Annotation guidelines:

CONTENT WARNING

Please be advised that this codebook contains offensive and hateful language which you could find harmful or may otherwise affect you negatively. Please seek advice and support if working with hate speech and be aware of the potential for harm at all times.

Introduction:

Welcome, and thank you for participating in this task.

In this task, you will be asked to label a set of social media posts (mostly from Twitter). For each post, you will be asked to determine whether the post is either neutral (annotated as 0), offensive (1) or hateful (2). If you consider the post hateful, you will be asked to determine the target of the hateful post.

We present below guidelines for your annotation task, adapted from Vidgen et al. (2021) and Ousidhoum et al. (2019). During annotation, we ask you to **please stick to these guidelines**.

1. Labels:

a) Hate speech

Definition: Based on Twitter's definition from September 2022, we define hate speech as the promotion of violence or a direct attack, abuse or threat against an individual or a group based on the perceived belonging of a certain characteristic. These characteristics include but are not limited to: race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.

We underline important aspects of this definition of hate speech:

1. Hate involves expressing something negative, such as contempt, disparagement, derogation, demonization, harm or bias.
2. Hate is directed against the identity of a group. Identities can be understood as the social groups and affiliations that individuals belong to and are associated with.
3. Hate is an intentional action. Whilst much research has drawn attention to the spread of 'microaggressions', hate establishes a higher bar, requiring that the speaker intends to express something hateful and that it is not just a wilful mistake or an act of ignorance. Note that 'intention' is hard to discern in many online contexts and that it is the most conceptually difficult part of understanding hate.

b) Offensive content

Definition: We follow Zampieri et al. (2019) and define offensive content as “ containing any form of non-acceptable language (profanity). [...] This includes insults, threats, and posts containing profane language or swear words”.

Note that according to this definition, hate speech is offensive but the other way around is not true. Offensive content differs from hateful content in the sense that it does not target individuals or groups based on their perceived belonging to a certain identity group.

c) Neutral content

We define neutral content as any content that is neither hateful nor offensive.

2. Examples

Hate speech

To give you a sense of what you may encounter during annotation, we provide common forms of hate speech as well as examples for each form below. **Note that this list is not exhaustive and that you should annotate as hateful anything that matches the definition of hate speech described above.** Common forms of hate speech are:

- Threatening language: defined as (a) taking ACTION against a group which (b) inflicts either serious or imminent HARM on them. Both these aspects must be there for content to be threatening. Harm can include physical harm/violence, criminal damage, intimidation/harassment, emotional abuse and mental health problems, political exclusion and denial of important/fundamental rights, financial harm, doxing (e.g. sharing of private information online). Examples include:
    - Direct threat: “I am going to attack every X”
    - Support for harmful action against a group: “We should attack these X”
    - Advocating that others inflict harm against a group: ““you would be better off blowing them up”
- Dehumanisation, defined as language which describes groups as insects, animals, trash or explicitly compares them to these. Dehumanization must express maliciousness, showing evidence of extreme prejudice and hostility against the group. Examples include:
    - “All X are cockroaches”
    - “All X are trash”
- Support for hateful entities, defined as language which glorifies, embraces, justifies or supports hateful actions, events, organizations, tropes and individuals. Examples include:
    - Denial of historical atrocities (Holocaust, Rwandan genocide, Apartheid)
    - Support for hateful figures (e.g. Hitler, Mussolini, Pol Pot or David Duke)

- Derogation, defined as language which explicitly derogates, demonizes, demeans or insults a group. Examples include:
    - “X are all terrorists”
    - “I hate X”
    - “X want to take over the country and change our way of life”
- Animosity, defined as language which expresses abuse against a group in an implicit or subtle manner. The lynchpin of this category is that (1) the group is treated negatively but (2) this is not expressed explicitly.
    - “Sometimes I think that political correctness has gone too far, why on earth is so much money given to refugees?”
    - “My friend had a horrible experience with letting out her flat. The family, obviously Indians, left it in a right state. It was full of rubbish and completely filthy, I guess they should've expected it really.”

Offensive content

Common categories of offensive content are:

- Insults:
    - "You're so stupid for believing that. #idiot"
    - "Your opinion is so dumb. Do us a favor and stay silent. #ignorant"
- Threats:
    - "If you don't shut up, I'll make sure you regret it."
    - "Cross me again, and you'll see what happens."
- Profanity:
    - "This is absolute bull****!"
    - "I can't believe this s*** is happening again. #frustrated"

Again, this list is not exhaustive **and you should annotate as offensive anything that matches the definition of offensive content described above.** As discussed in the definition section, hate speech is offensive which explains why some categories overlap between hate speech and offensive content (e.g. threat). We emphasize again though that offensive content is not always hateful. For example, “F*ck you” is offensive whereas “F*ck you, you stupid Muslim” is hateful and the target is “Muslims”.

3. Task description

You will be provided with a sheet in a Google Sheets document (normally called `{language/country code}_{your first name}`, e.g. `fr_manuel` for me) containing a list of social media posts (mostly from Twitter). In this document, there will be the following 4 columns:

- a `text` column, containing the text of the post to annotate
- a `label` column you will fill in after determining if the `text` is neutral, offensive or hateful
- a `hate_target` free text column, where you will indicate the target in case a tweet is hateful
- a `flag_unsure` column

Your task consists in reading the post in the `text` column, and annotating it, using the `label` , `hate_target` and `flag_unsure` columns. The possible values for the `label` column are:

- 0 if you think the post is “neutral”
- 1 if “offensive”
- 2 if “hateful”

We ask you to always give your best guess in the `label` column. In case you give your best guess but you are unsure if it is correct, please additionally give the value 1 to the column `flag_unsure`. Even though hateful content may be offensive as described above, we ask you to **assign each post to one single category** (either hateful, offensive or neutral).

In case you consider the post as hateful, please indicate in plain text the target of the tweet in your opinion in the `target` column. Please note that one hateful tweet may have multiple targets and we ask you to please indicate all targets that apply.