Dataset for the "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior" paper, published in ICWSM 2018. The full text of the paper can be found [here](https://arxiv.org/pdf/1802.00393.pdf).

The dataset provided here includes an updated version of the original dataset, with ~100k tweets annotated using the CrowdFlower platform:

* hatespeech\_id\_label\_PUBLIC\_100K.csv: contains ~100K rows, where every row consists of a unique Tweet ID.
* hatespeech\_text\_label\_vote\_RESTRICTED\_100K.csv: contains ~100K rows, where every row consists of the tweet text, its label according to majority annotation and the number of majority annotators.
* retweets.csv: contains ~2K rows, where every row consists of the row number in the hatespeech\_text\_label\_vote\_RESTRICTED\_100K.csv file which is the first occurrence of a Tweet text followed by comma-separated row numbers of all other occurrences of the same Tweet text in the same file. There are ~8K other occurrences due to retweets.

UPDATE: It has come to our understanding that a number of the tweets are not available anymore for download on Twitter. Therefore, we provide the hatespeech\_text\_label\_vote\_RESTRICTED\_100K file with the full ~100K tweet text, their associated majority label, and the number of votes for the majority label. The tweets are shuffled so that there is no connection between tweet IDs and texts (in order to be in line with the T&C of Twitter).

***Please cite the paper in any published work that uses any of these resources.***

@inproceedings{founta2018large,  
 title={Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior},  
 author={Founta, Antigoni-Maria and Djouvas, Constantinos and Chatzakou, Despoina and Leontiadis, Ilias and Blackburn, Jeremy and Stringhini, Gianluca and Vakali, Athena and Sirivianos, Michael and Kourtellis, Nicolas},  
 booktitle={11th International Conference on Web and Social Media, ICWSM 2018},  
 year={2018},  
 organization={AAAI Press}  
}

For any further questions contact a.m.founta at gmail dot com AND markos.charalambous at eecei dot cut dot ac dot cy