Learning From the Worst | Codebook

CONTENT WARNING

Please be advised that this codebook contains offensive and hateful language which you could find harmful or may otherwise affect you negatively. Please seek advice and support if working with hate speech and be aware of the potential for harm at all times.

1. What is hate?

Hate speech directs abuse, negativity, contempt and derision at an identity. Following, Warner and Hirschberg (2012) our paper defines hate as: "abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation".

Given there is not a well-established single definition of hate, the other references should provide some useful context.

- Facebook defines hate as "a direct attack on people based on what we call protected characteristics". In turn, they define "attack" as "violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation" (Facebook 2020).
- Google's YouTube "removes content promoting white violence or hatred against individuals or groups" (YouTube 2020),
- Reddit updated its rules in 2020 to commit to banning "communities and users that incite violence or that promote hate based on identity or vulnerability" (Reddit 2020).
- Gagliardone et al., define hate as "expressions that advocate incitement to harm based upon the targets being identified with a certain social or demographic group" (Gagliardone et al., 2015).
- Simpson describes hate speech as 'things like identity-prejudicial abuse and harassment, certain use of slurs and epithets, some extremist political and religious speech (e.g. statements to the effect that all Muslims are terrorists, or that gay people are second-class human beings), and certain displays of 'hate symbols' (e.g. swastikas or burning crosses).' He further explains that 'We classify such activities as hate speech if, and insofar as, they convey the idea that belonging to a particular social group warrants someone's being held in or treated with contempt.'
- Nockleby defines hate as "any communication that disparages a person or a group on the basis of some characteristic" (2000, p).
- Jay defines hate as; 'bias-motivated speech aimed at a person identified as a member of a historically victimized group based on gender, sexual orientation, race, ethnicity, religion, national origin, or disability.'

Despite differences in wording and focus, there are some important commonalities across definitions of hate speech.

1. Hate involves expressing something negative, such as contempt, disparagement, derogation, demonization, harm or bias.
2. Hate is directed against the identity of a group. Identities can be understood as the social groups and affiliations that individuals belong to and are associated with [

3. Hate is an *intentional* action. Whilst much research has drawn attention to the spread of 'micro-aggressions', hate establishes a higher bar, requiring that the speaker intends to express something hateful and that it is not just a wilful mistake or an act of ignorance. Note that 'intention' is hard to discern in many online contexts and that it is the most conceptually difficult part of understanding hate.

2. What is 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:

a. Religion
b. Race
c. Ethnicity
d. Gender
e. Sexuality and Sexual preference
f. Immigration status
g. Nationality
h. Ableness / disability
i. Class

3. Types of online hate

3.1 Threatening language

Threatening language is defined by (a) taking ACTION against a group which (b) inflicts either serious or imminent HARM on them. Both these aspects must be there for content to be threatening.

Harm can include:

- Physical harm/violence
- Criminal damage
- Intimidation/harassment
- Emotional abuse and mental health problems
- Political exclusion and denial of important/fundamental rights
- Financial harm
- Doxing (e.g. sharing of private information online).

As a guide: if you see a verb being used that expresses intent or encourages action then you are most likely looking at threatening language:

- I will…
- I want to…
- I am going to…
- They should be…
- We should…

It can be difficult to separate Threatening language from Derogation. You should always think about whether ACTION is either being explicitly stated or strong implied.

Threatening language (a) often involves an 'intimidating' or 'menacing' tone, and (b) usually there will be some immediacy to the action (i.e. there should be a sense that action is going to be taken in the near-future) OR it will involve a very substantial type of harm (e.g. murder or rape or other physical assault).

<u>Making a threat</u>

The archetypal form of 'threatening language' is a statement which makes an explicit threat; it expresses *intention* to inflict pain, injury, damage or harm on a group.

For instance, it may involve expressive statements like:

1. "I am going to attack every X"
2. "I will break all of these X, they have no right to be here"
3. "Me and my friends are going to tie up the next n*gga we see."

Threatening language also involves expressing a *desire* to inflict harm on a group:

4. "I want to stab all X"
5. "I wish I could expunge all gays from round here."
6. "What I would give to ram my knife into every trannies' face."

Both intention and desire to inflict harm might be targeted at an individual, with clear reference to their identity:

7. "You goddam n*gga, Ima string you up like they did in the good old days"
8. "Fucking fag. I wish I could punch his stupid fucking face in."

Note that threats against individuals (with no reference to their identity) are NOT considered hate speech. For instance, "I'm going to punch you in the face Joey" would not count as hate. Equally, "I want to shoot everyone at my school" is worrying -- but not hateful under this taxonomy as the threat is not against a group.

Normative support for harmful action against a group

Sometimes threatening language involves a normative statement. These are remarks which morally justify, support or otherwise establish the 'rightness' of inflicting harm on a group. The normative support for threats should be very explicit in that the author is stating that harmful *action* should be taken against the group. Note that the harmful action still needs some sense of immediacy/urgency. For instance:

1. "We should attack these X"
2. "We should do something to fuck up these X"
3. "I think gassing X wouldn't be such a bad idea"
4. "X deserve to be kicked out of this country for all the harm they've done"

Threatening language can also involve *celebrating harmful actions* taken against a group. But remember that there must be either a clear sense of ACTION and imminent harm or serious harm:

1. "I hope that X drown"
2. "I love hearing about X being slaughtered"
3. "We should all be grateful that those freak Xs have had the beating they deserve."
4. "I love the idea of hitting X"
5. "I love seeing Xget bashed on the street"
6. "Does anyone else love the idea of beating all X to death or is it just me?"

Calls for harm: advocating that others inflict harm against a group

Often hate is expressed by advocating that others inflict harm against a group. If this takes the form of *explicit incitement* then it can be considered threatening language. For instance, calling for another person to attack a group, either now or in the future, is threatening. Often this content will have a menacing tone (but not always). This includes statements like:

1. "Why don't you do something about it? X need to suffer to realise how sick they are."
2. "you should attack X"
3. "you would be better off blowing them up"

Clarifications on threatening language

Finally, we want to outline a few cases that are not threatening language and explain why:

1. "I just heard some great news on the radio, they said that black people are much more likely to die from covid than any other group!" -- *this is not threatening as although there is harm there is no implied or actual ACTION. This is considered Derogation.*
2. "Saw some good news today - PHE found that disabled people are two to three times more likely to have been diagnosed with Covid-19 compared to normal people" -- *Same as the previous example.*
3. "There is so much to be proud of if you are british. I love how the majority of us collectively just will not tolerate lesbians!" -- *The link to action is very minimal, and this is Derogation.*
4. "Muslims are not welcome in our country" - *this expresses a viewpoint rather than calls for action or implies action. As such it is Derogation. if it said "Muslims are not*

welcome in this country and we should kick them all out." then it would be Threatening.

5. "I'm a huge fan of those that overtly express what we all think: jews are not welcome in our country" - *Same as the previous example.*

3.2 Dehumanisation

Dehumanisation is language which describes groups as insects, animals, trash or explicitly compares them to these -- it is a form of *dangerous hate* and is a special form of particularly concerning derogation. Dehumanization must express *maliciousness*, showing evidence of extreme prejudice and hostility against the group.

> Google defines it as: 'the process of depriving a person or group of positive human qualities.'

Dehumanisation must be explicitly negative. It does *not* refer to cases where implicit content contains references, jokes or is otherwise difficult to deconstruct. Implied dehumanisation (e.g. objectifying women by denying them individual autonomy and agency) does 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. As the folks at the Dangerous Speech project put it: "Not all language comparing people to animals or other non-human creatures is dehumanizing or dangerous, of course – it's possible to compare a person to an animal in a way that doesn't lower social barriers to violence"

Furthermore, dehumanisation is not just about making literal comparisons between a group and a non-human object (such as inanimate objects, technology, …). e.g., comparing black people with black laptops or black phones will, in most cases, NOT count as dehumanisation - although it would probably be Derogation. With Dehumanization we are looking for egregious comparisons, rather than clever word play.

Some general categories of Dehumanization to help guide you:

- Biologically subhuman – *usually, comparisons are to creatures that are considered repulsive, threatening or deserving of violence.*
    1. Vermin (rats, snakes)
    2. Beasts (apes, gorillas, pigs)
        - Although the term 'bitch' has a non-human aspect (i.e. a female dog), we do *not* count this as a form of dehumanisation.
    3. Biologically subhuman (Virus, disease, infection, bacteria, sickness, microbes, cancerous)
    4. Insects (Cockroaches, leeches, insects, 'swarm' (in some contexts))
    5. natural disaster/environmental threats [be very careful with these] (flood, tidal wave, weeds)
- Waste
    1. Trash, garbage, e.g. "cum dumpster" for dehumanization of women
- Non-human -- *please be very careful with these, and really make sure that the content is genuinely dehumanizing.*
    1. Evil spirits, monsters, demons, unnatural, inhuman
    2. Treating members of a group as property (be careful about the tcontext)

Useful resources for understanding dehumanisation:

1. https://dangerousspeech.org/guide/
2. https://theconversation.com/the-slippery-slope-of-dehumanizing-language-97512
3. https://www.vox.com/science-and-health/2018/5/17/17364562/trump-dog-omarosa-dehumanization-psychology
4. https://www.rehumanizeintl.org/badwords -- *this is a particularly helpful resource, the image below is taken from there.*

3.3 Support for hateful entities

Support for hateful entities 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). Hateful entities includes:

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

The 'target' in Support 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 Supporting Hateful Entities are:

1. Hate organizations, strictly defined (e.g. the Ku Klux Klan): 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.
2. Hateful events (e.g. the Rwandan genocide). This includes rejecting that well-established historical atrocities took place (e.g. Holocaust denial).
3. 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.
4. Hateful individuals (e.g. Hitler, Mussolini, Pol Pot or David Duke).
5. Hateful tropes, e.g. 'They Will Not Replace Us' and 'White Sharia' – respectively, the Alt-right and white supremacists.

3.4 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.

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

1. 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:
   Absolute statements of negativity (e.g. 'X are lowlife' or 'X are thick as pigshit)'. This can include the use of negative stereotypes.
   Hateful use of Slurs, most uses of slurs will be Derogation -- if there isn't a clear reason to think that the slur is either non-hateful (i.e. counter speech) then it will usually be Derogation.
   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').
   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.
2. Negative emotions: Expressing intensely negative feelings or emotions about a group (e.g. 'I hate X' or 'I just really dislike X').
3. Negative impact: Portraying a group as having a negative impact. This includes:
   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').
   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.
   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.

3.5 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 negatively but (2) this is not expressed explicitly. If the negativity is explicit then the content is Derogation.

Some types of Animosity (which you do not need to annotate separately) are:

1. *Undermining the experiences and treatment of groups*, often by mocking or ridiculing them, usually by using humour.
    a. "The Tesco advert is very diverse and everyone is celebrating it!? You can bet next time all the other supermarkets will be including disabled muslims in their adverts and being celebrated too.. Yawn"
2. *Accusations that a group receives 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.
    a. "Sometimes I think that political correctness has gone too far, why on earth is so much money given to refugees?"
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 nonetheless still implicitly attack the whole group and cast aspersions about them.
4. *Implying that the speaker has a negative view about a group* or that the author is 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.
    a. "The UK is such a wonderful country. The best universities in the world. The best music festivals you can wish for. The loveliest cottages and little villages. The friendliest people. We are so wonderful that we accept all kinds of refugees… see what I mean?"
    b. "My friend had a horrible experience with letting out her flat. The family, obviously Indians, left it in a right state. It was full of rubbish and completely filthy, I guess they should've expected it really."
    c. "My area is predominantly populated by Blacks, hence the high crime rates"
5. *Making factual or pseudo-factual statements* about groups which are prejudiced against them. If these statements are explicitly hateful then it is Derogation -- but online, prejudicial facts are often used with caveats and discussions which can make their hateful nature less obvious and would fall into Animosity.
    a. "New survey shows that 8/10 blacks in the UK do not feel like they are British."
6. *Jokes about a group which play off stereotypes.* Some entries will use distasteful jokes about a group, such as implying they are lazy or incompetent. If jokes are explicit/serious enough then they can be Derogation but many jokes, especially more subtle ones, are going to be Animosity.
7. *Attacking a specific trait of a group*. A common line of attack is to denigrate just one aspect of a group, such as claiming that breastfeeding is wrong, that Muslims praying in public is bad or that black music is loud and unpleasant. This may be superficially

justified by the specificity of the attack -- it is similar to the ‘some but not all’ approach. Yet in many cases it is a way of insulting the entire group. That said, you should use your judgement as sometimes these attacks will not be Animosity but just non-hateful criticism. You will need to examine the style of attack, the justification (if any) and the tone that is used.

A note on 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.

A note on identity-based insults

Weaponized identity referents (e.g. 'that is so gay') and pejorative terms (e.g. 'silly cunt') can be quite complex; whether they are hateful often depends heavily on the specific term and the context in which it is used, as well as other signals in the entry. This guidance is intended to help clarify your annotation decisions.

As a heuristic: if the content uses an explicit slur then it is likely to be Derogation. If it refers to a person, multiple people or a group with a pejorative term or identity-based insult then it is likely Animosity. If it is completely colloquial use and not being used against people/groups then it is probably None.

Use of identity-based insults that are likely to be derogation

1. If you see use of a slur then you are probably looking at derogation (e.g. '**fag**', '**n*ggga**', '**retard**' etc….), unless it is reclaimed. It is only reclaimed if you can clearly show evidence from the text that the speaker is of the identity that the slur attacks; even if hateful terms like '**cripple**' or '**paki**' are used jovially then they should usually be considered derogatory.
2. In some cases, you will need to have some appreciation of the 'strength' of the slur. The examples just given are very clearly hateful and in nearly all cases their use will be considered derogation -- but other slurs may not cross the line into automatically being derogatory, such as '**baby boomer**' or '**chad**'
3. Use of '**pillow biter**' is likely to be derogation against gay people.
4. Use of '**bimbo**' is likely to be derogation against women.
5. Use of '**autist**' is likely to be derogation against disabled people.

Use of identity-based insults that are likely to be animosity

1. Weaponizing an identity (or something that is very closely related to a fundamental aspect of an identity) to personally attack someone, e.g. "that is so **gay**", "Be careful, that's something a **black guy** would do", "Pfft, what are you, a **woman**?"
    a. If the entry contains more like, "that's so **weak**, what are you, a woman?" -> then it would cross over into derogation.
2. Referring to someone using a pejorative term for an identity (or a pejorative term that is closely related to an identity) negatively or in order to attack them, e.g. "you're a **cunt** if ever I saw one", "stop being such a **pussy** fam", "you are such a fucking **bitch**", "stop being a **cow**".
3. Addressing or referring to someone from an identity in a negative way by using a pejorative term for that identity, e.g. "these stupid **bitches** were causing me trouble", "this **slut** came up to me and tried to get it on, but i dont fuck **whores**", "some **slappers** were walking through town".
4. Use of terms which explicitly other and ostracize a group, e.g. "the **goyem** know the truth"
5. Use of "tard" as a suffix, e.g. "**libtard**", "**fucktard**", "**femtard**". This is weaponization of the anti-disability slur "retard" to attack another identity. Such terms do two things: (1) attack the targeted person/group (respectively, liberals, a person, feminists) [which is not hateful if the person/group is not an in-scope identity!] and (2) attack people with disabilities. In general, we would not consider this to be *derogation* given that the anti-disability prejudice is implicit in the attack against the person/group
6. Use of **gyp** is likely to be animosity against gypsies, e.g. "I got gyped off when i bought this car".

7. Use of **japs eye** is likely to be animosity against East Asian people, "my jap's eye hurt like hell."
8. Terms which are likely to be animosity against Disabled people: **cretin**, **lunatic,**
9. Terms which are likely to be animosity against people with mental health problems (assigned to the Disability target): **lunatic, psycho, wacko, nutter.**
10. Use of **cock sucker** is likely to be derogation against gay people. It could easily become *derogation* if these is some other attack against gay people in the content. See: https://www.thegayuk.com/is-cocksucker-a-homophobic-insult/

Use of identity-based insults that are *by themselves* likely to be None

1. Use of terms which are pejorative to an identity but are used in a non-attacking context, such as "quit your **bitching**", "**bitch**, please", "what a **cunty** thing to say", "stop being a **moron**", "what an **idiot**", "the economy was **crippled** by poor management", "those celebrities are **whoring** for money",
2. Use of terms which have an identity-based element but it is not clear that they are prejudicial, often because the lynchpin of the attack is not usually the identity but rather a perceived deficiency in the person -> *there are fairly few of these out there but they do exist!* e.g. "you are a **mother fucker**", "stop being such a **twat**" or "look at these **sons of bitches**"
3. Use of terms which are somewhat pejorative to an identity -- but are used in a clearly positive / non-negative way, e.g. "You have such an amazing sex life, you are such a **slut**."
    a. Note that in our meeting on Friday 11th December we discussed a similar example but it was not clearly positive enough; I believe it said, "You have such an active sex life, you are such a slut" -> this is ambiguous and, on balance, there is not enough positivity here so it should be considered animosity.
    b. Note that 'positive' use of explicit slurs (e.g. paki, fag, nigga) is still considered derogation. This rule only applies to pejorative referents where there is some question mark over their hatefulness (e.g. bitch, whore, cunt).

Note that even if an entry contains identity-based insults which are by themselves None or Animosity the entry itself could easily still be considered Derogation, Threatening or even Dehumanization; it depends on what else is in the entry. *You must assess every entry as a complete statement.*
The examples we have provided here are based on the (probably unfaithful) assumption that there is nothing else in the entry which is hateful in any way. For instance, all of these uses cross over from animosity into other types:

1. "I wouldnt even rape her, the **bitch**" (Derogation)
2. "Those **gays** are such nonces" (Derogation)
3. "Your wife is a dumb **cunt**, like all women" (Derogation)

Always apply your judgement and critically evaluate what you are annotating at all times!