**Instructions for Online Hate Annotation**

Attached to this email is a CSV with 1,000 tweets. We would like you to annotate whether you believe they contain hate speech. Please assign each tweet a value of '1' if you think it is hateful and '0' if you think it is not.

Hate speech is defined as *“language that is used in reference to certain groups that*

*expresses hatred towards the group or is intended to be derogatory, to humiliate, or to*

*insult the members of the group.”* Hate speech often targets people based on their race, ethnicity, gender, sexualpreference, or religion. This might (but does not necessarily) include calls for action

against a targeted group, such as discrimination, deportation, or physical violence.

It is important to note that the use of certain words (e.g. the n-word or “bitch”) does not

necessarily indicate hate speech - they must be used within the context described in the

above definition. The following are actual examples taken from Twitter messages of what should be

considered as hate speech:

* “I hate white trash.”
* “There really needs to be a Holocaust for faggots, y'all taking shit too far wearing skirts
* and leggings, go to Hell.”
* “Only thing niggers are good at is basketball #run #shot #steal”
* "Need to send these wetbacks back to their country”

Please note, we are NOT interested in whether content is ‘offensive’ or ‘incivil’. You are ONLY marking up whether it is hateful. Bear in this in mind at all times and make sure that you read and follow the attached instructions carefully.

Please do not try to guess about how what proportion of the tweets should be considered hateful and do not make any assumptions about whether these tweets are representative of Twitter in general. Judge each tweet solely on its own merits.

If you have any questions, please reach out to the study authors.