Annotation Instructions

In the following, you are asked to annotate a set of small sentences. We ask you to decide for every individual sentence whether you think it contains **offensive language** or not.

By offensive language we define utterances that are meant to hurt a single person or a group of persons:

- you stupid bitch nuclear power kills shut up with your lies
- You have the face of someone only a mother could love.
- Just want to slap the stupid out of these bimbos!!!

Many of the offensive sentences will also focus on specific identity groups (i.e. gay people, Jews, Muslims and women). The data you will be given will focus on: antisemitic remarks, homophobic remarks, Islamophobic remarks and misogynistic remarks.

Antisemitic remarks display hostility to, prejudice, or discrimination against Jews. Typical examples are:

- I cannot believe how perverted and sick Jews are.
- These jews see babies as sex toys to be exploited at will.
- When Walt Disney died in 1966, the last barrier to the total Jewish domination of Hollywood was gone, and Jews were able to grab ownership of the company that Walt built.

Homophobic remarks display hostility to, prejudice, or discrimination against homosexual people. Typical examples are:

- GAY PEOPLE ARE ALWAYS TRYING TO FORCE GAYNESS ON YOU
- Put on a wig and lipstick and be who you really are.
- gay people are sooo dramatic

Islamophobic remarks display hostility to, prejudice, or discrimination against Muslims. Typical examples are:

- Islam forces societies to regress.
- Europe needs to expel their Muslim immigrants.
- Muslims train babies to hate and murder, then use them for propaganda.
- And the Quran says that the Muslims must steal and occupy the entire world.

Misogynistic remarks display hostility to, prejudice, or discrimination against women. Typical examples are:

- Women Are Less Intelligent Than Men.
- Women ruin your life completely.
- Marriage only benefits women at the expense of the man.
- Women are very manipulative.

The sentences given to you will be in a random order. This means that each sentence should be judged <u>in isolation</u> without considering the preceding or following sentence.

For each sentence, you will have to assign exactly one of the following 3 categories:

1. The first category is **OFFENSIVE**. It should be used if the sentence is either antisemitic, homophobic, islamophobic or misogynistic. Please **<u>also</u>** use this label if you think other

groups of people or even individuals are meant to be offended by that sentence, or you are just faced with profane language:
    - `Go fucking kill yourself and die already useless ugly pile of shit scumbag`
    - `Im tired of people complaining about the little shit when I lost my father to that cancer bitch`
    - `Hope one of those bitches falls over and breaks her leg`
2. The second category is **OTHER**. This label should be applied if the sentence contains no offensive language at all:
    - `I hate Mondays.`
    - `Jews are concerned by the recent antisemitic attacks in Europe.`
    - `Muslims fight against prejudice.`
    - `Gay people dislike being stereotyped.`
    - `Women fight against domestic violence.`
3. The final category is **NO_PROPER_ENGLISH**. You should always use this category if you think that the sentence is no proper English, either because it is ungrammatical or you think that the wording of the sentence does not sound like proper English. **If you do not understand the sentence**, then you should also use this label.
   Note that we have prefiltered all sentences that you are going to annotate, so we expect only few cases of this category in this survey.

Please bear in mind the following **advice**:

- **Just because a sentence contains a mention of some identity term (e.g. Jews, Muslims gay people, women) does not mean the sentence is automatically offensive! Consider, for example, the above examples for the category OTHER.**
- Offensive remarks (including antisemitic, homophobic and Islamophobic remarks) need not employ explicitly offensive terminology. In other words, just watching out for offensive expressions, such as *kike, faggot, goatfucker* or *bitch* is insufficient. While mentions of such terms may be a strong indicator of an offensive remark, there are other ways of expressing offensive. For example, often such remarks contain negative stereotypes.
    - Examples of antisemitic stereotypes: Jews are greedy, dishonest and manipulative beings.
    - Examples of homophobic stereotypes: Gay people are oversensitive beings, discriminate against transsexuals, want to impose their lifestyle to other people and are a threat to traditional values and institutions (e.g. marriage).
    - Examples of Islamophobic stereotypes: Muslims as brutal, misogynistic and only emigrate to Western countries to take advantage of social benefit.
    - Examples of misogynistic stereotypes: women are manipulative, deceitful, capable of using seduction to control men, and need to be kept in their place.
- **It is very important that you label news items (or sentences that read like news items) as OTHER.** News items, too, may address one of the identity groups. They may even be negative in tone. The major difference is that they refer to specific events that took place (or are just taking place):
  `a) Muslims killed a woman in Sweden.`
  News items would not refer to stereotypical properties of identity groups, such as the following sentence:
  `b) Muslims always kill wherever they go.`
  Therefore, b) should be labeled as OFFENSIVE.
  You should also try to figure out how the mention of the identity group is to be interpreted.

If it just refers to a subset of the group (e.g. *Muslims* in a) reads as *some Muslims*), then the sentence you annotate is more likely to be a news item. If the mention generalizes to the entire identity group (e.g. *Muslims* as in b) reads as *all Muslims*), then you are more likely to deal with a stereotype.

Our recommendation for the annotation is that you store the above guidelines or print them out. You should **use the above examples as a reference** for your annotation.

By the nature of this survey, you will be exposed to some deeply offensive language. This survey is intended to produce labeled data for subsequent linguistic analysis. The authors of this survey, **in no way**, share the views that are expressed in these offensive remarks.