

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 *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 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

5	<p>አንድን ማህበረሰብ በሌላው ላይ ለአመፅ ማነሳሰትና ጥሪ ማድረግ፣ የሌላን ማህበረሰብ እንዲጠቃ ማነሳሰት መጠላለፍ በማስፈራራት ማነሳሰት፣ ለጅምላ ጭፍጨፋ ማነሳሰት ወይም ሌሎች ሲጠቁ ማቃላል</p> <ul style="list-style-type: none"> ይሄንን የቁሮ መንግስት ከስሩ ነቅለህ ጣል፣ አለዚያ ግን ሲያስቃይህ ይኖራል። (Hate: Ethnic /Politics) የወያኔ ስርዓ መልሶ ሊመጣ አይገባዉም? ሞቷል፣ እንዳይነሳ እዛው ቅበረው። (Hate: Politics)
4	<p>በቡድን ማንነት ላይ ያነጣጠረ ማንኛውም ጥላቻ አዘል ስድብ ወይም ከሰባዊነት በታች አጋኖ በአሉታዊ መልኩ መግለፅ /extreme negative characterization/ ለምሳሌ አሳማ፣ አህያ፣ ዝንጀሮ ውሻ፣ ጭራቅ፣ ካንሰር ፤ ደንቆሮ ብሎ የመሳሰሉ ስድቦችን በያዘ መልኩ ሲገለፅ</p> <ul style="list-style-type: none"> ነፍጠኛ ጭራቅ ነው ለማንም አይመችም ከማንም ጋር አይስማማም። (Hate: Ethnic) እነዚህ ጁንታወች ድሮ ላይ ተቸንክረው የቆሙ የአሳማዎች ስብስብ ናቸው። (Hate: Ethnic)
3	<p>አንዲድን የማህበረሰብ ክፍል/ግለሰብ በቡድን ማንነቱ ምክንያት በበጎ አለማየት፣ ማዳላት፣ ንቅት፣ ምፀት፣ አሽመር፣ የመሳሰሉትን ጥላቻ አዘል ንግግሮችን የያዘ ነው። የ"እኛ" እና "እነሱ" ትርክት መጀመሪያ አይነት ነው። በሂደቱ ወደ ከፍተኛ አጥላቻ አይነቶች ሊያድግ የሚችል</p> <ul style="list-style-type: none"> ይህች እኮ የነፍጠኞች ሲራ ናት፣ መቸ ጠፋችኝና። (Hate: Ethnic) ፕ! ሀገር መምራትንማ አሮሞ በተግባር አሳየን እኮ፣ እስኪ ተወኝ። (Hate: Ethnic) ከብልፅግና የተሻለ ይህችን አገር ማስተዳደር የሚችል ድርጅት የለም። (Hate: Politics)
2	
1	

Less hate intensity

2. Offensive speech Severity levels:

Extreme offensive intensity

5	<p>እጅግ በጣም አስፀያፊ ስድብ (ዛቻ እና ማስፈራራያ አዘል) ወይም ከሰባዊነት በታች አጋኖ በአሉታዊ መልኩ መግለፅ፣</p> <ul style="list-style-type: none"> አንተ ደንቆሮ አህያ፣ አለቅህም ጠብቀኝ።
4	<p>እጅግ በጣም አስፀያፊ ስድብ ወይም ከሰባዊነት በታች አጋኖ በአሉታዊ መልኩ መግለፅ /extreme negative characterization/ ለምሳሌ አሳማ፣ አህያ፣ ዝንጀሮ ውሻ፣ ጭራቅ፣ ካንሰር ፤ ደንቆሮ ብሎ የመሳሰሉ ስድቦችን በያዘ መልኩ ሲገለፅ</p> <ul style="list-style-type: none"> ይሄ ሰይጣን የውሻ ልጅ ምን እያለ ነው?
3	<p>አንዲን ግለሰብ በበጎ አለማየት፣ ማዳላት፣ ንቅት፣ ምፀት፣ አሽመር፣ የመሳሰሉትን ነገሮች የያዘና የሚያስቅይም። የቡድን ማንነትን ሳያካትት / በግለሰብ ላይ ብቻ ያነጣጠረ/ ግለሰባዊ ነው።</p> <ul style="list-style-type: none"> አንተን ብሎ መምሀር! በጣም አቅመ ቢስ ነው። እግሯ ሸፈፍ አለ እንጂ ቆንጆ ናት።
2	
1	

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

@USER አዎ ዓቅማለም ትግረ ነው አየቅረጥክ ጣለው በኢትዮጵያ ከተማው ውስጥ የሚገኙ አጋሜ

What is the text category?

☐ Offensive

☒ Hate

☐ Normal

☐ Indeterminate

How hate is this tweet?

Very Hate ☒ ☐ ☐ ☐ ☐ Less Hate

What is the target of the hate?

☒ Ethnicity

☐ Religion

☐ Disability

☐ Gender

☐ Politics

Others

E.g. racism, sexual orientation, etc.

[Previous](#) [Submit](#)

Example:2

@USER @USER ሆኖ ሲያውቅ ጾሮ ማታ አንቺ አጋሜ አማራ አንኳን አንቺ ጣለያን ያውቀዋል ዘረኛ ገንጣይ አስገንጣይ የሴይጣን ልጆች::ዘመዶሽን አግር ለአግር አየተ

What is the text category?

☐ Offensive

☒ Hate

☐ Normal

☐ Indeterminate

How hate is this tweet?

Very Hate ☐ ☒ ☐ ☐ ☐ Less Hate

What is the target of the hate?

☒ Ethnicity

☐ Religion

☐ Disability

☐ Gender

☐ Politics

Others

E.g. racism, sexual orientation, etc.

[Previous](#) [Submit](#)