The Contextual Abuse Dataset | Codebook

Key advice for completing annotations

1. Entries which are marked as 'Neutral' can only be assigned to that one category.
2. For all other categories, each entry can fall into several of the primary categories and may, in some cases, fall into the same category several times (For example, one entry could contain different types of identity-directed abuse).
3. Annotations are made on any English text in the entry, do not translate text which is non-English. Look out for non-English terms which are widely used in English; such as 'uber' for 'very' and 'wunderbar' for wonderful.
4. Do not click on links. This is because (1) many people in the Subreddits don't click on the links, and (2) it introduces lots of 'noise' for subsequent analyses. However, reasonable inferences can be made about the content of links based on how they are discussed in the thread.
5. Consult sources such as such as Urban Dictionary, Conservapaedia, Reddit and Wikipedia. This is particularly important if you encounter internet-specific slang or terms you are unfamiliar with.
6. Your goal is to apply the coding guidelines presented here. Apply them dispassionately and try to mitigate any personal biases you hold about particular groups. The 'truth' of entries should not impact how you make the annotations. For instance, if you personally agree with a statement that is abusive about women, you should still mark it up as identity-directed abuse against women.

Reviewing images

Review all images that have been flagged in the Excel Sheet (these will all be provided to you separately in a .zip file). Images contain vital information and often set the tone for a conversation. Annotate images based on how they are deployed by the user. This will require careful examination. Because you cannot highlight an image in our interface, you must write out the relevant bits of text in the 'highlighted' field.

One special case to look out for is when images contain abusive content which has been shared by a user *who then distances/critiques that content*.

- For instance, someone might share a screenshot of a conversation between two users in which the users engage in abuse. This *by itself* should not be marked up as abusive. The person we are annotating for is the poster in the conversation thread. Sharing interpersonal abusive content produced by others is not necessarily abusive. However, the author might then comment on the image, and by doing so express abuse themselves.
- If an author shares an abusive screenshot, image or meme but does not say anything else (i.e. they just post the content on its own) then it should be marked up as abusive. *If an author shares abusive content but does not distance themselves from it then the entry should be marked up as abusive.*

Bots

In some cases, content has clearly generated by a bot (it may even have the term ‘bot’ in the author’s username). This content should be marked up as though it were from a real user. Sometimes this will be difficult because the content is less coherent.

Abusive usernames
Usernames can be a vehicle for expressing abuse (e.g. u/FuckN1G6A$). However, within these guidelines, usernames should not be marked up. Only mark up the content of entries, do not base your assessment on the username.

Codebook

1.1 | Neutral

- Content which does not fall into any of the other categories outlined in the codebook.
- It is usually entirely unrelated to abuse, hate, prejudice or intolerance (e.g. 'Had a great ice cream earlier' or 'Fossil fuel crisis won't solve itself!').
- Nothing further needs to be done with entries which fall into this category.
- You do not need to label the context or secondary categories for Neutral content.

1.2 | Identity-directed abuse

Identity-directed abuse directs abuse at an identity. An 'identity' relates to fundamental aspects of individuals' social position, community and self-representation. What counts as an 'Identity' is open-ended and is not based on what platforms moderate for or legal constraints (e.g., the 'protected characteristics' in the UK). It includes but is not limited to:

1. Religion
2. Race
3. Ethnicity
4. Gender
5. Sexuality and Sexual preference
6. Immigration status
7. Nationality
8. Ableness / disability
9. Physical appearance, including weight and baldness
10. Class

Key points for marking up identity-directed abuse:

- The targets of identity-directed abuse should be *inductively* identified and entered as free text. We will then clean them up afterwards for consistency.
- Identity-directed abuse also includes intersectional identities, such as 'White Women' or 'Black Gay Men'.
- Note that political affiliations are included under Affiliation-directed abuse.
- If you see a group being targeted which includes both an Identity element and an Affiliation element (e.g. Black Conservatives) then this should be marked up as Identity-directed abuse.
- Multiple facets of identity might be targeted in one entry (e.g. 'fucking women cause all our problems, and big darkies like Diane Abbot are the worse'). In this example, two annotations would be made, one for abuse against women and one for abuse against black people.

We distinguish between five secondary categories of identity-directed abuse, outlined below (1.2.1 to 1.2.5).

1.2.1 Threatening language

Threatening language is language which either expresses an intent/desire to inflict/cause a group to suffer harm, or expresses support for, encourages, advocates or incites such harm. Harm includes physical harm/violence, criminal damage, intimidation/harassment, emotional abuse, financial harm, systematic exclusion from public discourse, and doxing (e.g. sharing of private information online).

- The conceptual basis of 'threatening language' is *action*. If a verb and/or an expression of intent is used (e.g. 'I will…', 'I want to…', 'I am going to….') or a normative statement that *others* should (e.g. 'They should be…' or 'We should'), then it is likely that threatening language is involved.
- There must be a group that is 'under threat' for an entry to be considered threatening language. If a user generically uses threatening language (e.g. 'I want to shoot everyone') but a group is not identified, it does not fall within this category.
- If one entry expresses a threat (e.g. 'I want to shoot some Muslims') and later entries express support for it (e.g. 'Yes mate, go do it!' or 'Right on!! You know it') then they should also be marked up as threats. The support that subsequent entries show should be explicit and completely unambiguous, as with these examples, to be marked up as 'Threatening language'.

Subtypes of threatening language (which do not need to be annotated separately, but are useful to have in mind) are:

- Making a threat: personally expressing intent, desire or willingness to commit harm (e.g. 'I want to stab all X' or 'I want to rape every….'). This includes language which is threaten*ing*, even if an explicit threat is not made (e.g. 'Those X are unwelcome, if I see one I don't know what I'll do…').
- Advocating that harm should be committed against a group:
    - *Incitement*: calling for another person to commit harm, either now or in the future (e.g. 'you should attack X' or 'you would be better off blowing them up'). This can also be general and not addressed to a specific person (e.g. 'we should do something to get rid of these animals').
    - *Support*: endorsing or encouraging that harm is committed, either now or in the future, but not directed towards another individual as with incitement (e.g. 'I think gassing X wouldn't be such a bad idea' or 'X deserve to be shot'). Often, these might be normative statements (e.g. 'They had better fucking watch out' or 'They should be banned').

1.2. 2 Glorification of hateful entities

Glorification is language which glorifies, embraces, justifies or supports hateful actions, events, organizations, tropes and individuals (which, collectively, we call 'entities'). In all cases, the entities should be unambiguously hateful (e.g. Hitler, the Holocaust, Rwandan Genocide or Apartheid). When the entity is considered hateful by some people, but this is deeply contested by others, then it should not be considered identity-directed abuse (e.g. expressing support for Donald Trump). Glorification includes:

- Endorsing and supporting hateful entities.
- Denying that identity-based atrocities took place.
- Encouraging and advocating that hateful entities receive support, such as through recruitment and/or financial assistance.
- Uncritically/supportively using symbols associated with hateful groups (e.g. the Swastika).

Glorification does not need to be 'weaponised' for it to be considered a form of identity-directed abuse. For instance, glorifying Nazis is itself a form of identity-directed abuse, even if it is not accompanied by an explicit statement against minorities.

The 'target' in Glorification is the entity which is being glorified rather than the group being attacked. This is because it is often difficult to identify which group is being attacked (e.g. the

Nazis committed atrocities against not only Jewish people, but also Roma, Gypsy, the Disabled and racial minorities).

Sub-types of Glorification (which do not need to be annotated separately, but are useful to have in mind) are:

- Hate organizations, strictly defined (e.g. the Ku Klux Klan):
    1. Glorifying an explicitly neo-Nazi or white supremacist news provider/forum (e.g. the Daily Stormer or Stormfront) should also be included here. However, sharing the content of such sites would not necessarily be glorification on its own – there must also be a clear positive statement about the content. Additionally, glorification of more politically ambiguous news providers (e.g. Breitbart) would not count as Glorification on its own but should alert you to other forms of abuse that might be expressed concurrently.
- Hateful events (e.g. the Rwandan genocide). This includes rejecting that well-established historical atrocities took place (e.g. Holocaust denial).
- Hateful acts (e.g. hate crimes, rallies/protests by overtly hateful groups and terrorist attacks targeting particular groups). This should be *very* narrowly defined – comments which discuss voting for the populist right party UKIP, for example, should not be included.
- Hateful individuals (e.g. Hitler, Mussolini, Pol Pot or David Duke).
- Hateful tropes, e.g. 'They Will Not Replace Us' and 'White Sharia' – respectively, the Alt-right and white supremacists.

1.2.3 Derogation

Derogation is language which explicitly derogates, demonizes, demeans or insults a group. Most of this content will be *descriptive*: it describes how the author perceives things to be or expresses an opinion about how things are. Remember that you do not need to make any judgement about the truth or falsity of content.

Sub-types of Derogation (which do not need to be annotated separately, but are useful to have in mind) are:

- Negative representations: Representing/discussing/portraying a group in extremely negative terms, such as portraying the group as evil. It also includes moral statements (e.g. 'X are wrong' or 'it's not okay to be X'). Further sub-types:
    1. *Absolute statements* of negativity (e.g. 'X are lowlife scum' or 'X are thick as pigshit)'. This can include the use of negative stereotypes.
    2. *Relative statements* [Inferiority thesis]: (e.g. 'X are worse than the rest of us' or 'they aren't capable of developing those skills because their brains are less clever than ours').
    3. *Association with Negative behaviours/identities:* (e.g. 'X are all terrorists' or 'X like to fiddle kids'). This includes calling all members of an Identity Nazis or terrorists, claiming they are all prejudiced or are fascists.
- Negative emotions: Expressing intensely negative feelings or emotions about a group (e.g. 'I hate X' or 'I just really dislike X').
- Negative impact: Portraying a group as having a negative impact. This includes:
    1. *Incompatibility thesis:* Stating that a group is not welcome or is incompatible – whether due to cultural or natural reasons (e.g. 'they will not integrate and cannot be allowed' or 'you can't mix X and Y').
    2. *Evil intentions thesis:* Ascribing to a group *evil intentions*, goals and plans (e.g. 'X want to take over the country and change our way of life' or 'they will

bring about the downfall of western civilization'). It may also include stating that the group controls society; for instance, many anti-Semites claim that Jewish people control the media and big business. Primarily this line of reasoning states the outgroup poses a threat to the ingroup and/or to society as a whole.

3. *Conspiracy thesis:* Claiming that the outgroup is engaged in a well-organised global conspiracy to ruin/control/undermine society. Again, this is most widely observed in relation to Jewish people.

1.2.4 Animosity

Animosity is language which expresses abuse against a group in an implicit or subtle manner. The lynchpin of this category is that (1) the group is treated in an abusive or negative manner but (2) this is not expressed explicitly. If the negativity is explicit then the content most likely should be annotated as Derogation.

Sub-types of Animosity (which do not need to be annotated separately, but are useful to have in mind) are:

1. Undermining the experiences and treatment of identities/groups, often by mocking or ridiculing them, usually by using humour. Often, authors will mock the accent of a particular group by using phonetic spelling to imitate their accent. You should be attuned to genuine mis-spellings and one-off mistakes; we are only interested in cases where accents are mocked through obviously and intentionally mis-spelt content. When the content is read out loud, does it sound like a stereotypical mocking of the accent?
2. Accusations that a group receive special treatment and/or are undeserving of it (e.g. 'Muslims get loads of stuff in this country, it's wasted on them!'). This will often depend on the tone; some discussions about how (often ethnic minority) groups are treated are entirely non-abusive and contain no animosity.
3. Being derogatory against "Some but not all" members of a group – this is a common line of argument amongst hateful groups, which often make deeply negative statements against "some but not all" members of a group. Your judgment must be used to identify whether statements which seem non-prejudiced and justified, nonetheless still implicitly attack the whole group and cast aspersions about them.
4. Implying that a group is a threat (subtly) or that the author views them in a deeply negative way, such as by being suspicious of them. Such statements are often ambiguous and you must infer the latent meaning of the author. For example, 'You don't know the half of it, our street is full of Romanians and it's not been the same since' – with this statement you have to deconstruct the author's "dog whistle" to uncover that it is hateful. Remember that there must be some discernible negativity in the content for it to be considered animosity.

Drawing the line between animosity and non-abusive content

There must always be space for people online to discuss a group without their content automatically being labelled as Identity-directed abuse. – criticism, discussion and incivility are not the same as abuse. People online may talk about contentious subjects, such as race and religion, in critical and uncomfortable ways. However, if a discussion does not contain anything implicitly or explicitly negative against the group then it should not be marked as identity-directed abuse.

Example 1

The following post is very close to being Animosity:

> *But since we're talking about Trump, I'm like 80% sure you mean issues with immigrants, and while there are tensions, the Netherlands is one of the few West-European countries to have not suffered any major terrorist attacks this century, largely in part due to the efforts of our highly effective intelligence agency (which watched the Russians hack the DNC live via their own cameras)*

However, it does not cross the line because the author emphasizes the work of the security services for the lack of terrorism in the Netherlands, rather than linking it to immigration.

Example 2

The following post might appear as Animosity because it refers to the idea that Jewish people control society (a common anti-Semitic trope):

> *The JQ is mere distraction to divert the attention from the actual overlords, pfft….*

It could be argued that this content belittles the experiences of prejudice that Jewish people are regularly subjected to (which would be Animosity). However, this is an overreach – this content is not expressing abuse but undermining it and *mocking* the prejudicial conspiracy theory. That said, it is often the case that content like this is followed by actual abuse later in the thread so you should be alert to this possibility.

More broadly, humour is often a key part of how marginalised groups challenge hate speech and provide each other with support – we do not want to over-penalize such content by marking it up as abuse,

1.2.5 Dehumanisation

Dehumanisation is language which describes groups as insects, animals and non-humans or explicitly compares them to these. Authors must express maliciousness and show evidence of extreme prejudice and hostility against the group in question.

Dehumanisation must be explicitly negative and literal whereas Animosity is subtle and covert. Dehumanisation should not be used to implicit references, jokes or content that is difficult to deconstruct. Our goal is to capture the real extremes where there is a genuine sense that the authors views the group as less than human.

Implied dehumanisation (e.g. objectifying women by denying them individual autonomy and agency) would not count as dehumanisation on its own. For instance, ‘Women can’t think for themselves’ would not count as dehumanisation but ‘Women are basically fuckable cockroaches, you stamp one out and another will come along’ would.

- Terms that indicate dehumanisation: leech, cockroach, insects, germs, rats.
- Although the term ‘bitch’ has a non-human aspect (i.e. a female dog), we do not count this as a form of dehumanisation.

1.3 | Person-directed abuse

- Content which directs abuse against an *identifiable* person, who is either part of the conversation thread or is explicitly named in the conversation.
- Person-directed abuse is different from incivility or impoliteness and includes subcategories of annotations; these are listed below.
- In most cases, we don't know the personal relationships of people who we interact with online – assume that the content you annotate is shared by people who do not know each other.
- Sometimes, as with other types of abuse (Identity-directed abuses, in particular), people will share *another user* being abusive towards someone else: Person A shares a post in which they describe Person B being abusive towards Person C. We mark this up as abusive unless the author (Person A) has clearly distanced themselves from the abuse expressed by Person B; we need to identify a part of the entry which shows the distancing (e.g. "This is such bolox" or "Eurgh, how awful."). If we can't put our finger on something specific, then we have to annotate it as though Person A were the one expressing abuse against Person C.
- Person-directed abuse involves annotating the relational status of the person being abused. This is divided into two sub-types:
    1. Abuse about a person. Content which directs abuse at a person who is not a participant in the conversation thread. The person must be *identifiable to the people in the conversation* (i.e. there must be a genuine sense that the person under discussion is real and that the people talking about the person know of them). This is primarily identified in the text by the person being tagged (e.g. by using @ or the /u/ flag) or having their name written (e.g. "@DonaldTrump is such a w*nker" or "Helen Bonham Carter is a useless fucking c*nt").
        a. Referring to someone by their *relationship* would not qualify as abuse about a person if they are not also named. For example, statements such as 'My mum is a bitch' or 'I bloody hate my ex-girlfriend, she screwed me over' would not qualify. We do not know who the person is because they have not been named.
        b. One edge case to consider is when a very prominent public figure is described and not named – but we can easily induce their name. For instance, 'The president of the USA is a c*nt' makes a clear reference to a *nameable* person (Donald Trump). Cases such as this should be considered abuse about a person. Be strict here: only identify it as abuse 'about a person' if it is very clearly about someone who most people would be expected to know – esoteric and ambiguous references to people who are not well-known do not count.
            i. In general, be very cautious when you see a single name (e.g. Jessica, David, Gary) without a surname and if that name does not relate to someone else in the thread.
    2. Abuse to a person. Content which directs abuse at a person who is part of the conversation thread. This is usually directing attacks/insults or addressing a highly aggressive statement at them (e.g. "I hate you, you stupid bellend" or "@User has no clue what he's talking about, the daft twat").
        a. This differs from abuse 'about a person' on the basis that the victim of the abuse is part of the conversation thread, i.e. they are someone who has already made an entry.

1.4 | Counter-speech

- Content which challenges, condemns or calls out the abusive language of others. Counter-speech must be relational and a *response* to abuse already in the thread. For instance, if an opening post attacks/calls out someone in another setting (e.g. offline or from a subreddit) for what they have said then it is not counter speech for the purposes of this taxonomy. The content is only considered counter speech if it responds to what has been previously posted, whether that is either in response to a single entry or to a whole series of entries, such as a conversation thread.
- Counter speech can take several forms:
  1. Directly attacking/condemning abusive language in unambiguous terms (e.g. 'You should not say things like that' or 'that is just total nonsense').
  2. Calling out the original entry as abusive/hateful (e.g. 'that is seriously prejudiced').
  3. Offering an alternative viewpoint which is clearly meant to challenge and undermine the original post (e.g. 'That is not at all what I have experienced and I think you're completely wrong about this').
  4. Attacking the author for what they have said (e.g. 'You are a dickhead for sharing that sort of nonsense') Note that this also qualifies as person-directed abuse, 'against a person', and should also be flagged as such.
     - Mocking the original author or being sarcastic does not count as counter speech if it is expressed in a light-hearted manner or fails to seriously attack/criticise the author or viewpoint (e.g. 'Enlightening point from /r/user there!').
  5. Challenging the conclusions of the original author in clear and unambiguous terms (e.g. 'That isn't right. You don't have the evidence to say that').
- For an entry to be considered counter-speech the author must *not* engage in another form of abuse *against the same identity*. For instance, if the first post states, 'Women are all sluts who want to control men' and the second post states, 'Woah, you shouldn't be saying that!' then the second post would be counter speech. However, if the second post then continues, 'Women are not all sluts, women are just control freaks who want to ruin men's lives!' it would not be counter speech as the second author only "challenges" the abuse of the first author to then engage in an only-slightly-different form of abuse themselves. In this case, the second author has not rejected the expression of abuse but, rather, refined it.
- However, counter-speech *can co-occur with other categories in the taxonomy*, such as abuse directed against other identities or person-directed abuse. This is a nuanced point so let me elaborate further: if, for example, someone counters speech against Muslims but then is abusive against Jewish people that would be marked up as both counter speech and identity-directed abuse (e.g. 'Muslims aren't like that, you shouldn't say such things! Jews, however, actually are totally barbaric…').
- In many cases, counter speech may quote or reference abusive language. It should be clear that the author is only doing so in order to make an attack. Content which simply quotes a post and expresses shock/surprise/incredulity should not be viewed as counter-speech. Using lots of punctuation (e.g. '?!?!?!!?') or emojis (e.g. ':p') after an abusive post is still *ambiguous* as to whether the author is engaging in counter speech. We want to be very sure that what we identify as counter-speech really is counter-speech.

1.5 | Non-hateful Slurs

What counts as a slur?

- Slurs are collective nouns, or terms closely derived from collective nouns, which are pejorative. Slurs include terms which are explicitly insulting and derogatory (E.g. 'n*gga', 'spear chucker', 'kebabi' or 'retard') as well as terms which are not as explicit but nonetheless implicitly express negativity/animosity against a group (e.g. 'Rainy', 'Chad' or 'Nutter').
- By 'collective noun' we mean a term which refers to a well-defined group (e.g. women, disabled people or ethnic minorities) as well as subgroups (e.g. unattractive women, elderly black people, etc.) or affiliations (e.g. feminists, Labour party members). You must be able to identify a specific identity/subgroup/affiliation that the slur refers to. For instance, we do not include 'Cunt' or 'Dick' as they do not refer to an identifiable identity/subgroup/affiliation, despite being highly gendered.
- By 'closely derived from a collective noun' we mean terms such as 'retarded' and 'bitching'. The lemma of these words is a slur (respectively, 'retard' and 'bitch') and so they should be included in the framework.
- Some slurs are *very unusual* and may be *intentionally hard to identify*. For instance, 'Bings' is a widely used derogatory way of referring to Asian people, precisely because it is so hard to detect. Provided you are sure that they refer to an identifiable group then you should annotate such cases as Slurs.
- Slurs explicitly target the group itself and not just the activities of that group. For instance, a commonly used term amongst misogynistic communities is 'cock carousel', which is a reference for the promiscuous behaviour of women who 'ride the cock carousel' by having sex with many men. This is a *pejorative term* and its use would usually qualify as a form of identity-directed 'derogation' (see above). However, 'cock carousel' is not a slur because it references the supposed *behaviour* of women rather than women as a *group* per se.
- A slur is not the same as a profanity or 'curse word' (e.g. 'fuck', 'bugger', 'shit', 'wanker'). Such terms should not be annotated as slurs.

Non-hateful use of slurs

For this category we are only interested in marking up non-hateful uses of slurs. Hateful uses of slurs should be marked up under Identity-directed abuse. Non-hateful uses are likely to either be:

1. *Reclaimed uses*. People from the targeted group using the term self-referentially. For instance, women writing "me and all of my bitches". You should be sure that it is likely that the speaker is from the group which the slur otherwise attacks.
2. *Counter speech*. People using the slur to challenge its hateful use and/or comment on its use. For instance, "I hate being called a n*gga".

Some examples of slurs

- A non-exhaustive list of relevant slurs:
  - Rainy, Chad, Bitch, Retard, Psycho, Nutter, Loony, Lunatic, The Afric, Skanks, Whores, N*ggas, N*ggers, Spear Chucker, Kebabi, Kyke, Pale princess, Paki, Abbo, Soyboy, Chink, Coolie, Gypo, Sandn*gger, Cracker, Wog, Hick, Mulatto, Maniac.
- This list is useful for anti-disability slurs, although we do not include everything they have here, such as 'idiot': http://web.augsburg.edu/english/writinglab/Avoiding_Ableist_Language.pdf

- Not slurs:
    - Wanker, Dick, Cunt, ‘Cock Carousel’, Fakers, Losers, Insane, Crazy, Autistic (*note that Autist* would count as a slur), “Dodgy Jew”, Idiot, Moron.
- Remember that deliberately disguised or mis-spelled slurs should still be marked up (e.g. ‘Pale princess’ versus ‘Pail princess’).

1.6 | Affiliation-directed abuse

- Abuse directed against people who have a (more or less) voluntary affiliation with a profession, membership, association, ideology or other well-defined group/collective. Affiliation includes but is not limited to:
  1. Profession/occupation (e.g. Doctor, Soldier, Pilot, Cleaner, Prisoners, Retired people, the Unemployed). This includes celebrities, where they are discussed as a group.
  2. Association/membership (e.g Student Union, Sports club, trade union).
  3. Political organisation or party (e.g. Labour, the Democrats, Extinction Rebellion)
  4. Political affiliations (e.g. Feminism, Liberal, Conservatives, Socialists, Capitalists, Communists, Anarchists).

- For a statement to be affiliation-directed negativity, the negativity must be directed *at the group of people* with the affiliation and not the affiliated organisation/institution itself. To clarify what this means in practice:
  - Extreme negativity against soldiers would count as affiliation-directed abuse – but not if the extreme negativity is only directed against the Army.
  - Extreme negativity against Doctors, Nurses and other Hospital staff – but not against Hospitals or the NHS.
  - Extreme negativity against Priests, Cardinals and Bishops – but not against the Church.
  - Extreme negativity against Labour party supporters, members, candidates and representatives – but not against the party itself.
  - Extreme negativity against civil servants – but not against Government itself, including specific departments.
  - Extreme negativity against sports club supporters and players (e.g. for a Football team) – but not against the club itself.
  - Extreme negativity against activists campaigning to stop climate change, whether that is activists in general or those part of a specific group (such as Extinction Rebellion) – but not against climate change campaigns in general.
    a. This gets tricky in one use case – sometimes, people will linguistically refer to an institution or abstract concept e.g. "the police" but from the way that they are discussing the entity, it is clear that it refers to the underlying group (i.e. police people) – in which case, if the content is abusive, you should mark it up under the affiliation-directed abuse category (the same logic operates for identity-directed abuse).

- Note: Often you will find that negativity against the group of people with an affiliation overlaps with negativity against the affiliated organisation/institution itself (e.g. 'The Labour Party are wasting everyone's time with this bolox. Their members should pack up and fuck off'). So long as there is negativity against the membership as well as the Party itself then it counts as affiliation-directed abuse.
- You need to be careful about the line between abuse and incivility/criticism. Many people will be uncivil about certain affiliations but not abusive.
- We are using the same sub categories as for 'identity-directed abuse' and applying them in exactly the same way – please refer to Section 1 for details.

1.7 | Context

- For each entry, the full conversation thread is shown, including all previous posts and replies. We record whether your annotation was made based on the post alone or whether it also took into account the previous entries. This will be useful for our analysis and will enable better training of ML systems.
- For each annotation there will be an option of 'was this annotation entirely independent on the content of the previous posts?' (i.e. if you had seen this post on its own, would you have made this annotation?). If you need the content of previous posts to understand the nature of the abuse, then it should be marked up as needing context. Whether that is to understand who/what is being targeted or the way in which it is abusive, both count as requiring context.