Annotation Guideline: Offensive Tweet Classification

Introduction:
The purpose of this annotation guideline is to provide instructions and examples for annotators to classify tweets as offensive or non-offensive. Offensive tweets often contain language or content that is disrespectful, discriminatory, vulgar, or harmful. Annotators will categorize each tweet into one of the following classes: Offensive, Non-Offensive, or Can't tell / not annotable.

Target Class:
The target class is "Offensive." We label a post as offensive if it contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. Tweets in this class contain insults, offensive language, explicit content, hate speech, threats, profane language, swear words or any other form of harmful or disrespectful communication. Annotators should focus on identifying tweets that clearly fall into this category.

Non-Offensive Class:
Tweets that do not contain any offensive language or harmful content should be classified as "Non-Offensive." These tweets can be neutral, informative, positive, or unrelated to offensive topics.

Can't tell / not annotable Class:
This class is used for tweets that cannot be definitively classified as offensive or non-offensive due to factors such as ambiguous language, incomplete context, or content that is difficult to interpret. Annotators should assign this class sparingly and only when it is genuinely not possible to determine the tweet's offensive or non-offensive nature.

Examples:
To illustrate the classification process, here are some examples of tweets along with their respective classes:

Example 1:
Tweet ID: 1410492618790817793
Text: YOU BETTER SUCK HIS DICK KOZY I SEE YOU WITH KNUCKLES GET EM GYAAAAL
Class: Offensive

Example 2:
Tweet ID: 1410492618790686720
Text: You should raise the webform....how would they know then that you completed ur medicals
Class: Non-Offensive

Example 3:
Tweet ID: 1410492618803335174
Text: im tired too but this is so entertaining i cant
Class: Non-Offensive

Example 4:
Tweet ID: 1410492618778157059
Text: Fuckoff
Class: Offensive

Example 5:
Tweet ID: 1410492622972588038
Text: Even if they didn’t exploit people to acquire their riches, how are you gonna be okay literally wasting thousands and thousands of dollars while there are still people who are homeless? While there are people skipping life-saving medical treatments bc of the cost?
Class: Non-Offensive

Guidelines:

1. Read the tweet thoroughly and consider the overall meaning and intent.

2. Determine if the tweet contains offensive language, hate speech, discrimination, threats, or explicit content.

3. Classify the tweet as Offensive if it meets any of the criteria mentioned in step 2.

4. If the tweet does not contain offensive language or harmful content, classify it as Non-Offensive.

5. If it is not possible to determine the tweet's offensive or non-offensive nature due to ambiguous language, incomplete context, or other factors, assign the Can't tell / not annotable class.

6. Avoid personal biases and judgments. Focus on the content of the tweet rather than the identity of the user.

7. If a tweet contains mixed content, where part of it is offensive and part of it is non-offensive, classify it based on the dominant tone or intent.

8. If there are any specific terms or expressions that may be context-dependent, consider the immediate context within the tweet or consult additional sources if necessary.

9. If you encounter URLs, hashtags, or other non-textual elements that are not directly related to offensive content, ignore them and focus on the text of the tweet itself.

10. If you have any doubts or questions about a tweet's classification, do not hesitate to seek clarification from the supervising team.

Note: The Can't tell / not annotable class should be used sparingly and only when there is genuine uncertainty in classifying a tweet.

Conclusion:
This annotation guideline provides instructions for classifying tweets as Offensive, Non-Offensive, or Can't tell / not annotable. It is crucial to maintain consistency and follow the guidelines accurately to ensure high-quality annotations. Remember to focus on the content of the tweet, consider the overall meaning and intent, and avoid personal biases.