<u>Annotation Guidelines</u>

Background:
We are creating better ways of testing models for the automated detection of online hate speech. For this purpose, we have compiled a dataset consisting of a few thousand short text entries. Each entry reflects a particular aspect of online hate.

Your tasks:

1. Mark whether each entry assigned to you is **hateful** or **not hateful.**
2. Flag any entries which you think are **unrealistic.**

What do we mean by "hate"?
We define **hate** as abuse that is targeted at a protected group or at its members for being a part of that group. Protected groups are based on age, disability, race (colour, nationality, ethnic or national origins), religion or belief, sex and sexual orientation as well as gender identity.

What do we mean by "unrealistic"?
We intentionally leave this up to you. We did not construct any entries to be unrealistic, so this optional flag is merely meant to ensure data quality.

Some things to keep in mind while annotating:

- **<u>Language that is hateful in some contexts can have non-hateful uses</u>** (e.g., counter speech that references hateful slurs: *"It's not okay to call people niggers"*)
- **<u>The target of abuse matters for whether something is hateful</u>**. Interpersonal abuse and abuse against non-protected groups such as professions and affiliations is not hate. (e.g. *"I hate Jews"* is hateful but *"I hate you"*, *"I hate my table"* and *"I hate doctors"* are not).

Other key bits of guidance:

1. Do not flag entries as unrealistic just because you think they are unlikely to appear often online. Only flag nonsensical and grammatically incorrect entries.
2. Please complete your annotations independently and don't talk about them with others.
3. Most statements are quite short so please don't overthink them.

Thanks a lot for your work on this!