Annotation Guidelines for Multimodal Hate Speech Detection

Objective
The goal of this annotation task is to classify each tweet as either "Hate Speech" or "Not Hate Speech." A tweet is considered hate speech if it contains text and/or images that attack, demean, or hurt specific individuals or groups.

Definition of Hate Speech
Hate speech includes any language and/or imagery that:
- Attacks, demeans, or hurts individuals or groups based on attributes such as race, religion, ethnicity, gender, sexual orientation, disability, or other characteristics.
- Uses derogatory language, slurs, offensive stereotypes, or hateful imagery (including aggressive remarks presented in a sarcastic tone).
- Incites violence or hostility against individuals or groups.

Annotation Categories
1. **Hate Speech (HS)**
2. **Not Hate Speech (NHS)**

General Guidelines
1. **Examine Both Modalities**: Consider both the text and the image in each tweet.
2. **Context Matters**: Consider the context in which the tweet was written and the imagery used. Look for implicit meanings and references.
3. **Neutral or Ambiguous Content**: If the content is neutral or ambiguous, classify it as Not Hate Speech.
4. **Satire and Sarcasm**: Pay attention to satire and sarcasm. Even if a tweet seems humorous, it may still be classified as hate speech if it demeans or attacks someone.
5. **Direct vs. Indirect Hate Speech**: Direct hate speech includes explicit attacks, while indirect hate speech may include coded language, implied hostility, or imagery.

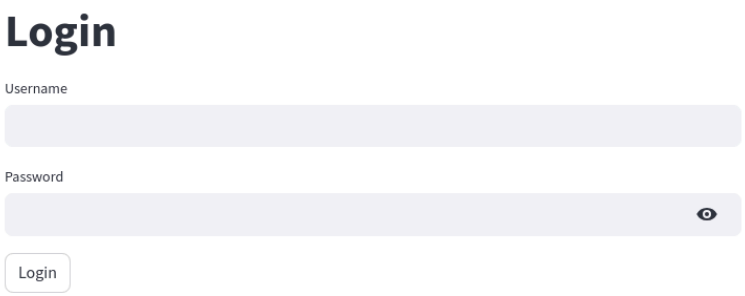
Specific Guidelines
1. **Target Identification**
    - Please double-check whether the given target is already appropriate
2. **Language and Tone**
    - Derogatory words could become a signal, but please check the context since there is also a possibility of false positives.
3. **Imagery Analysis**
    - Evaluate images for hateful symbols, gestures, or depictions that demean or attack individuals or groups.
    - Look for memes, cartoons, or edited images that may convey hostility or derogatory messages.
    - Assess the context in which images are used alongside the text.
4. **Text Information**
    - Put on the notes part or words of sentences or tweets that represents emotions related to hate speech. It can be also swear words.

- Example
  - “Saya **membenci** [specific group]”
  - “[specifig group] memang yang **paling top**, **sampe ngga bisa bedain mana yang bener mana yang salah**”

5. **Examples of Hate Speech**
   - Text with explicit statements like " Saya benci dengan [specific minorities]" or "Semua [specific group] harus [violent action]."
   - Images containing slurs, offensive symbols, or hateful depictions.
   - Combined text and imagery promoting violence or hostility, e.g., " Kita harus memusnahkan semua [specific group]" with an image of a violent act.
   - Derogatory jokes or memes targeting a specific group.
6. **Examples of Not Hate Speech**
   - Criticism without derogatory language or hostile imagery, e.g., " Saya tidak setuju dengan keyakinan [specific group] "
   - Neutral statements about groups with non-offensive imagery, e.g., "Saat ini banyak sekali [specific group] yang hidup di daerah ini " with a neutral photo.
   - General discussions on controversial topics without hostile language or imagery.

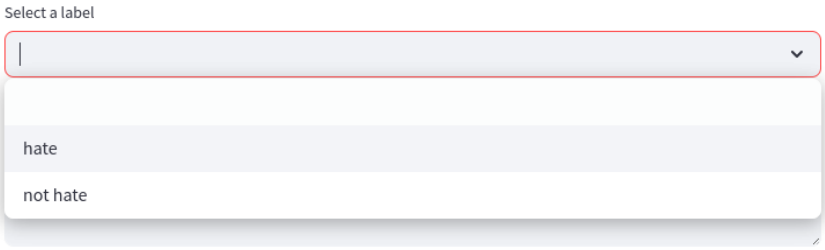
Steps for Annotation

1. **Log in and Access the Annotation Tool**
   Open the annotation tool by navigating to the designated platform where the tweets are hosted. Log into the platform using your username and password.

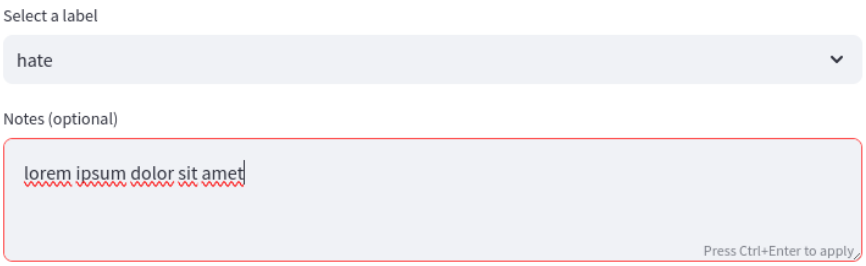


2. **Read Each Tweet and Examine the Image**
   Once logged in, you will be redirected to the main dashboard. Carefully read the content of each tweet displayed on the main screen. Examine the accompanying image by clicking on the arrow icon located at the top right corner of the image to view it in full size.

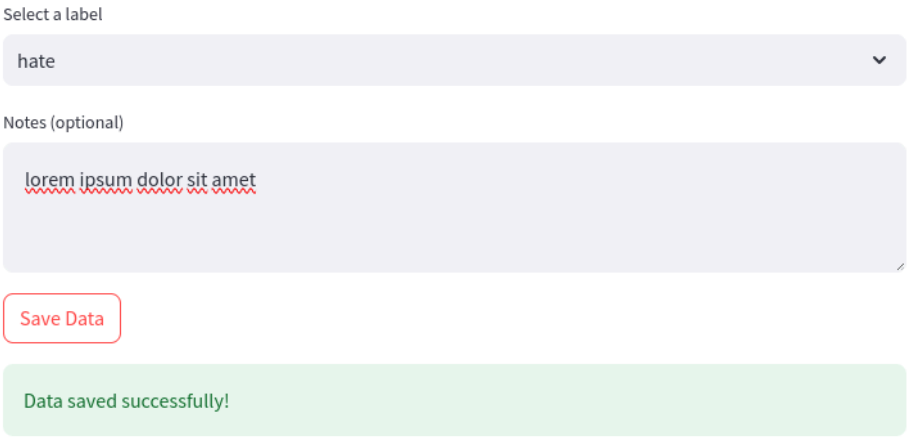
3. **Annotate the Tweet**
   Decide if the tweet falls under "Hate Speech" or "Not Hate Speech" based on the guidelines for both text and images. Select the appropriate label for each tweet using the dropdown menu next to the tweet.
   - Click on the dropdown menu labeled "Select a label".
   - Choose either "hate" or "not hate" based on your assessment.



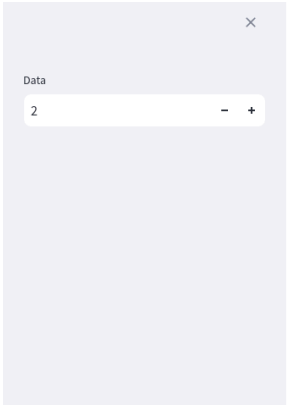
4. **Add Notes (Optional)**
   If you find any tweet particularly challenging to classify, you can add notes explaining your decision. Use the text area labeled "Notes (optional)" to provide additional context or reasoning for your annotation. This step is optional but highly encouraged for ambiguous or borderline cases.



5. **Save Your Annotations**
   After annotating the tweet and adding any notes, click the "Save Data" button to save your annotations. A success message "Data saved successfully!" will be displayed, indicating that your annotations have been saved.



6. **Navigate to the Next Tweet**
   Use the navigation controls on the left sidebar to move to the next tweet.
   - Click on the "+" button to load the next tweet.
   - Click on the "-" button to return to the previous tweet.
   - Repeat the process for each tweet in the dataset.



Quality Assurance

- **Consistency**: Ensure consistent application of these guidelines across all tweets.
- **Review**: Periodically review your annotations to maintain high quality.
- **Feedback**: Provide feedback if you encounter any unclear or ambiguous tweets.

Contact Information

If you have any questions or need further clarification, please contact Endang Wahyu Pamungkas at ewp123@ums.ac.id.