Annotation Guideline for Offensive Tweets

Target Class

The target class is tweets that are offensive. Offensive tweets are those that are intended to cause harm or offense to others. This can include tweets that are racist, sexist, homophobic, or otherwise discriminatory. It can also include tweets that are threatening, violent, or hateful.

Context

It is important to consider the context of a tweet when determining whether it is offensive. For example, a tweet that contains a swear word may not be offensive if it is used in a joking or ironic way. Similarly, a tweet that makes a reference to a sensitive topic may not be offensive if it is done in a respectful way.

Intent

It is also important to consider the intent of the tweet's author when determining whether it is offensive. For example, a tweet that contains a swear word may not be offensive if the author is simply using it as a form of expression. However, if the author is using the swear word to insult or offend someone, then it would be considered offensive.

Audience

The audience of a tweet can also play a role in determining whether it is offensive. For example, a tweet that would be considered offensive in one context may not be offensive in another context. For example, a tweet that makes a reference to a sensitive topic may not be offensive if it is directed at a group of people who are familiar with the topic. However, if the tweet is directed at a group of people who are not familiar with the topic, then it could be considered offensive.

"Can't tell / not annotatable" class

This class should be used for tweets that are not clear enough to be annotated. This could include tweets that are too short, tweets that are in a foreign language, or tweets that are simply not understandable.

Examples

- **Offensive:**
  - "Black people are lazy."
  - "Women should stay in the kitchen."
  - "I'm going to kill all the Jews."
  - "This is a threat to kill someone."
  - "This is a tweet that promotes violence against a group of people."
- **Ambiguous:**
  - "That's a gay joke."
  - "I'm not sure if this is offensive or not."
  - "This tweet makes a reference to a sensitive topic, but I don't think it's offensive."

Additional considerations

- **Cultural context:** It is important to consider the cultural context of a tweet when determining whether it is offensive. For example, a tweet that would be considered offensive in one culture may not be offensive in another culture.
- **Humor:** It is important to consider the intent of the tweet's author when determining whether it is offensive. For example, a tweet that contains a swear word may not be offensive if the author is simply using it as a form of expression. However, if the author is using the swear word to insult or offend someone, then it would be considered offensive.
- **Frequency:** The frequency of a tweet can also play a role in determining whether it is offensive. For example, a tweet that contains a swear word may not be offensive if it is used infrequently. However, if the swear word is used frequently, then it could be considered offensive.