# Resources for Annotation Experiment

This document provides the annotation prompts and guidelines used for the illustrative annotation experiment in our article on “Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks”.

In the experiment, annotators are randomly assigned to one of three equal-sized groups. Their task is to distinguish between hateful and non-hateful content. Each group will label the same dataset of 200 short Twitter text posts, but with different *annotation prompts*.

Group 2 (the ‘prescriptive paradigm’) will also be given *annotation guidelines*, which contain more detailed instructions on how to annotate content. Group 2 will be asked to review the guidelines before and during the annotation task.

## Group 1 - DESCRIPTIVE PARADIGM

**Prompt**: **“**Imagine you come across the post below on social media. **Do you personally feel this post is hateful?** We want to understand your own opinions, so try to disregard any impressions you might have about whether other people would find it hateful.”

## Group 2 - PRESCRIPTIVE PARADIGM

**Prompt**: **“**Imagine you come across the post below on social media. **Does this post meet the criteria for hate speech?** We are trying to collect objective judgments, so try to disregard any feelings you might have about whether you personally find it hateful.”

**Click here to view the criteria: LINK**

## Group 3 - CONTROL (prescriptive w/o guidelines)

**Prompt**: **“**Imagine you come across the post below on social media. **Does this post meet the criteria for hate speech?** A post is considered hate speech if it is 1) abusive and 2) targeted against a protected group (e.g., women) or at its members for being a part of that group.”

# Annotation Guidelines (only shown to Group 2)

We define **hate speech / hateful content** as abuse that is targeted at a protected group or at its members for being a part of that group. Protected groups are those based on race (e.g., black people), gender (e.g., women), sexual orientation (e.g., gay people) and disability.

More **specific criteria** are given below. Please refer to them for guidance as you work through the data. They will help you make the right label choices.

## Criteria for HATEFUL Content

The **use of abusive slurs** to refer to a protected group is defined as **HATEFUL**. Abusive slurs include "**faggot**", "**fag**", "**retard**"/"**retarded**", "**dyke**", "**tranny**", "**nigger**", "**coon**", insulting people as "**pussy**" or "**queer**", "**cunt**", "**whore**", "**slut**", "**spic**", "**beaner**", "**chink**", and "**kike**".
*Ex.: "Too many faggots in this country" → HATEFUL*
*Ex.: "Shut up, pussy" → HATEFUL*
*Ex.: "She is such a slut whore" → HATEFUL*

## Criteria for NON-HATEFUL Content

The **use of reclaimed slurs** such as "**nigga**" is defined as **NOT HATEFUL**. Reclaimed slurs are slurs that are used within the communities they were originally targeted at, to refer to other members of their community. A good indicator is the use of spelling variants ("**nigga**", "**nicca**", "**nigguh**"), rather than the likely hateful "nigger". The content itself might still be negative, but is not hate speech.
*Ex.: "Niggas been on my shit all day" → NOT HATEFUL*
*Ex.: "Me and my friends are bad bitches" → NOT HATEFUL*

**Abuse against white people** is defined as **NOT HATEFUL.** This includes the use of slurs such as "**white trash**", "**redneck**", "**honky**", "**whitey**" or "**hillbilly**".
*Ex.: "Fuck those honky rednecks" → NOT HATEFUL*

**Interpersonal abuse** that does not reference membership in a protected group and does not use hateful slurs (see "hateful slurs" above) is defined as **NOT HATEFUL**. This includes uses of the word "**bitch**" that are not directed at women.
*Ex.: "He's such a little bitch" → NOT HATEFUL*
*Ex.: "Fuck off, you piece of trash" → NOT HATEFUL*

**Abuse against non-protected groups** is defined as **NOT HATEFUL**. Non-protected groups include professions and affiliations (e.g., political ideology) and reviled groups (e.g., racists).
*Ex.: "I hate the god damn communists" → NOT HATEFUL*