ANNOTATION GUIDELINES FOR TWEETS

- Sai Hemanth Kilaru

Labels:

1. Hate/Offensive:

The main purpose is to classify the tweets that use language which contains harmful, hostile or discriminatory intentions. This category includes the content aimed at offending or degrading specific people or groups or things.

What I look for:
- **Discriminatory or Hostile Language:**
  The words or phrases that negatively refer to someone based on characteristics such as gender, nationality, race, ethnicity, religion, or other identities.
- **Aggressive or Threatening Tone:**
  Messages that promote, support,encourage, or justify violence or harm against individuals or groups or countries.
- **Stereotypic content:**
  Content that makes negative assumptions ,harsh generalizations or labels a whole group in a harmful way.

2. Neutral:

The main purpose is to classify the tweets that have no negative purpose and do not direct hostility towards any specific group or individual. This category includes neutral, informative, or casual language.

What I look for:
- **General Statements:**
  Content that shares the opinions, ideas, facts, or observations without any negative intent or discriminating language.
- **Constructive Criticism :**
   Tweets that provide opinions or critiques without an aggressive or offensive tone.
- **Non-Hostile:**
  Content that does not encourage violence, discrimination, or hatred of any sort.

These are the main things I consider for my annotation process. I am the Annotator-1 and will try my best to accurately label my data according to the annotation guidelines.

References:

1. https://help.x.com/en/rules-and-policies/hateful-conduct-policy
2. https://transparency.meta.com/policies/community-standards/hate-speech/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fhate_speech
3. https://aclanthology.org/2023.woah-1.21.pdf