# HATE SPEECH

## What is Hate Speech?

"Public communication that expresses hatred or advocates violence against a person or group based on something such as race, religion, sex, or sexual orientation" is described as hate speech. "Usually understood to comprise expressions of enmity or disparagement of a person or a group on account of a group feature such as race, color, national origin, sex, handicap, religion, or sexual orientation," according to the definition of hate speech.

## Modules:

1) https://dl.acm.org/doi/abs/10.1145/3292522.3326032
2) https://prohic.nl/wp-content/uploads/2021/05/213-17mei2021-InternetOnlineHateSpeechtSystematicReview.pdf
3) https://www.frontiersin.org/articles/10.3389/fdata.2020.00003/full

## Summary of the Modules:

### Article 1 (Prevalence and Psychological Effects of Hateful Speech in Online College Communities):

Hate speech in online collegiate forums is discussed in this article. Hate speech has severe consequences and is especially harmful in college settings. Hate speech on college campuses is a challenging socio-political problem, and interventions to reduce the consequences must assess the phenomenon's pervasiveness on campuses as well as the repercussions on students' psychological well-being.

The goal is to research the online element of hate speech in a dataset of 6 million Reddit comments posted in 174 campus groups, given the increased usage of social media among college students. The Collegiate Hate Index (CHX) was created to assess the prevalence of hostile discourse in an online college community. Then it's looked at in terms of hateful speech, conduct, class, disability, ethnicity, gender, physical appearance, race, religion, and sexual orientation to see how it's distributed. The psychological impacts of hostile speech are then studied using a causal-inference paradigm, namely in the form of individuals' online stress expression. Finally, psychological endurance to hateful speech is determined by examining their language, discriminating keyword usage, and personality features.

According to the findings, hateful speech is common among college subreddits, with 25% of them having more hostile speech than non-college subreddits. In addition, being exposed to hate causes more tension to be expressed. However, not everyone who is exposed is affected

in the same way; some people have less psychological endurance than others. Individuals with low endurance are more prone to emotional outbursts and neurotic than those with better endurance.

**Article 2 (Internet, social media and online hate speech):**

The goal of this literature review was to look at research publications on how the Internet and social media may or may not provide opportunities for online hate speech. Out of 2389 publications retrieved in the searches, 67 research were deemed to be appropriate for analysis. Between 2015 and 2019, there were articles about online hate speech or cyberhate. Because of the wide range of studies and measurement units, a meta-analysis was not possible. Exploratory data regarding the Internet and social media as spaces for online hate speech, categories of cyberhate, terrorism as an online hate trigger, online hate expressions, and the most prevalent methodologies to measure online hate speech were supplied by the analyzed research. As a general definition of cyberhate, it is defined as the use of violent, aggressive, or offensive language directed at a specific group of people who share a common property, such as religion, race, gender, or sex, or political affiliation, through the use of the Internet and Social Networks, based on a power imbalance, and which can be carried out repeatedly, systematically, and uncontrollably, often motivated by ideologies.

The article also discusses the many sorts of hate speech. It comprises the following:

1) **Online hate speech**
    a. **Online religious hate speech:** This sort of hate speech is described as the use of provocative and sectarian rhetoric in cyberspace to advocate hatred and violence against people based on religious affiliation.
    b. **Online racism:** Racism appears to be magnified in online communities. The anonymity and improved accessibility of the Internet provided a forum for utterances of online racist sentiments, as well as identity protection. The examination of 51,991 public comments on 119 news items on the Canadian Broadcasting Corporation News Facebook page concerning race, racism, or ethnicity revealed the spread of hate against indigenous and black people, maintaining dominating discourses on white identities.
    c. **Political online hate:** Through social media, many democratic and political institutions may intensify hate and prejudice toward others. Rather of generating significant democratic discourse about policy options and rationales, referendums, for example, are typically focused on influencing other arguments around a subset of frequently unrelated problems to what is being considered for a popular vote.
    d. **Gendered online hate:** While the majority of online hate speech targets people based on their race or nationality, hate speech based on gender and sexual

orientation is on the rise, since digital media may worsen current patterns of gendered violence and bring new forms of abuse.

2) **Terrorism as an online hate trigger:** Terrorism occurrences are commonly linked to observable public social media, according to the second group. Following the March 2016 terrorist bombings in Brussels, the #StopIslam hashtag became popular on Twitter, where it was used to promote racist hate speech and misinformation directed towards Islam and Muslims. Following the June 2017 London Bridge terrorist incident, 200,880 anti-Muslim hate tweets were detected in the United Kingdom, and in the aftermath of the 2015 terrorist attacks in France.
3) **Online hate expressions:** This third category depicts how social media platforms are used to spread hate online. Vicarious observation, racist comedy, negative racial stereotyping, racist internet media, and racist online hate organizations were all discovered to be used in the instance of racism. Shaming is a common tactic used by those who detest women online.

**Article 3 (The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring):**

This article describes and focuses on an action research project including multi-organizational collaboration that was carried out during the 2017 Finnish municipal elections, in which a digital infrastructure was created to automatically monitor candidates' social media updates for hate speech. The setting allowed for a two-pronged investigation.

First, the partnership provided a new perspective on how hate speech manifests as a technological issue. Using supervised machine learning, the research generated a sufficiently well-working algorithmic solution. It ended up employing a mix of Bag-of-Words feature extraction and Support-Vector Machines after assessing the performance of different feature extraction and machine learning approaches. However, an automated method necessitated significant reduction, such as the use of simple scales for categorizing hate speech and a dependence on word-based techniques, although hate speech is a linguistic and social phenomenon with many tones and forms.

Second, the action-research-oriented environment allowed for the observation of emotive responses to machine learning technologies, such as hopes, dreams, and anxieties. We identified participants' aspirations for effective automation as well as the level of neutrality and objectivity introduced by an algorithmic system based on participatory observations, project artifacts and documents, interviews with project participants, and online reactions to the detection project.

After the monitoring procedure, however, the participants were more critical of the system. The findings show how enormous technological expectations may readily dominate a project dealing with a contentious, current societal problem.

## Understanding hate speech with the help of a dataset:

Let's look at the ETHOS hate speech dataset to learn more about it. Having a data visual aids our understanding of the subject.

ETHOS is a shortened form for multi-labEl haTe speecH detectiOn dataSet (ETHOS). The dataset has been generated from various social media platforms. ETHOS is a Greek word meaning “character” that is used to describe the guiding beliefs or ideals that characterize a community, nation, or ideology. The Greeks also used this word to refer to the power of music to influence emotions, behaviors, and even morals.

The dataset contains a total of 433 items. The following is a representation of the instances:

| | comment | violence | directed_vs_generalized | gender | race | national_origin | disability | religion | sexual_orientation |
|---|---|---|---|---|---|---|---|---|---|
| 0 | You should know women's sports are a joke | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.000000 |
| 1 | You look like Sloth with deeper Down's syndrome | 0.000000 | 0.833333 | 0.000000 | 0.000000 | 0.000000 | 1.0 | 0.0 | 0.000000 |
| 2 | You look like Russian and speak like Indian. B... | 1.000000 | 0.500000 | 0.142857 | 0.142857 | 0.714286 | 0.0 | 0.0 | 0.000000 |
| 3 | Women deserve to be abused, I guess. | 0.666667 | 0.166667 | 1.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.000000 |
| 4 | Women are made for making babies and cooking d... | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.000000 |
| 5 | why haven't you faggots taken the bi-pill yet | 0.500000 | 0.500000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 1.000000 |
| 6 | Why don't you go beck your god to murder you? ... | 0.833333 | 0.500000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 1.0 | 0.000000 |
| 7 | why does that african american monkey hold gun... | 0.000000 | 0.500000 | 0.000000 | 1.000000 | 0.000000 | 0.0 | 0.0 | 0.000000 |
| 8 | Why does Bulgaria exist as a country? It's one... | 0.000000 | 0.500000 | 0.000000 | 0.000000 | 0.428571 | 0.0 | 0.0 | 0.571429 |
| 9 | Why do negroes tear down their own neighborhoo... | 0.250000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 0.0 | 0.0 | 0.000000 |

There are several fields to comprehend the data. Each field is described in detail below:

1) **Comment:** This section comprises comments gathered from various social media networks.
2) **Violence:** This section indicates whether or not the statements promote violence. If it incites aggression, the value is 1, otherwise it is 0.
3) **Directed vs Generalized:** This section indicates whether the statement is addressed towards a specific person or a group of people. The value for an individual is 1, whereas the value for a group is 0.
4) **Gender:** This field contains the hate speech's Gender category. It can carry either a 0 or a 1 value. The number is 1 if the comment is about a certain gender, else it is 0.
5) **Race:** This field indicates the hate speech's race categorization. It can only carry 0 or 1 values. If the comment is about a certain race, the value is 1, otherwise it is 0.

6) **National Origin:** This element specifies the hate speech's country origin category. It can only carry 0 or 1 values. If the comment is about a country, the value is 1, otherwise it is 0.
7) **Disability:** This field specifies the hate speech's handicap category. It can only carry 0 or 1 values. If the comment is about a specific handicap (for example, any syndrome), the value is 1, otherwise it is 0.
8) **Religion:** This parameter specifies the hate speech's religion categorization. It can only carry 0 or 1 values. If the comment is about a certain religion, the value is 1, otherwise it is 0.
9) **Sexual Orientation:** This field specifies the hate speech's sexual orientation categorization. It can only carry 0 or 1 values. If the comment is about sexual orientation, the value is 1, otherwise it is 0.