Annotation guidelines for hate speech detection in tweets.

Welcome to my thesis annotation and thank you for choosing to participate in this project.

You're asked to read a given set of tweets in English on the data-labelling platform Swivl. These tweets are having issues related to immigration, religion, nationalities and countries as the main topic, and, for each tweet, you're asked to answer some questions regarding the presence or not of hate speech.

What is hate speech?

Online hate that is composed of the use of language that contains either hate speech targeted toward individuals or groups, profanity, offensive language, or toxicity – in other words, comments that are rude, disrespectful, and can result in negative online and offline consequences for the individual, community, and society at large.

What does hate speech include?

- insults, threats, denigrating or hateful expressions
- incitement to hatred, violence or violation of rights to individuals or groups perceived as different for somatic traits (e.g. skin color), origin, cultural traits, language, etc.
- presumed association of origin/ethnicity with cognitive abilities, propensity to crime, laziness or other vices
- references to the alleged inferiority (or superiority) of some ethnic groups with respect to others
- de-legitimation of social position or credibility based on origin/ethnicity
- references to certain backgrounds/ethnicities as a threat to the national security or welfare or as competitors in the distribution of government resources
- dehumanization or association with animals or entities considered inferior

For the scope of this thesis project, only specific **target groups** of hate speech in tweets are relevant. These groups are can be divided into four categories:

- IMMIGRANTS, REFUGEES&MIGRANTS
- NATIONALITIES &COUNTRIES
- RELIGION
- POLITICAL INDIVIDUALS

From this point onward the definition **target group (X)** will be used to refer to one of the categories of the targeted groups.

On the next page, the **three** labels are described that can be assigned to each tweet. Every tweet can be assigned once and a tweet should **always** be assigned to a label.

Hate Speech

While answering the question "Is this tweet hateful?", the following aspects have to be taken into account:

A. the tweet content MUST have X as main TARGET, or even a single individual, but considered for his/her membership in that category (and NOT for the individual characteristics)
B. we must deal with a message that spreads, incites, promotes or justifies HATRED OR VIOLENCE TOWARDS THE TARGET, or a message that aims at dehumanizing, hurting or intimidating the target. A tweet can be considered hateful if condition A is true AND at least one of the following statements is true:

- it implies or legitimates **discriminating attitudes** or **policies** against the given target
- there is an allusion to a **potential threat** posed by the presence of the target, or its **alleged outnumbering** with respect to the native population
- there is a sense of **dissatisfaction** and **frustration**, which may also result in **overt hostility**, due to the (perceived) **privileged treatment** granted to the target group by the government
- there is the reference (whether explicit or just implied) to **violent actions** of any kind perpetrated against the given target of the message

Offensive language

While answering the question "Is this tweet offensive?", you must take into account the following aspects:

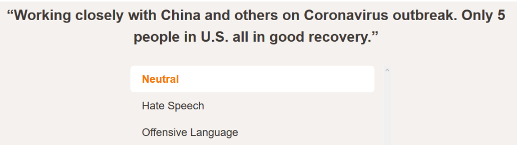
A. the tweet content MUST have X as main TARGET, or even a single individual, but considered for his/her membership in that category (and NOT for the individual characteristics)
B. we must deal with a message that focuses on the potentially HURTFUL EFFECT of the tweet content on a given target. A tweet can be considered offensive if condition A is true AND at least one of the following statements is true:

- the given target is associated with **typical human flaws**, or in **general negative characteristics**
- the **status** of the target group is **questioned**
- the **members** of the target group are described, or just considered, **unpleasant people**, or just the kind of people you better have **nothing to do with**
- there is **derisive intent**
- the target group is addressed to by means of **outrageous** or **degrading expressions**
- An overtly insulting language is used

Neutral

If none of these previous conditions apply, then your answer will be 'Neutral'.

Example of a task



“Working closely with China and others on Coronavirus outbreak. Only 5 people in U.S. all in good recovery.”

When is a tweet NOT **hateful** or **offensive**?

A. It does not incite hatred or offense:

The Hong Kong thing is a very tough situation. Very tough[1].

B. It does not have X as main target:

The only way to change it would be to kill the problem [the LGBTQ community] out. I know it's bad to say but without killing them out there's no way to fix it[2].

Remember: HATE or OFFENSE often is implicit.

Always ask yourself: what is the **intention** of this tweet? A sentence can contain positive words, while the intention might be hurtful for the target group.

You're now ready to start annotating. **Good luck**! ☺