Annotation guidelines for Offensive Speech Detection

Introduction

This document presents guidelines on how to annotate potentially harmful tweets that can cause emotional distress to individuals, incite violence, or discriminate against, and exclude social groups.

As an annotator, it is important to approach this task with objectivity (as much as possible). We welcome your feedback on how we can update the guidelines based on the peculiarity of the language you are annotating, your background, or any socio-linguistic knowledge that we may have overlooked. Consider the following when performing the task:

Always use the guidelines and you should be objective and consistent in your annotation.

- Focus on the message conveyed in the tweets and try not to focus on your personal opinion on the topic.
- Do not rush to finish the task and always reach out to your language coordinator with questions when in doubt.

Mental health risk and well-being:

Annotating harmful content can be psychologically distressing. We advise any annotator who feels anxious or uncomfortable during the process to take a break or stop the task and seek help. Early intervention is the best way to cope.

Definitions:

- **Abusive language** is any form of bad language expressions including rude, impolite, insulting or belittling utterance intended to offend or harm an individual.
- **Hate speech** is language content that expresses hatred towards a particular **group or individual** based on their race, ethnicity, religion, gender, sexual orientation, or other characteristics. It also includes threats of violence
- **Normal** is any form of expression that does not contain any bad language belonging to any of the above classifications.
- **Indeterminate** is any tweet that is not readable or is **completely** written in another language other than your language of annotation.

Task Description:

Task 1: Given a tweet, select the option that best describes it:

- Abusive language
- Hate speech
- Normal
- Indeterminate

Task 2: If the label in task 1 is hate, select the target of the hate speech

- Ethnicity
- Religion
- Disability
- Gender
- Others

Detailed Task Description:

1. Tweet Category Detection (TCD): You will be required to label a tweet as either **abusive, hate, normal or indeterminate**. See examples below:

Table 1: Examples of tweets classified as offensive, hate, normal and indeterminate.

Label	Tweet
Abusive	● You are very stupid! Who do you think you are? ● People like you don't deserve five star hotels. ● Ariza is a snake and a coward"" but at least he isn't a cripple like your hero Roach lmaoo” ● Go lick a pig you arab muslim piece of scum ● What would y'all lil ugly bald headed bitches do if they stop making make-up and; weave?" ● how come you never bring me food" i dont have a car retard
Hate	● alsarabsss" hes a beaner smh you can tell hes a mexican ● We should let women do more housework, and encourage their husbands' work ● Chuck Hagel will shield Americans from the desert animals bickering. Let them kill each other, good riddance! ● happy birthday nigs" ● LMFAOOOO I HATE BLACK PEOPLE This is why there's black people and niggers ● "Our people". Now is the time for the Aryan race 2 stand up and say "no more". Before the mongerls turn the world into a ghetto slum ● We hate niggers, we hate faggots and we hate spics ● California is full of white trash ● At least I'm not a nigger
Normal	● Much love my nigger. ● I'm an early bird and I'm a night owl, so I'm wise and have worms." ● Banning Muslim immigrants will not change anything
Indeterminate	● Tweets that are completely written in languages other than English and your language of annotation ● Tweets that make no sense or do not have any meaning

2. Hate Target Detection (HTD): If the tweet is marked as `hate speech', you will be asked to identify the associated target attribute. The target attribute can be one of the following:

a. **Ethnicity:** This includes tweets that target an individual or group because of their ethnicity.

b. ***Religion:*** This kind of tweets targets individuals or groups because of their religious belief.

c. **Disability:** This is when a tweet is hateful towards individuals with special needs.

d. **Gender:** This includes hateful tweets towards a particular gender.

e. **Others:** Hateful tweets that do not belong to any of the above categories.

Target	Tweet	Translation
Ethnicity	● Arabs are shit! ● The Kilan tribe are terrorist	
Religion	● No Hindu should be giving resident permit ● All Muslims are terrorists	
Disability	● A crippled man can never be my husband ● We don't employ disabled people in our company	
Gender	● Women belong in the kitchen ● I don't like this gender	

Annotation Procedure Summary

Carefully read and understand the meaning and content of each tweet, then follow the steps outlined in the flowchart below to annotate each tweet.

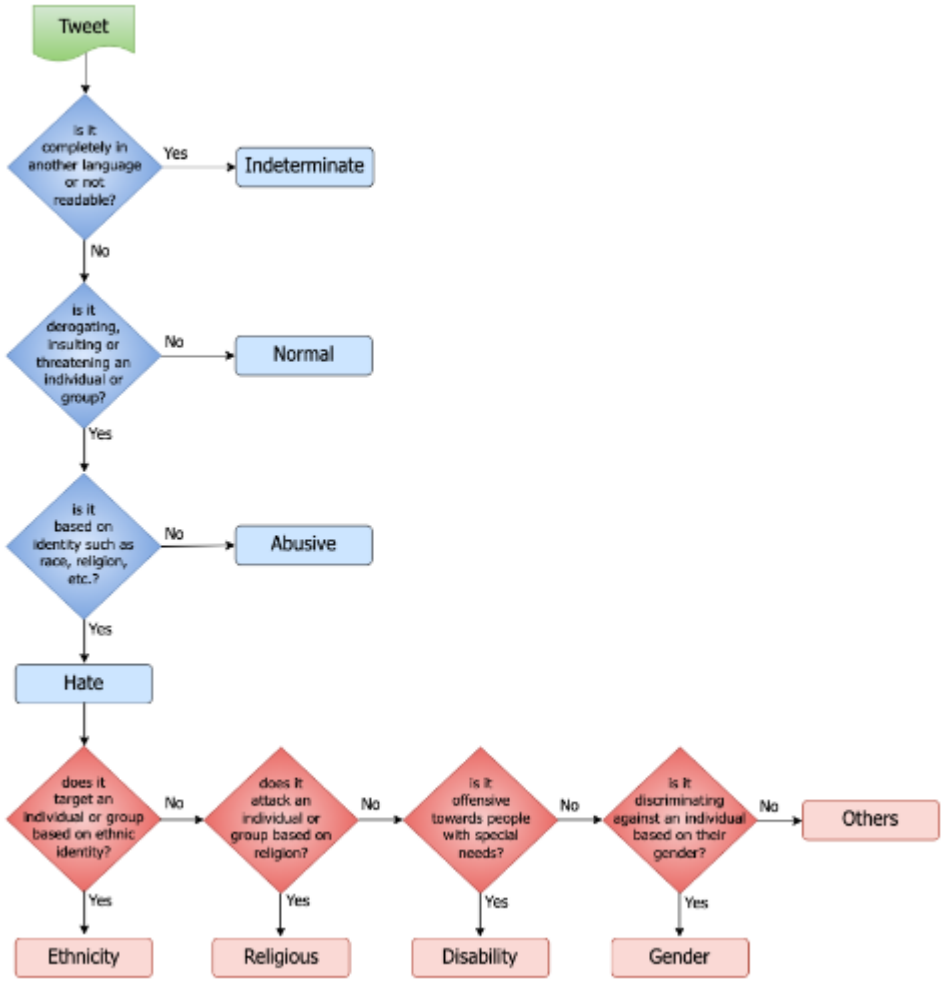

Figure 1: Annotation flowchart