

Guideline for Annotators (Version 3)

A note on the tool :

- The tool is trying to capture those instances of online gender based violence that target persons of marginalised gender who are active on social media and are often at a receiving end of hate speech and harassment due to their profession such as journalists, celebrities or are targeted for their political views to create a chilling effect. This kind of harassment doesn't usually infringe the community guidelines of a social media platform.
- The annotation process is an exercise to capture the connotations and context of the slur. In our opinion, the data set can have multiple use cases:

Possible use-cases of the Lexicon:

1. Help Uli effectively capture instances of online gender-based violence that targets persons of marginalised gender.
2. Create awareness on slurs that have been normalized in their everyday use but reflect deep social, cultural and historical hierarchies.
3. Create an archive of coded offensive words/phrases (which are otherwise not slurs) that are being used on social media.

Guidelines:

- **Please rate the severity of offence for the word on a scale:**
 - **High:** The word is very offensive.
 - **Low:** The word is mildly offensive even when used in a derogatory manner
 - **Medium:** If the word doesn't fit as either highly offensive, or slightly offensive, please mark it as medium offensive.
- **Each offensive word, can also be used in manner that is either:**
 - **Casual**, where an offensive word can be used to tease friends, or in a casual manner to express harmless anger. Such as *bitch*, *saali*, *chutiya*, *kameenay*.

- **Appropriated**, where an otherwise offensive word is used to rework the previous derogatory meaning into something positive (reclamation). For instance, slut in slut walk, or queer in LGBTQIA+ politics.

A word can be both of the above as well.

- **Annotators can note if an appropriated use is facilitated by the**
 - **affected community**, where affected community is the community or members of a group or a community which is often at a receiving end of the offensive word/phrase. For instance a feminist campaign's use of the phrase "pink chaddi" would be marked as appropriated by 'affected community'. Or the word 'slut' has been reclaimed by certain groups.
 - **others**, where others refers to groups or individuals who might not be at the receiving end of the offensive word/phrase as an identifier. For instance, a word play on *chutiyapa* as *Qtiyapa* by a media house, or use of the word *Kameenay* as a movie title would be captured as appropriated by 'others'.

- **What makes a word problematic:**

Factors that contribute to a word being offensive should be noted as text. Some cues are:

- **Social/Cultural connotations**, that refer to certain established social or cultural values. For instance, promiscuity among men and women is morally judged in certain societies and expressed through slurs such as *chinaal etc.*
- **Meaning**, where certain words are derogatory because they signify certain negative concepts in the language system and as such do not refer to any particular person, group or an activity or a group. Such as *dirty, filthy, polluted, contaminated.*

It is important to note that social/cultural connotations and the linguistic meaning of a word are often entangled. For instance, a 'filthy' woman can connote a 'promiscuous woman'.

- **How it is said**, where certain words become offensive because of the manner in which they are uttered. For instance, the word ‘queen’ which otherwise has no social/cultural connotations or any derogatory meaning can be used offensively. For instance a tweet such as , “Madam queen demands respect”.

This factor can be difficult to identify on social media, but can be gauged in the way capitalization is used, accompanying emojis, the intended tone, or frequency etc. Annotators should use discretion and should mark only when they have seen a repetitive use of a regular word uttered in a certain way to disparage a person or a group.

- **What it’s made to do**, some regular words or phrases can become derogatory or threatening when they are used in a manner that signals an intent to offend, threaten, silence or provoke an offensive action from the observers or the targeted individuals. For instance, a troll campaign to mark profiles of activists can use a very neutral phrase such as “package picked-up” to create a chilling effect.
- **Finally, the annotator can classify the slur according to a set of categories already identified such as gendered, sexualized, religion, sexual identity, ethnicity, political affiliation, caste, class, body shaming and ableist. For any unidentified category, the annotators can select others and add their own in the available space.**

When marking the category of the offensive word, please don’t think about whether the word can be used to target that identity type- an offensive word can be used to target individuals of different identities. This categorization is referring to the identity surfaced in the origin or meaning of the word.

At present you can add at most four intersected identities in a word.

Remember: None of these categories are mutually exclusive. They overlap with each other.

Following the initial annotation process, the Tattle team would welcome feedback from the annotators regarding the exercise to improve the overall annotation process.