**Summary of literature on HS/CS**

**(1) Definitions of Hate Speech**

UN:

*“****Offensive******discourse******targeting*** *a* ***group*** *or an individual* ***based******on******inherent******characteristics*** *(such as race, religion or gender) and that may threaten social peace.”*

🡪 very vague, implicates that hate speech is grounded in real characteristics

🡪 threaten social peace??

🡪 doesn’t explain much

Brittanica:

*“speech or expression that denigrates a person or persons on the basis of (****alleged****) membership in a social group identified by attributes such as* ***race****,* ***ethnicity****,* ***gender****,* ***sexual orientation****,* ***religion****,* ***age****,* ***physical*** *or* ***mental disability****, and others.”*

🡪 a lot more specific, shifts focus on the effect it has on the target of HS  
🡪 alleged: doesn’t have to be grounded in reality

Wroclaw University of Science and Technology:

*“ form of direct discrimination which involves dissemination of verbal and non-verbal* ***statements*** *and messages that* ***spread****,* ***propagate*** *and* ***justify******hatred******against*** *specific individuals or groups, most often distinguished by characteristics such as: skin colour (****racism****), ethnic origin (****xenophobia****), nationality (****chauvinism****), gender (****sexism****), gender identity (****transphobia****), psychosexual orientation (****homophobia****,* ***heterophobia****) age (****ageism****), world view (e.g.* ***anti-Semitism, Islamophobia****)”*

* overview of the most prevalent forms of HS
* functions of HS: “Spread, propagate and justify hatred”

Combining Brittanica and Wroclaw:

“form of **direct** **discrimination** which involves dissemination of **verbal** and **non-verbal** statements and **messages** that **spread**, **propagate** and **justify** **hatred** against specific individuals, most often on the basis of (alleged) membership to a social group, most often distinguished by characteristics such as: skin colour (**racism**), ethnic origin (**xenophobia**), nationality (**chauvinism**), gender (**sexism**), gender identity (**transphobia**), psychosexual orientation (**homophobia**, **heterophobia**) age (**ageism**), world view (e.g. **anti-Semitism**, **Islamophobia**)”

**(2) “Gruppenbezogene Menschenfeindlichkeit (GMF)”**

At this point it makes sense to distinguish two different general forms of hate (speech), based on how the addressee of hate (speech) is targeted: (1) targeting the individual through characteristics or behaviors associated (not with their (alleged) membership to a social group but) with the individual itself and (2) targeting the individual or group based on the (alleged) membership to a social group. In German discourse the second kind of hate (speech) is often referred to as “Gruppenbezogene Menschenfeindlichkeit” (GMF) which loosely translates to “group-focused enmity”. The Amadeu Antonio foundation is regarded as an expert in german-speaking discourse on fighting against right-wing extremism, racism and antisemitism through initiatives and projects. On their website, they present an overview including definitions of the different forms of GMF, which can function as a starting point for the prompts used in our media lab project.

1. **Racism**
2. **Generally**

UN: “*The term "racial discrimination" shall mean any distinction, exclusion, restriction, or preference based on race, colour, descent, or national or ethnic origin that has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life*.”

1. **Anti-muslim racism**

bpb: “Islamophobia is any distinction, exclusion, or restriction towards, or preference against, Muslims (or those perceived to be Muslims) that has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life.”

1. **Anti-black racism**

Boston-University: “specific kind of racial prejudice directed towards Black people, or those perceived to be Black. More than merely a product of a few bad actors or individual prejudice, antiBlack racism is systematic and structural. It systematically marginalizes Black people and communities, and devalues Blackness in all contexts”

1. **Anti-asian racism**

Mediendienst Integration: “Anti-Asian racism refers to prejudices and racist narratives against people who are considered to be of Asian origin. It usually involves references to South, Southeast and East Asian countries. The prejudices against Asian people go in different directions: On the one hand, they are stigmatized as "different", "dangerous" or "homogeneous" and, for example - as in the corona pandemic or the SARS pandemic in 2002/2003 - blamed for the spread of diseases. Other narratives speak of "hard-working model migrants". They are therefore often denied that they experience racism.“

1. **Antisemitism**
2. **Generally**  
   IHRA: Antisemitism is a certain perception of Jews, which may be expressed as hatred toward Jews. Rhetorical and physical manifestations of antisemitism are directed toward Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities.

* Calling for, aiding, or justifying the killing or harming of Jews in the name of a radical ideology or an extremist view of religion.
* Making mendacious, dehumanizing, demonizing, or stereotypical allegations about Jews as such or the power of Jews as collective — such as, especially but not exclusively, the myth about a world Jewish conspiracy or of Jews controlling the media, economy, government or other societal institutions.
* Accusing Jews as a people of being responsible for real or imagined wrongdoing committed by a single Jewish person or group, or even for acts committed by non-Jews.
* Denying the fact, scope, mechanisms (e.g. gas chambers) or intentionality of the genocide of the Jewish people at the hands of National Socialist Germany and its supporters and accomplices during World War II (the Holocaust).
* Accusing the Jews as a people, or Israel as a state, of inventing or exaggerating the Holocaust.
* Accusing Jewish citizens of being more loyal to Israel, or to the alleged priorities of Jews worldwide, than to the interests of their own nations.
* Denying the Jewish people their right to self-determination, e.g., by claiming that the existence of a State of Israel is a racist endeavor.
* Applying double standards by requiring of it a behavior not expected or demanded of any other democratic nation.
* Using the symbols and images associated with classic antisemitism (e.g., claims of Jews killing Jesus or blood libel) to characterize Israel or Israelis.
* Drawing comparisons of contemporary Israeli policy to that of the Nazis.
* Holding Jews collectively responsible for actions of the state of Israel.

1. **Structural anti-Semitism**

Antisemitismus WTF: does not openly agitate against "the Jews", but uses codes such as "high finance" or "the East Coast" and engages in an abbreviated, personalized and moralizing criticism of economic and power relations. In doing so, it uses the same prejudiced structures, forms of argumentation and old anti-Jewish images as classic anti-Semitism and refrains from taking a differentiated view of economic and political structures.

1. **Secondary Antisemitism**

Anders Denken: In secondary anti-Semitism, a reversal of the perpetrator-victim relationship often takes effect, either attributing Jewish people with complicity in rejection and persecution or ascribing a perpetrator role to them. This also becomes clear in the insinuation that Jews would deliberately use the memory of the Holocaust to gain financial advantages or to immunize themselves from any criticism. Motives of a secondary anti-Semitic perpetrator-victim reversal are often also found in combination with motives of Israel-related anti-Semitism, for example in the form of inadmissible Nazi comparisons, one-sided accusations of guilt or accusations of an unlawful instrumentalization of the Holocaust.

1. **Antigypsyism**

IHRA: Antigypsyism/anti-Roma discrimination is a manifestation of individual expressions and acts as well as institutional policies and practices of marginalization, exclusion, physical violence, devaluation of Roma cultures and lifestyles, and hate speech directed at Roma as well as other individuals and groups perceived, stigmatized, or persecuted during the Nazi era, and still today, as “Gypsies.” This leads to the treatment of Roma as an alleged alien group and associates them with a series of pejorative stereotypes and distorted images that represent a specific form of racism.

* Distorting or denying persecution of Roma or the genocide of the Roma.
* Glorifying the genocide of the Roma.
* Inciting, justifying, and perpetrating violence against Roma communities, their property, and individual Roma.
* Forced and coercive sterilizations as well as other physically and psychologically abusive treatment of Roma.
* Perpetuating and affirming discriminatory stereotypes of and against Roma.
* Blaming Roma, using hate speech, for real or perceived social, political, cultural, economic and public health problems.
* Stereotyping Roma as persons who engage in criminal behavior.
* Using the term "Gypsy" as a slur.
* Approving or encouraging exclusionary mechanisms directed against Roma on the basis of racially discriminatory assumptions, such as the exclusion from regular schools and institutional procedures or policies that lead to the segregation of Roma communities.
* Enacting policies without legal basis or establishing the conditions that allow for the arbitrary or discriminatory displacement of Roma communities and individuals.
* Holding Roma collectively responsible for the real or perceived actions of individual members of Roma communities.
* Spreading hate speech against Roma communities in whatever form, for example in media, including on the internet and on social networks.

1. **Sexism**

* Dorsch: “prejudiced attitudes and discriminatory behavior towards people based on their gender. In the most widespread version, sexism is discrimination against women.”.

1. IDEOLOGICAL AND INEQUALITY: The text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.
   * "I think the whole equality thing is getting out of hand. We are different, that’s how were made!".
2. STEREOTYPING AND DOMINANCE: The text expresses false ideas about women that suggest they are more suitable to fulfill certain roles (mother, wife, family caregiver, faithful, tender, loving, submissive, etc.), or inappropriate for certain tasks (driving, hardwork, etc), or claims that men are somehow superior to women.
   * "Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to.".
3. OBJECTIFICATION: The text presents women as objects apart from their dignity and personal aspects, or assumes or describes certain physical qualities that women must have in order to fulfill traditional gender roles (compliance with beauty standards, hypersexualization of female attributes, women's bodies at the disposal of men, etc.).
   * "Don't get married than blame all woman for your poor investment. You should of got a hooker but instead you choose to go get a wedding ring.".
4. SEXUAL VIOLENCE: Sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made.
   * "fuck that cunt, I would with my fist".
5. MISOGYNY AND NON-SEXUAL VIOLENCE: The text expressses hatred and violence towards women.
   * "Some woman are so toxic they don't even know they are draining everyone around them in poison. If you lack self-awareness you won't even notice how toxic you really are".
6. **Hostility towards the homeless**

* bpb: “The discrimination of homeless people. It is seen by social scientists as a typical element of an extreme right-wing world view. According to consistent reports from those affected and social organizations, homeless people are exposed to a wide range of discrimination. The spectrum ranges from stigmatization and discrimination in everyday life to threats and verbal abuse to physical attacks and murder. Hostility towards the homeless can be described as an expression of Social Darwinism.”

1. **Queer hostility**
2. **Generally**

Queer-Lexikon: Queerophobia is discrimination against queer people. It manifests itself, for example, through rejection, anger, intolerance, prejudice, discomfort or physical or psychological violence. However, queer hostility can also affect non-queer people who are perceived as queer or who (seem to) deviate from social rules and norms on sexuality and gender.

1. **Homophobia**

bpb: “The devaluation and harsh rejection of non-heterosexual people. The most frequent target of homophobia is male homosexuals, which is why the term is simplistically translated as homophobia. Homophobia manifests itself in numerous forms: in overt or covert rejection, in discriminatory acts up to and including violent attacks and murders or even in the assertion that homosexuality is a mental illness and can or must be treated. Academics classify homophobia (as well as racism, anti-Semitism and Islamophobia) as an extreme right-wing "ideology of inequality".”

1. **Trans hostility**

Queer Lexikon: “ Trans hostility refers to discrimination against trans people. It manifests itself, for example, through rejection, anger, intolerance, prejudice, discomfort or physical or psychological violence towards trans people. Trans hostility can also affect people who are not trans but are perceived as trans.

1. **Lookism**

Online Lexikon für Psychologie & Pädagogik: „Stereotypisierung bzw. Diskriminierung von Menschen allein auf Grund des Aussehens, kennzeichnet also eine Hierarchisierung von Menschen auf der Basis von Körpermerkmalen, die allgemein positiv oder negativ bewertet werden und somit den Wert des Individuums steigern oder mindern“

1. **Fatphobia**

World obesity org: “Weight stigma refers to the discriminatory acts and ideologies targeted towards individuals because of their weight and size. Weight stigma is a result of weight bias. Weight bias refers to the negative ideologies associated with obesity. These can include laziness, lack of will power, a lack of moral character, bad hygiene, low level of intelligence and unattractiveness. Stigmatizing beliefs and ideologies can lead to stigmatizing acts. These acts can manifest themselves in various different ways. People with obesity may experience negative verbal commentaries, teasing, or physical assault.”

1. **Ableism**

Not only anti-disabled, discriminatory attitudes and actions, but also an ideological discourse that fundamentally presupposes and demands non-disabled normality, autonomy and usefulness, and which is deeply embedded in social structures and in the subjectivity of all.

1. **Classism**

Gleichstellungsportal: Classism means the categorization and often associated with this the devaluation of people based on their social background (e.g. what educational qualifications and professions their parents have or in what environment they grew up), their economic position (e.g. how much they earn in their job or whether they have assets) and their milieu (e.g. what school qualifications they have or what interests they have). This means that classism is usually directed against people with low incomes, the unemployed and the homeless. Classism has an impact on life expectancy and limits access to housing, educational qualifications, healthcare, power, networks, participation, recognition and money.

1. **Hostility towards Refugees**

bpb: grundsätzliche Ablehnung von Geflüchteten als Gruppe. Flüchtlingsfeindlichkeit generell ist eine Summe an abwertenden Einstellungen, die auf der „Ideologie der Ungleichwertigkeit“ fußt. Menschenfeindliche Vorurteile gegen Geflüchtete markieren, wer „die Fremden“ sind. Menschen werden aufgrund ihrer zugeschriebenen sozialen Gruppe als „anders“, „fremd“ oder „unnormal“ stigmatisiert. Um Geflüchtete als Gruppe herabzuwürdigen, verbreiten Personen mit flüchtlingsfeindlicher Einstellung Unwahrheiten und leugnen Fakten. Auch gezielte Desinformations-Kampagnen spielen hier eine zentrale Rolle. So soll die Angst der Menschen vor einer vermeintlichen „Überfremdung“ durch den Zuzug von Geflüchteten geschürt werden.

What will be missing from this list is the already mentioned hate speech which targets individual based on the characteristics or behaviors associated (not with their (alleged) membership to a social group but) with the individual itself. In the study “Analyzing the Targets of Hate in Online Social Media” by Silva et al. (2021) this form of hate speech is categorized as “Behavior” and “Physical” and together make up 40% of encountered hate speech on Twitter, showing its significance for online hate speech which stands in stark contrast to who online hate crimes are generally documented. This underlines the need for an adequate definition of such kinds of hate speech which can be used in our prompt engineering. A possibility for filling this gap would be to include a definition of cyberbullying for the analysis of messages. The American Psychological Association (APA) defines cyberbullying the following way:

*“Cyberbullying occurs when someone uses technology to demean, inflict harm, or cause pain to another person. It is “willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices.” Perpetrators bully victims in any online setting, including social media, video or computer games, discussion boards, or text messaging on mobile devices.”*

**(3) Definition of Counter Speech (Mathew et al., 2019)**

A direct response/comment (not reply to a comment) that counters hateful or harmful speech.

**(4) Types of Counter Speech (Mathew et al., 2019)**

1. **Presenting facts to correct misstatements or mis-perceptions:** The counterspeaker tries to persuade by correcting misstatements.

**Example:** *“Actually homosexuality is natural. Nearly all known species of animal have their gay communities. Whether it be a lion or a whale, they have or had (if they are endangered) a gay community. Also marriage is an unnatural act. Although there are some species that do have longer relationships with a partner most known do not”*

1. **Pointing out hypocrisy or contradictions:** The counterspeaker points out the hypocrisy or contradiction in the user’s (hate) statements. In order to discredit the accusation, the individual may explain and rationalize their previous behavior, or if they are persuadable, resolve to avoid the dissonant behavior in the future.

**Example:** *“The ‘US Pastor’ can’t accept gays because the Bible says not to be gay. But...he ignores: The thing about eating shrimp or pork, The thing about touching the skin of a dead pig (Football), The thing about mixing fabrics, The thing about torn clothes, The thing about tattoos, The thing about planting two crops in one field, The thing about working on the Holy Day (Saturday or Sunday depending)...for any and all of those sins one should burn for an eternity, yet is ignored. But when it comes to loving the wrong person (gays) this will not do! Christians only follow the parts of the bible that supports their bigotry. YOUR A HYPOCRITE.”*

1. **Warning of offline or online consequences**: The counterspeaker warns the user of possible consequence sof their actions. This can sometimes cause the them to retract from their original opinion.

**Example**: “I’m not gay but nevertheless, whether You are beating up someone gay or straight, it is still an assault and by all means, this preacher should be arrested for sexual harassment and instigating!!!”

1. **Affiliation:** Affiliation is “...establishing, maintaining, or restoring a positive affective relationship with another person or group of persons” (Byrne 1961). People are more likely to credit the counterspeech of those with whom they affiliate, since they tend to “evaluate ingroup members asmore trustworthy, honest, loyal, cooperative, and valuableto the group than outgroup members” (Kane, Argote, andLevine 2005). In the dataset used in this study, counterspeakers who use Affiliation received the highest number of likes for their comments among all of the counterspeech types.

**Example:** *“Hey I’m Christian and I’m gay and this guy is so wrong. Stop the justification and start the accepting. I know who my heart and soul belong to and that’s with God: creator of heaven and earth. We all live in his plane of consciousness so it’s time we started accepting one another. That’s all”*

1. **Denouncing hateful or dangerous speech:** The counterspeakers denounce the message as being hateful. This strategy can help the counterspeakers in reducing the impact of the hate message.

**Example:** “Please take this down YouTube. This is hatespeech.”

1. **Humor and sarcasm:** Humor is one of the most powerful tools used by the counterspeakers to combat hate speech. It can deescalate conflicts and be used to garner more attention toward the topic. Humor in online settings also eases hostility, offers support to other online speakers, and encourages social cohesion (Marone 2015). Often, the humor is sarcastic.

**Example:** “HAHA-HAHAHAHAHAH...oh you were serious. That’s even funnier :)”.

1. **Positive tone:** The counterspeaker uses a wide variety of tones to respond to hate speech. In this strategy, we consider different forms of speech such as empathic, kind, polite, or civil. Increasing empathy with members of opposite groups counteracts incitement

**Example**: “I am a Christian, and I believe we’re to love everyone!! No matter age, race, religion, sex, size, disorder...whatever!! I LOVE PEOPLE!! We are not going to go anywhere as a country if we don’t put God first in our lives, and treat EVERYONE with respect”.

1. **Hostile language:** The counterspeaker uses abusive, hostile, or obscene comments in response to the original hate message. Such a response can persuade an original speaker to delete his message or even a whole account, but is unlikely to either de-escalate the conversation or persuade the original speaker to recant or apologize.

**Example:** “This is ridiculous!!!!!! I hate racist people!!!! Those police are a\*\*holes!!!”.

**(5) Effectiveness of Counter Speech**

**Empathy-based counterspeech**: Can increase the retrospective deletion of xenophobic hate speech (by 0.2 SD) and reduce the prospective creation of xenophobic hate speech over a 4-week follow-up period by 0.1 SD. However, no consistent effects for strategies using humor or warning of consequences were found (Hangarnter et al., 2021)

**Example:** “Language like this is just unnecessarily hurtful towards immigrants”

**Community acceptance** (as observed by % agreement) of the type of counter-speech varies (Mathew et al., 2019).

A screenshot of a report

Description automatically generated

In the end, the effectiveness of **counterspeech** itself depends on the **goal** that is set: (1) Making the target(s) of hate speech feel better for which it makes sense to offer different kinds of counter-speech possibilities, not only in wording but in type or (2) Decreasing the likelihood of hate speech in the future for which specifically empathy-based counterspeech might be effective.

**(6) Hate Speech Detection through LLMs**

AI models inherently contain biases, since they are based on the data, they are fed. Abid et al. (2021) for example show for GPT-3 a clear anti-Muslim bias, severe even compared to biases about other religious groups. The same problem also extends to the task of detecting hate speech. When training classifiers on five datasets containing tweets written either in African American English or in Standard American English, the model predicts tweets written in African American English to be abusive at substantially higher rates (Davidson et al., 2019). Since we won’t have direct access to any reweighing or fine-tuning process for the LLM we’re using, we should provide it with as much information (also about possible biases) as possible to counteract any inherent biases, especially through added context. This is made apparent in Yu et al. (2022) analyses, showing that human judgments change for most comments depending on whether annotators are shown the context. They identify that adding *Parent* comments as context can fix errors caused by “Lack of information”, “Negation”, “Sarcasm and Irony” and hate speech not containing any swear words. An additional and interesting approach to added context called Rule-By-Example (RBE) is presented by Clarke et al. (2023). On a very basic level, they created rules out of substrings extracted from datasets of hateful words or hates speech. Then, they pair each rule with an example. This architecture not only seems to outperform state-of-the-art deep learning classifiers, but it also provides explainability, as each decision can be traced back to rules. As we don’t have to train our own model, the exact architecture of RBE is not as important. However, providing the LLM with prompts that contain annotated examples (parts of speech that make it hateful) would provide it with something similar to these rules on which it could base its decisions during hate speech detection. The same approach could be used in providing the model with annotated counter speech examples for better generation.

**(7) Counterspeech Generation**

First of all, it is shown that people tend to prefer counter speech that shows a good understanding of hate speech (Zheng et al., 2023). This once again underlines the importance of well though through and contextualized hate speech detection, which is capable to also answer why a comment was flagged as hateful. Mun et al. (2024) identify multiple barriers for users to engage in writing counter speech: Limited personal resources, lack of training, unclear impact, and personal harms. These can all be addressed by an automated counter speech generation through an LLM which would reduce effort, provide guidance and possibly also increase the impact and redirect personal harms of writing counter speech. However, counter speakers voiced strong concerns about credibility (“If people figure out it’s a bot then it loses all credibility“), agency (“I like using my words not being tied to what AI says“), authenticity (“If you are trying to be genuine, then you have to, you know, be a genuine human.”) and the long-term functionality of using AI tools for counterspeech. While the functionality is of course yet to be tested, the first three concerns hint towards the importance of involving the user actively in the counter speech. This should at least be reflected in a feature of our extension which gives users the possibility to edit the generated counter speech suggestions.

models that use rationale and named-entity features as part of the training input outperform

our baseline model.

**(8) Pipeline for Counterspeech Generation against Online Hate Speech**

A person with a counter speaker

Description automatically generated with medium confidence

**(9) Hatespeech Examples**

**DATASETS:**<https://github.com/marcoguerini/CONAN?tab=readme-ov-file#conan>

<https://www.dropbox.com/sh/g9uglvl3cd61k69/AACEk2O2BEKwRTcGthgROOcWa?dl=0>  
<https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data>

<http://nlp.uned.es/exist2023/>

**OTHER DATA:**https://zentralrat.sintiundroma.de/wp-content/uploads/2018/12/report\_antiziganismus\_online.pdf

https://www.jugendschutz.net/fileadmin/daten/publikationen/praxisinfos\_reports/report\_queerfeindlichkeit\_online.pdf

https://www.jugendschutz.net/fileadmin/daten/publikationen/praxisinfos\_reports/praxisinfo\_antimuslimischer\_rassismus\_online.pdf

1. **Racism**
2. **Generally**

* **HATE:** “Some races have lower physical and cognitive abilities, the sooner we accept this, the better it will be for humanity.”  
  **COUNTER:** “We all belong to one - scientifically determined - species: the Homo sapiens. The sooner you accept this, the better it is.”
* **HATE:** “Many philosophers argue that humans are not born equal, especially when it comes to race.”  
  **COUNTER:** “This is a so-called 'argument from authority' and thus a logical fallacy. I would like to see some real arguments to support this claim.”

1. **Anti-muslim racism**

* **HATE:** “Cultures have never been fair and this is a fact. Muslims for example are just a bunch of misogynist monkeys!”   
  **COUNTER:** “Again, we can't point the finger at a religion/culture for some old commandment. Even Christianity has rules that may seem crazy, but rightly people do not hate Christians for this.. a small example: Ephesians 5:22, "Wives, submit to you husbands as the Lord"
* **HATE: “**Every Muslim is a potential terrorist.”  
  **COUNTER:** “I think this is a very inconsiderate thing to say, would you say the same thing about non-Muslims as well??”

1. **Anti-black racism**

* **HATE:** “I went to a shit load of therapy. I still don't feel safe in my own house with all these black people around.”

**COUNTER:** Maybe the black community also feels unsafe in their own houses with so many people glaring at them...

* **HATE:** “I'm proud to be white, blacks are a disgrace to humanity.”  
  **COUNTER:** “You are not ashamed of your heritage, and that is good. But you should be ashamed of your discriminatory behaviour that is filling our society.”

1. **Anti-asian racism**

* **HATE:** “Asians are basically just a bunch of slobs and I don't like how they are accepted here.”  
  **COUNTER:** “There is a wide range in income and education, as well as religious and ethnic background, amongst people from Asia. What exactly is the reason why you don't like them all?”
* “I'm losing my mind 😵 quarantine is making me mad. #Chinese douchebags ate shit infected everyone with this #ChineseVirus and now the whole world is suffering for them. Bloody shitbags! Bloody ban #china ban Chinese people!😤 #Coronavirus”
* “We were told by @Stanford Doctor buddy ours January 24 when asked about Wuhan Virus... "Prepare for the worst" ...and we closed our ad agency on Monday - February 17 and sent everyone home. China covered this whole Damn thing up and the Fascists in Beijing are Pure Damn Evil! https://t.co/XRpVYXrBDQ

1. **Antisemitism**

* **HATE:** Freedom of speech is under attack. It's being taken away by Zionist Occupation Government.  
  **COUNTER: “**Freedom of speech means that you can disagree with ideas. Freedom of speech means that you cannot be censored. Freedom of speech means that you can have your ideas promoted. In the end you can have a lot of different beliefs, but there is no such thing as a Zionist Occupation Government, otherwise you would not be allowed to freely express these antisemitic ideas.”
* **HATE:** “The Holocaust is a business. It has been and is still used to extort millions from European countries in reparations and to create a wall of impunity behind which Jews can operate with impunity, free from criticism.”  
  **COUNTER:** “Considering all the criticisms of the state of Israel (however right or wrong, does not matter) I don't see how you can say that Jews are operating "with impunity, free from criticism."
* **HATE:** “Go and read the Protocols of the elders of Zion if you haven’t. Everything is orchestrated..”  
  **COUNTER:** “You are right, everything is orchestrated: the protocols were concocted by the czarist police in the early part of the 20th century to counter a hankering for constitutional government among Russian people. They needed a scapegoat to divert people's attention. Sadly the protocols worked so well that you are still citing them here today.”

1. **Antigypsyism**

* **HATE:**„Nach Auschwitz hätte man sie abschieben sollen. DIE ZIGEUNER SIND EUROPAS UNGLÜCK!“
* **HATE:** „Selbst die Zigeuner in meinem Block grüßen mich weil sie wissen wie asozial ich bin
* **HATE:** „Es ist wissenschaftlich bewiesen das die meisten Zigeuner Unmenschen sind und sie müssen genauso behandelt werden wie sie unser Land und Gesetze“
* **HATE:** „Gypsies are literally destroying Europe. They are literally invading and destroying our beautiful continent. They are a cancer on our race and they must be stopped.“  
  **COUNTER:** “Your statement is FALSE. They are guiltless victims of the 'influx of refugees' paranoia. They are European and have been persecuted for centuries. They have always fought for their rights, and for the protection of their culture and way of life, as any european would do. In fact there are many good Gypsy activists who are working to improve their lives. And what we are doing to these people is a malevolent smear campaign, with no common motivation. They are part of European history.”

1. **Sexism**
2. IDEOLOGICAL AND INEQUALITY:

* **HATE:** "I think the whole equality thing is getting out of hand. We are different, that’s how were made!".

1. STEREOTYPING AND DOMINANCE:

* **HATE:** "Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to.".

1. OBJECTIFICATION:

* **HATE:** "Don't get married than blame all woman for your poor investment. You should of got a hooker but instead you choose to go get a wedding ring.".

1. SEXUAL VIOLENCE:

* **HATE:** "fuck that cunt, I would with my fist".

1. MISOGYNY AND NON-SEXUAL VIOLENCE:

* **HATE:** "Some woman are so toxic they don't even know they are draining everyone around them in poison. If you lack self-awareness you won't even notice how toxic you really are".

1. **Hostility towards the homeless**

/ Prevalent mainly in real-life actions

1. **Queer hostility**
2. **Generally**

* **HATE:** “LGBTs just want to brainwash our children. They should not be taught about gender identity and sexual orientation in school. This is a too polarized debate.”  
  **COUNTER:** “Many points in history are also subjects of polarized debate, but they still can be, and are, taught objectively.”

1. **Homophobia**

* **HATE:** “We should isolate gays because they are spreading AIDS all around the world.”  
  **COUNTER:** “If you want to isolate someone then there is only one "high-risk group" to focus on. It is made of all those who have frequent unprotected sex with different partners. Being gay does not make you a part of this by default.”
* **HATE:** “Gay sex is a private kink that gays can perform in their gay houses. Instead, love is possible only between a man and a woman.”  
  **COUNTER:** “There are men in love with men too. Some people are heterosexual, and other homosexual. What you are saying is that loving men is not the same as loving women. Still, both are forms of love.”

1. **Trans hostility**

* **HATE:** “Wir leben in einer Welt, in der Kinder sich nicht als Indianer verkleiden dürfen, aber Männer als Frau“; „Dachau…“
* **HATE:** „I tolerated fag pride. I tolerated fags on tv and in every town square. I draw the line at tranny demons reading books to children.

**COUNTER:** Not sure how reading a book is a problem, sounds pretty nice for the kids.

* **HATE:** “There should be restriction on trans actors in action movies. This cannot continue.”  
  **COUNTER:** “There are very few transgender actors in films and almost no evidence of transgender characters according to GLAAD SRI index (at least for 2019). Gender discrimination in the mainstream movie industry is endemic and this is what cannot continue.”

1. **Lookism**

/Cyberbullying

1. **Fatphobia**

* **HATE:** Fat people should be isolated from society, so we will eventually have a healthier world.  
  **COUNTER:** This makes fat shaming directly analogous to other forms of discrimination, like racism and sexism.
* **HATE:** “The real problem is that fat people do not want to change.” **COUNTER:** “The narrative of 'you can choose not to be obese' is much, much more hurtful to the obese than it is helpful.”

1. **Ableism**

* **HATE:** “Downy children should be isolated, they cannot cope with education in normal schools and are detrimental to their peer's education.”  
  **COUNTER:** “Coeducation benefits neurotypical children as much as Down Syndrome children, as they are surrounded by positive examples of care and attention from authority figures in the knowledge that they, too, will be looked after if in need.”
* **HATE:** “People with disabilities are just privileged spoiled children.”

**COUNTER:** “It is important to remember that disabled people often face discrimination. This could have serious long-term consequences, for example making it harder for people to have a serene life. This is why it's important to fight for their rights.”

1. **Classism**

/ Cyberbullying

1. **Hostility towards Refugees**

* **HATE:** “Foreigners on UK benefits leaps 41% in 5 years with over 400K now handed payouts. This nonsense has got to stop.”

**COUNTER:** “Maybe the reason for such a rise in UK benefit claimants is due to unequal opportunities and lack of job opportunities, and this goes for everyone, not just 'foreigners'.”

* **HATE:** “Following the structure of our population we have to stop and reverse all migration. We have got all the doctors and rocket scientists we need. We will never need any workforce from outside the country.”
* **COUNTER:** “As a matter of fact, we couldn't manage without immigrants. Most of our NHS staff are immigrants. The health system would collapse without migrant workers.”

1. **Cyberbullying**

* **HATE:** “You sad fuck. I'm going to restore my edit, and you are going to leave it, or else I will knee you in the groin yourself, prick.” **COUNTER:**
* **HATE:** “WOULDN'T BE THE FIRST TIME BITCH. FUCK YOU I'LL FIND OUT WHERE YOU LIVE, SODOMIZE YOUR WIFE AND THEN BURN YOUR HOUSE DOWN. FUCK YOU YOU FUCKING QUEER.” **COUNTER:**

**Resources:**<https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech>

<https://rowna.pwr.edu.pl/en/discrimination/discriminationtypes/hate-speech>

<https://www.britannica.com/topic/hate-speech>

<https://www.amadeu-antonio-stiftung.de/gruppenbezogene-menschenfeindlichkeit/>

<https://www.bu.edu/antiracism-center/files/2022/06/Anti-Black.pdf>

<https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-elimination-all-forms-racial>

<https://www.bpb.de/themen/infodienst/302514/was-ist-antimuslimischer-rassismus/>

<https://mediendienst-integration.de/fileadmin/Dateien/Factsheet_Anti_Asiatischer_Rassismus_Final.pdf>

<https://antisemitismus.wtf/hintergrund/struktureller-antisemitismus/>

<https://holocaustremembrance.com/resources/working-definition-antisemitism>

<https://holocaustremembrance.com/resources/working-definition-antigypsyism-anti-roma-discrimination>

<https://www.anders-denken.info/informieren/sekund%C3%A4rer-antisemitismus-1>

<https://lexikon.stangl.eu/22396/lookism>

<https://dorsch.hogrefe.com/stichwort/sexismus>

<https://queer-lexikon.net/2017/06/08/transfeindlichkeit/>

<https://queer-lexikon.net/2020/04/29/queerfeindlichkeit/>

<https://www.bpb.de/themen/rechtsextremismus/dossier-rechtsextremismus/500785/homophobie/>

<https://www.bpb.de/themen/rechtsextremismus/dossier-rechtsextremismus/500802/obdachlosenfeindlichlkeit/>

<https://www.worldobesity.org/what-we-do/our-policy-priorities/weight-stigma>

<https://www.gleichstellungsportal.de/abc-der-gleichstellung/klassismus/>

<https://www.bpb.de/themen/rechtsextremismus/dossier-rechtsextremismus/544438/gewalt-gegen-gefluechtete/#node-content-title-0>

<https://www.apa.org/topics/bullying/cyberbullying-online-social-media>

<https://punyajoy.github.io/talk/crrash_talk/talk.pdf>

<https://zentralrat.sintiundroma.de/wp-content/uploads/2018/12/report_antiziganismus_online.pdf>

<https://www.jugendschutz.net/fileadmin/daten/publikationen/praxisinfos_reports/report_queerfeindlichkeit_online.pdf>

<https://www.jugendschutz.net/fileadmin/daten/publikationen/praxisinfos_reports/praxisinfo_antimuslimischer_rassismus_online.pdf>

<https://www.dropbox.com/sh/g9uglvl3cd61k69/AACEk2O2BEKwRTcGthgROOcWa?dl=0>

<https://github.com/marcoguerini/CONAN?tab=readme-ov-file#conan>

<https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data?select=test.csv.zip>

<http://nlp.uned.es/exist2023/>

Abid, A., Farooqi, M. & Zou, J. (2021). Persistent Anti-Muslim bias in large language models. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2101.05783>

Clarke, C., Hall, M., Mittal, G., Yu, Y., Sajeev, S., Mars, J. & Chen, M. (2023). Rule By Example: Harnessing Logical Rules for Explainable Hate Speech Detection. *Proceedings Of The 61st Annual Meeting Of The Association For Computational Linguistics*. https://doi.org/10.18653/v1/2023.acl-long.22

Davidson, T., Bhattacharya, D. & Weber, I. (2019). Racial bias in Hate Speech and Abusive Language Detection datasets. *Proceedings Of The Third Workshop On Abusive Language Online*. https://doi.org/10.18653/v1/w19-3504

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J. F., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Muñoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F. & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. Proceedings Of The National Academy Of Sciences Of The United States Of America, 118(50). https://doi.org/10.1073/pnas.2116310118

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P. & Mukherjee, A. (2019). Thou shalt not hate: Countering Online Hate speech. *Proceedings Of The International AAAI Conference On Web And Social Media*, *13*, 369–380. <https://doi.org/10.1609/icwsm.v13i01.3237>

Mun, J., Buerger, C., Liang, J. T., Garland, J. & Sap, M. (2024). Counterspeakers’ Perspectives: Unveiling Barriers and AI Needs in the  Fight against Online Hate. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.00179

Silva, L. A., Mondal, M., Correa, D., Benevenuto, F. & Weber, I. (2021). Analyzing the Targets of Hate in Online Social Media. *Proceedings Of The International AAAI Conference On Web And Social Media*, *10*(1), 687–690. <https://doi.org/10.1609/icwsm.v10i1.14811>

Yu, X., Blanco, E. & Hong, L. (2022). Hate Speech and Counter Speech Detection: Conversational Context Does  Matter. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2206.06423

Zheng, Y., Ross, B. & Magdy, W. (2023). What makes good counterspeech? A comparison of generation approaches and evaluation metrics. *Proceedings Of The 1st Workshop On CounterSpeech For Online Abuse (CS4OA)*.