ANNOTATOR GUIDELINES: CLASSIFYING TWEETS AS OFFENSIVE OR NON-OFFENSIVE

1. Purpose:

The purpose of these guidelines is to assist annotators in accurately identifying and classifying tweets as either offensive or non-offensive. The proper categorization of tweets is crucial for developing a safer, more respectful environment on the platform.

2. Definition:

Offensive: An offensive tweet contains language, imagery, or any form of communication that may cause discomfort, harm, or insult to an individual or group based on race, gender, ethnic origin, religion, disability, age, or any other characteristic.

Non-Offensive: A non-offensive tweet does not contain any such elements. It respects the dignity and sensibilities of all individuals and groups.

For every tweet you annotate, please label this way:

1 : **offensive**

0 : **neutral**

3. Guidelines:

Here are general guidelines you can refer to if whenever you have a doubt about a tweet.

3.1 Direct Offensive Language: This includes slurs, derogatory terms, insults, and obscene language targeted at individuals or groups. These are always classified as offensive.

Example: "You are a [racial slur], go back to where you came from." -> 1

3.2 Hate Speech: Any form of communication that vilifies, marginalizes, or discriminates against a group or individual based on their characteristics or identities is deemed offensive.

Example: "All [religion] people are terrorists." -> 1

3.3 Threats and Violence: Tweets promoting violence, harm, or illegal activities are considered offensive.

Example: "Anyone who supports [political group] deserves to be punched." -> 1

3.4 Sexual Harassment: Any unsolicited sexual comments, explicit content, or innuendos targeted at someone are offensive.

Example: "Hey [username], you look so hot in that photo. Want to hang out sometime?" -> 1

3.5 Non-Derogatory Swearing: Not all instances of swearing are considered offensive. Casual, non-derogatory use of swear words in a tweet does not necessarily make it offensive.

Example: "Fuck, I spilled coffee on my new shirt!" -> 0

4. Indirect Offensive Language:

4.1 Innuendo: This involves indirect hints or references to offensive or inappropriate matters. Annotators should be aware of the context to classify such tweets.

Example: "Sure, everyone from [Country X] isn't bad, but you know what they're like..." -> 1

4.2 Stereotypes and Prejudices: Promoting harmful stereotypes or prejudices, even if not directly offensive, should be considered offensive.

Example: "It's no surprise she can't park, she's a woman after all." -> 1

5. Gray Areas:

5.1 Sarcasm and Irony: The use of sarcasm and irony can often disguise offensive content. Please consider the context and potential harmful implications.

Example: "Yeah, because all [race] people are soooo smart." -> 1

5.2 Controversial Topics: Statements on contentious topics like politics, religion, etc., can be difficult to classify. They should be deemed offensive only if they involve hate speech, discrimination, or harmful stereotypes.

Example: "I don't agree with [political party's] policies." (Not offensive) -> 0

6. Note

Remember: Context matters: Check for the tweet's context. A word or phrase could be offensive in one context but not in another. When in doubt, consult: If a tweet seems ambiguous, consult with your team or supervisor.