Annotation Guideline for Amharic Hate Speech Annotations using Rating Scales

Introduction

This guideline presents the concepts and rules on how to annotate and rate potentially harmful tweets that can cause emotional distress to individuals, incite violence, or discriminate against, and exclude social groups.

Annotators are expected to be **objective (as much as possible).** We welcome your feedback on how we can update the guidelines based on your feedbacks.
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 own opinion on the topic.
- **Do not rush to finish the task** and always reach out to the researchers 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.

Category:

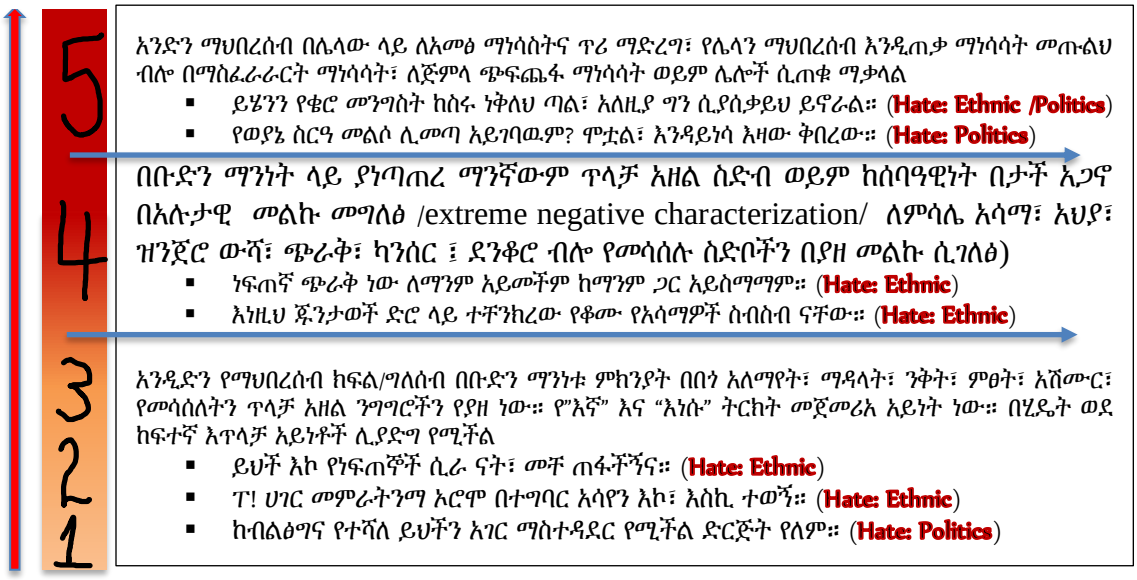
- **Offensive speech** is any form of bad language expressions including r*ude, impolite, insulting, or belittling utterance intended to offend or harm an individual.*
- **Hate speech** is language content that expresses **hatred towards a particular *group or individual*** based on their group identities such as **race**, **ethnicity**, **religion**, **gender, disability, political affiliation, or other characteristics.** It also includes **threats of violence** associating group identities.
- **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.

For each tweet, you are required to identify the category and for **hate** and **offensive** tweets, you are requested to indicate the severity of hate or offensive tweets by rating from **1-5** in **a Likert scale** as shown in the example.

Severity level:

1. Hate speech Severity levels:

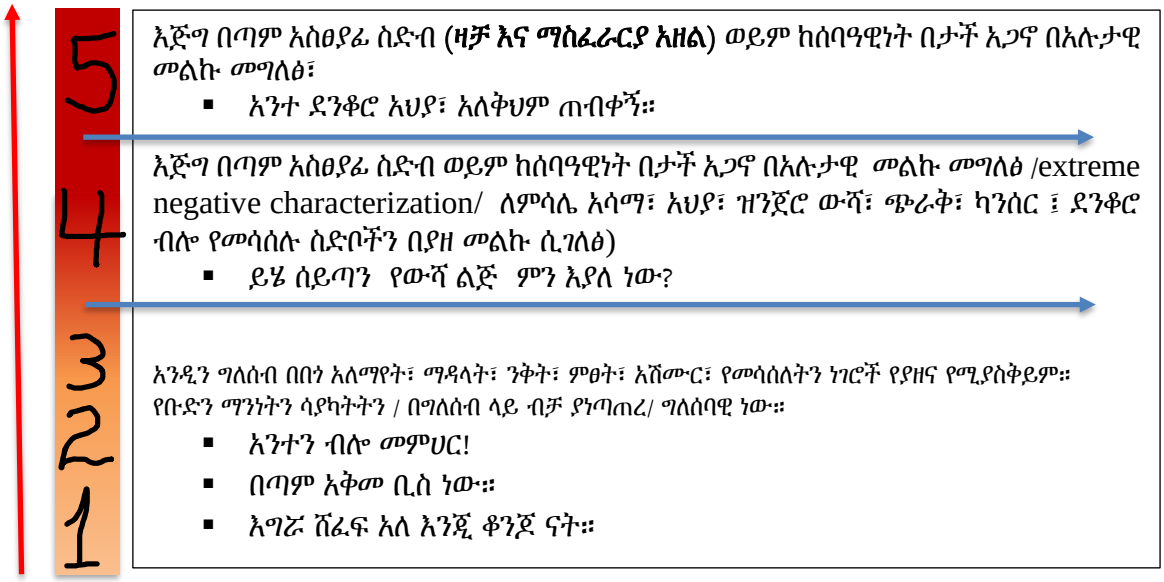
Extreme hate intensity



Less hate intensity

2. Offensive speech Severity levels:

Extreme offensive intensity

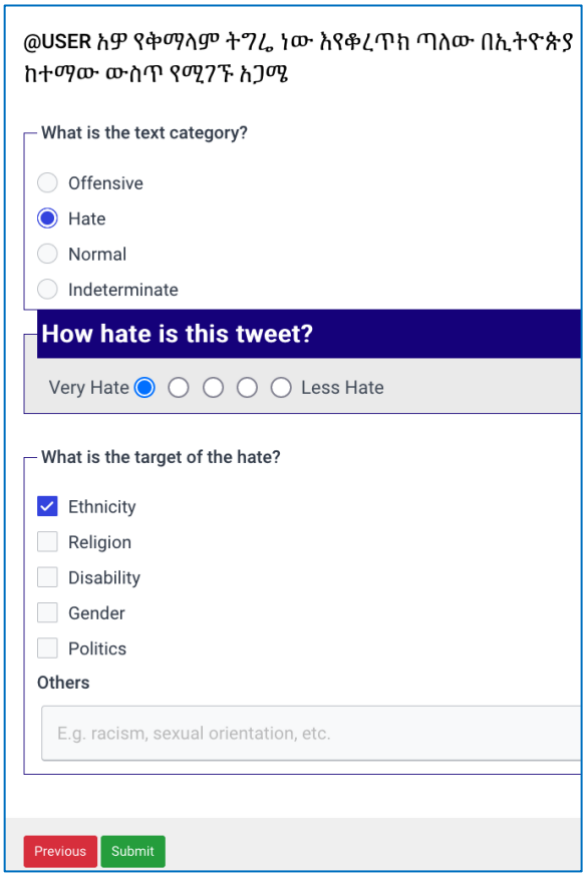


Less offensive intensity

For all hate speech categories, you are requested to label the target of the hateful content as follows:

- Ethnic: if the hatred tweet targets an ethnic group identity
- Religion: if the hatred tweet targets a religious group identity
- Gender: if the hatred tweet targets a particular gender group identity
- Disability: if the hatred tweet targets disabled group identity
- Politics: if the hatred tweet targets people/entities due to political ideology
- Unidentified target: if the hatred tweet's target is not clearly identified/known.
- Other: if the hatred tweet targets other groups/group identities such as sexual orientation, racism etc.

Example:1



Example:2