

Preparation for Annotation Task

Please read this document carefully. Before we begin explaining the task, there are a few things you must keep in mind.

1. Please ensure that you have filled out and submitted the [consent form](#).
2. Please note that many of the tweets you will be annotating contain foul language, and hostile sentiments sometimes towards religious, gender, and racial minorities.
3. If at any point you do not wish to continue, please do not hesitate to quit.
4. You will be paid £30 on completion of 200 tweets. We understand that reading malicious social media content can be distressing. Hence, if you are unable to annotate 200 posts, you will be paid on a pro-rata basis.
5. The task does not need to be completed in one sitting. You can take breaks as and when you need. We do however request that you finish the task within 3 days.
6. The task should take approximately 2 hours to complete.

You will need to classify tweets based on the definitions we offer, using your own judgement.

As the first step in our project, we are trying to build a new dataset for the hostile tweet classification task. The tweets are all related to UK politicians. They are either replies to or retweets of a politician's tweet, or they mention the politician.

Annotation Guidelines

In this task, you will be shown tweets. For each tweet, you will be asked to perform 2 steps.

1. Part A: you will be asked to decide whether you think the given tweet is **Hostile** or **Not Hostile**, and how confident you are about your decision.
2. Part B: you will be asked to decide if the tweet refers to identity characteristics - **Race/Ethnicity**, **Gender**, and **Religion**, and how confident you are about your decision.

Below you will find detailed definitions and descriptions of the whole task.

Part A

After reading the tweet thoroughly, you will need to decide whether a tweet is hostile or not hostile based on the following definitions.

Hostile: a hostile tweet expresses hostility towards a targeted group or individual. It is intended to be derogatory, abusive, threatening, humiliating, insulting, or inciting violence or hatred towards an individual/members of the group. Criticism need not always be hostile.

Examples:

- @johnmcdonnellIMP Fuck off back to your cubbyhole you vile, terrorist-enabling, antisemitic swine.
- @RishiSunak and @COP27P Put back on your leash were you? There's a good boy
- @10DowningStreet @SuellaBraverman and @ukhomeoffice I've spewed up ropey kebabs after a pub crawl, that are more honourable or legally qualified than Sue Ellen Braverman Fernandes
- @trussliz Honey. We should have known you would fail, you couldn't even curtesy for the queen properly - you looked like you was trying to head a ball 🤡

Not hostile: posts that are not hostile. Think not just about the words appearing in a given tweet but about the context in which they were used; the presence of a particular word, however offensive, does not necessarily indicate that a tweet is hostile.

Examples:

- @DavidLammy and @UKLabour Do you agree with Starmer that Brexit is safe in Labour's hands? Genuine question. It also appears you've purged your timeline of any mention of the Brexit you fought so hard against. If so that's probably the most unprincipled, underhand thing I've come across on Twitter.
- @Nus_Ghani You have to be fucking kidding me, wtf is going on.

Part B

If you selected “not hostile” in part A, please select the option “none of the above” in this step.

If you selected “hostile” in part A, in this step, you will need to decide whether a tweet refers to an identity characteristic or not, based on the following definitions. If a tweet refers to multiple characteristics, please select all that apply.

Race/ethnicity: hostility directed at a person/group based on racial background/ethnicity. This includes discrimination based on somatic traits (e.g. skin colour), origin, cultural traits, language, nationality, etc.

Examples:

- @Nus_Ghani Your in England 🇬🇧 speak bloody ENGLISH!
- @HackneyAbbott They were NOT abandoned, they LIED as usual, they said they had relatives in the area and were dropped off so their "Relatives" could take responsibility!! But of course they didn't have anyone, liars, rapists and murderers, send them back home
- @RishiSunak An indian man in charge of the UK? Europe is dying and this is just one of its very first symptoms. Please lay the decision on the people and get this immigrant out of the presidency. Labours are shameful but this? This is just a humiliation for this country.
- @sajidjavid You you exactly what we don't want. This once proud country is being destroyed by multiculturalism. Then we are then called racists because we don't want it.

Gender: hostility directed at a person/group based on their gender. This includes negative stereotyping, objectification, using gendered slurs to insult, and threats of sexual nature.

Examples:

- @HackneyAbbott How is someone as racist and vile as u permitted on social media ? U have clearly never been raped and no doubt had a blessed life and think I have a clue #shameful come on karma where r you!
- @DrRosena You disgusting #Bitch how dare you claim to know what poverty is like. Watch this <https://t.co/VBZupAjqw7>
- @trussliz If you can't stand the heat get the hell out of the kitchen next time elect a man to be prime minister, Liz Truss just showed us there are things women can't do.
- @AnumSNP the only good thing about your country is that women are put in there place and they arent allowed to do the dumb shit you do

Religion: hostility directed at a person/group based on their religious beliefs. This includes misrepresentation of the truth and criticism of a religious group (without a well-founded argument).

Examples:

- @AnumSNP @theSNP Anum qaisar fu@\$king leave scotland go to isis in your country we dont need threat
- @RishiSunak Just make sure Hindus are safe there. I have my family residing in London, they all were scared by the constant Islamists attack in UK. For the first time I heard them complaining why did they leave India back then. We don't want any favour

- from Britain just look after our people
- @Nus_Ghani sick of you tweeting about muslims or any other religion. You are my mp & i want answers on #cummings not that bullshit we had to pay our tv licence for tonight. You all just bullshit & your silence speaks the same bullshit, but its ok as Ramadan is over?!?!
- @Nus_Ghani For once can our government put the people of this country first? I can already guess what your solution is, you want to bring them to the Uk, if so you can fuck right off mate, Britian will never be an Islamic Nation.

None of the above: if the posts are not hostile or are hostile but do not refer to gender, race/ethnicity or religion.

Examples:

- @trusslis is the worst human being. I wish someone would shoot her (**hostile**)
- @RishiSunak will make a bad PM. Please don't turn this into a race war. Please notice that he is a terrible politician (**not hostile**)
- @pritipatel @za_63 Oh fuck off Patel you couldn't give a shit (**hostile**)
- @Conservatives and @Jeremy_Huntand having the highest level of taxation in over 70 years. You've lost my vote! (**not hostile**)

Confidence Rating

In this step, please indicate how confident you are about your labelling for a tweet. The confidence scores range from 1 to 5 and hold the following meaning.

5 - extremely confident (I'm certain without a doubt.)

4 - fairly confident (I'm confident, but there might be a small chance other annotators may label it as a different category)

3 - pretty confident (I'm pretty sure, but there might be a high chance other annotators may label it as a different category)

2 - not confident (I'm not sure, it could belong to this or another category)

1 - very low confidence (I'm really unsure, it might belong to another category instead)

Please comment on your decision in the Comment box if your confidence score is 3 or lower. For example, if the tweet might contain sarcasm or humour (that is difficult to identify) or might lack the context that would make it clear.

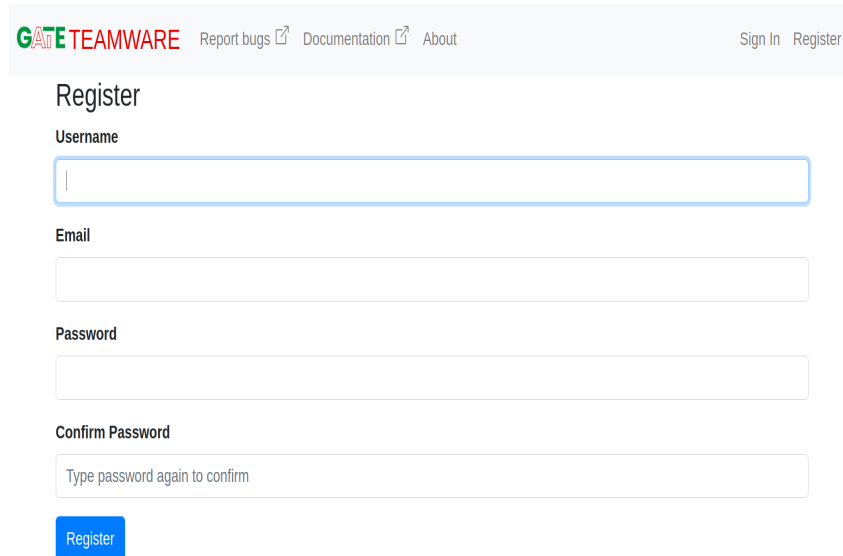
More examples with labels and explanations can be found [here](#).

If you have any questions or would like a demonstration please email Mugdha Pandya at mugdha.pandya@sheffield.ac.uk.

Platform Explanation

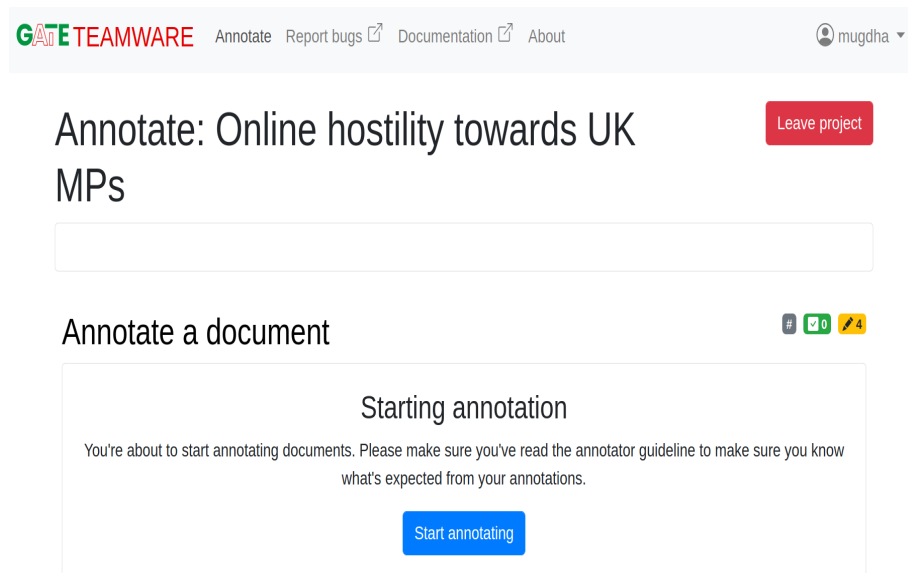
This task is run on Teamware which is an annotation platform created by the GATE team. Here are the steps to start annotating.

1. Register on [Teamware](#)



The screenshot shows the GATE TEAMWARE registration page. At the top, there is a navigation bar with the GATE TEAMWARE logo, links for 'Report bugs', 'Documentation', and 'About', and buttons for 'Sign In' and 'Register'. The main heading is 'Register'. Below it, there are four input fields: 'Username', 'Email', 'Password', and 'Confirm Password'. The 'Confirm Password' field has a placeholder text 'Type password again to confirm'. At the bottom, there is a blue 'Register' button.

2. Send your chosen username to Mugdha Pandya at mugdha.pandya@sheffield.ac.uk
3. You will receive an email informing you that you have been added as an annotator
4. Sign in, navigate to the annotate tab and select the 'Online Hostility Towards UK MPs' project.



The screenshot shows the GATE TEAMWARE annotation interface. At the top, there is a navigation bar with the GATE TEAMWARE logo, links for 'Annotate', 'Report bugs', 'Documentation', and 'About', and a user profile icon labeled 'mugdha'. The main heading is 'Annotate: Online hostility towards UK MPs'. Below the heading, there is a red 'Leave project' button. Underneath, there is a large empty text box. Below the text box, there is a section titled 'Annotate a document' with a small icon showing a document with a pencil. Below this, there is a box titled 'Starting annotation' with the text 'You're about to start annotating documents. Please make sure you've read the annotator guideline to make sure you know what's expected from your annotations.' and a blue 'Start annotating' button.

5. Go through the test phase (20 tweets)
6. Start annotating

Test phase

After going through the definitions and example sheet you will need to annotate 20 tweets as a test. This is to ensure that you have understood the definitions as we hope. Once you pass you can automatically start annotating.

Below is a screenshot of the Teamware interface for the task. Please note that the page does not have any refresh animation but the tweet changes once you click “submit”.

<p>TWEET TO ANNOTATE</p>

Hostility categorisation

Please select the hostility type

☐ Hostile ☐ Not Hostile

Hostility Label Confidence

Please select your confidence

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Comment on what made you uncertain

NOTE: Please fill this if you selected confidence 3 or below

Identity Characteristic Selection

Please select the target identity characteristic(s)

☐ Religion ☐ Gender ☐ Race/Ethnicity ☐ None of the above

Identity Characteristic Confidence

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Comment on what made you uncertain

NOTE: Please fill this in if you selected confidence 3 or below

Submit **Clear**

Extra Information

1. [Here](#) is some information about the MPs you will come across in the tweets.
2. Please feel free to look up any unfamiliar slang or words that you come across in the tweets.