Data sets

All data for the competition is collected from Twitter and manually annotated mainly via the Figur8 crowdsourcing platform. They are organized in four datasets especially released for the competition according to the languages and targets involved. More specifically, they will include TWO datasets, containing tweets about hate against women and immigrants, in English and Spanish, respectively.

A sample of each dataset is made available to participants from 08-20-2018, during the 'Practice' phase.

Format

According to the need of the task and related subtasks, for each tweet each dataset will include:

1. a numeric ID that uniquely identifies the tweet within the dataset
2. the text of the tweet
3. a binary value (1/0) indicating if HS is occurring against one of the given targets (women or immigrants)

4. if HS occurs (i.e. the value for the feature at point 2 is 1):
    - TS = a binary value indicating if the target is a generic group of people (0) or a specific individual (1)

5. if HS occurs (i.e. the value for the feature at point 2 is 1):
    - AG = a binary value indicating if the tweeter is aggressive (1) or not (0)

An annotated tweet is a tab-separated line with the following pattern:

id[tab]text[tab]HS[tab]TR[tab]AG

where 'id' is a progressive number denoting the tweet, 'text' is the given text of the tweet while the other parts of the pattern (given in trial and training data and to be predicted in testing data) are: whether Hate Speech (HS) is present (1) or not (0), whether the Target Range (TR) is the whole group (0) or a single individual (1), and whether Aggressiveness (AG) is absent (0) or present (1). An example of annotation is reported in the following:

42648663[tab]*USER_NAME Stupid ugly cunt who needs to die*[tab]*1*[tab]*1*[tab]*1*

Notice that aggressiveness is not a mandatory characteristic of all hateful texts and some text can express hate against a target in terms of disrespect but without using an aggressive language.

Our datasets are also available in this GitHub repository:

https://github.com/msang/hateval/tree/master/SemEval2019-Task5/datasets