

### 1. Preprocessing and vectorizaion.

Justify your choices and explain possible alternatives (e.g. removing stopwords, identifying bi/tri-grams, removing verbs or use of stemming, lemmatization etc.)

- Create a bag-of-words representation, apply TF-IDF and dimensionality reduction (LSA-topic modelling alternatively simply PCA or SVD) to transform your corpus into a feature matrix.

### 2. Explore and compare the 2 "classes of interest" - hate speech vs offensive language.

- Can you see differences by using simple count-based approaches?
- Can you identify themes (aka clusters / topics) that are specific for one class or another? Explore them using, e.g. simple crosstabs - topic vs. class and to get more detailed insights within-cluster top (TF-IDF) terms. (This step requires preprocessed/tokenized inputs).

### 3. Build an ML model that can predict hate speech

Use the ML pipeline (learned in M1) to build a classification model that can identify offensive language and hate speech. It is not an easy task to get good results. Experiment with different models on the two types of text-representations that you create in 2.

Bonus: Explore missclassified hate speech tweets vs those correctly predicted. Can you find specific patterns? Can you observe some topics that are more prevalent in those that the model identifies correcly?

The best-reported results for this dataset are.

| Class | Precision |
|---|---|
| 0 | 0.61 |
| 1 | 0.91 |
| 2 | 0.95 |
| Overall | 0.91 |

Here advanced NLP feature engineering has been used, and thus everything around an overall accuracy of 85 is fine. You will see that it is not easy to lift class 0 accuracy over 0.5

Good Luck!