Hate Speech on Social Media: Trends, Impacts, and Responses

by Nachiket Patil

December 16, 2025

Hate Speech:

Definition: Hate speech is any communication that attacks or uses derogatory language about a person or group based on a protected identity (religion, ethnicity, nationality, race, gender, sexual orientation, disability, etc.)[1]. It can be expressed in **words, images, videos, symbols, or gestures**, and spread both online and offline[2]. Hate speech is by nature *discriminatory* (bigoted or intolerant) or *pejorative* (prejudiced, demeaning), aiming to dehumanize or demean its targets[1][2]. Victims of hate speech often feel **traumatized, excluded, unsafe, angry or sad** because of these attacks[3], and communities can be unsettled by the spread of hateful content. Notably, experts warn that online hate speech can fuel real-world violence; historical cases (from lynchings to mass shootings) have been linked to spikes in hateful online rhetoric[4][5].

Key Characteristics of Hate Speech

- **Any Medium:** Hate speech can take many forms – slurs or insults in text, hateful memes or cartoons, malicious emojis, harassing videos, or even hostile gestures/symbols[2]. For example, sharing a meme mocking a racial group or using a pejorative term in a tweet constitutes hate speech because it disparages a group's identity[1][2].
- **Identity-Based:** It specifically targets people based on *identity factors* like race, religion, gender, sexual orientation, disability, nationality, etc. Hate speech "calls out" these factors to demean the person or group[2][1]. It does **not** include criticism of governments, institutions, or beliefs (which may be allowed speech) but focuses on individuals/groups.
- **Intent and Effect:** By definition, hate speech is harmful – it stigmatizes and can **legitimize discrimination** or even incite violence. As the UN Special Rapporteur notes, hate speech "dehumanizes the people who are targeted" and normalizes prejudice, often making extreme violence against them seem acceptable[5][4]. This is why many laws and platforms explicitly ban it.

Prevalence and Global Trends

Hate speech is **widespread** on social media. Surveys and studies worldwide report very high exposure: - **Global exposure:** In a 2023 UNESCO-Ipsos survey across 16 countries, **67% of internet users said they had encountered hate speech online** (74% of users under 35)[6]. In the European Union, around **80% of people have seen hate speech on social networks**, and 40% felt personally attacked or threatened on these sites[7]. These figures show most social media users come across hate content regularly. - **Youth impact:** Young people are especially exposed. In the United States, the Anti-Defamation League found that 52% of Americans report experiencing harassment on social media. In a large UCLA study of children (ages 10–18), **80% of youths reported encountering hate speech on social media in the prior month**[8][9]. These children commonly saw hate related to race/ethnicity (71%), gender (72%) and religion (62%)[9]. - **Platform hotspots:** People perceive certain platforms as more prone to hate speech. In the UNESCO survey, 58% believed **Facebook** was the biggest source, vs. 30% for TikTok and 18% for Twitter/X[10]. (This reflects user perception; actual rates depend on moderation effectiveness.)

Overall, research indicates hate speech on social media has been **increasing**. For example, machine-learning analyses show a steady rise: one study found hate speech nearly **doubled** on Twitter from 2015 to 2020[11]. Another 2025 study reported a ~50% spike in hateful posts on Twitter/X after late-2022 changes in management[12]. These trends suggest that without strong checks, hateful content can grow with platform usage.

Vulnerable Groups

Hate speech disproportionately targets **historically marginalized communities**. Surveys highlight the most affected groups: - **Ethnic and racial minorities:** Roughly 70% of hate speech victims belong to minority groups[5]. Immigrants, refugees, and racial or ethnic minorities frequently face online hatred (e.g. anti-Black, anti-Asian, Islamophobic, antisemitic slurs).
- **Religious minorities:** Groups like Jews, Muslims, and other religious minorities are regularly singled out in online hate campaigns[5]. During conflict spikes, hate against religious groups often surges.
- **LGBTQ+ communities:** Survey data shows **LGBT+ people are often top targets of online hate** (33% of respondents identified them as main victims)[13]. Transphobic and homophobic speech is a common form of cyberhate.
- **Women, immigrants, disabled:** Women (especially those speaking out on social media) often endure gendered harassment. Immigrants and refugees are blamed or vilified in hateful posts. Persons with disabilities and other characteristics (age, language, social class, etc.) can also be attacked, though less frequently[2].
- **Intersectional targeting:** People who fall into multiple protected categories (e.g. a Muslim woman of color, LGBTQ+ person of a minority ethnicity) often face compounding hate.

These groups share a common thread: hate speech reinforces existing **power imbalances**. By targeting vulnerabilities like race or gender, perpetrators exploit societal prejudice. This is why experts stress that combating hate speech is also about protecting human rights and dignity.

Examples of Hate Speech

Hate speech on social media can be blatant or coded, but some clear examples include:
- Posting a derogatory slur to insult a protected group. For instance, a tweet using a known racial slur against a minority or a derogatory term for a religious group is hate speech because it disparages them for who they are[1].
- Sharing memes or images that stereotype a group negatively. A meme that, for example, mocks immigrants with hateful captions targets ethnicity/national origin. Images with swastikas or other hate symbols that glorify violence are also hate speech[2].
- Text or video that calls for harm or exclusion of a group. Posts advocating "throwing out" a community, denying their rights, or inciting violence (e.g. "kill [group]") clearly cross into hate speech. Even seemingly "jokey" content can be hate speech if it ridicules or degrades a protected characteristic.

In practice, platforms define hate speech broadly: **any** content that "[attacks or uses] pejorative or discriminatory language" about protected groups is banned[1]. This includes code words or slang commonly understood as slurs, not just overt insults.

Impacts on Individuals and Society

Hate speech does serious harm. Victims often experience **emotional distress, anxiety, and depression**[3]. Being singled out in a public forum can make someone feel unsafe and unwelcome in their online (and offline) communities. This constant exposure can reduce self-esteem and sense of belonging, especially for young people.

Beyond individual harm, hate speech threatens social cohesion. It **normalizes prejudice** and can encourage others to adopt biased attitudes. Studies have linked surges in online hate to increases in real-world hate crimes and violence against minorities[4]. For example, inflammatory memes and posts have fueled mob attacks in some regions, as social media provides echo chambers that amplify extremist views. In short, unchecked hate speech not only injures targets but can **erode trust and peace** in society.

Trends and Projections

- **Rising volume:** Analyses suggest hate speech content is growing as social media usage increases. One machine-learning study found that the percentage of hate tweets roughly doubled from 2015 (0.53%) to 2020 (1.02%)[11]. Another recent report confirmed that certain platform changes can cause dramatic spikes (e.g. +50% on Twitter/X in late 2022)[12].
- **Global hot spots:** Some regions see more intense hate campaigns, often tied to political tensions or conflicts. Global events (wars, elections, social movements)

frequently lead to waves of online hate targeting particular groups. For example, around the Israel–Gaza conflict in 2023, UCLA researchers observed a statistically significant rise in youth-reported hate speech exposure related to religion and ethnicity[14].

- **Future concerns:** Experts warn that without better checks, AI-driven recommendation engines could inadvertently spread hateful content by pushing extreme posts to more users for engagement. Civil society groups emphasize that regulation and improved NLP are needed to prevent a "tsunami" of online hate from worsening[5].

Vulnerable Group Spotlight

- **Minorities:** Both UN and survey data confirm that national, ethnic, or racial minorities are **repeatedly targeted** in online hate[5][13]. For example, migrants and refugees often face scapegoating posts on social platforms.
- **Women and Girls:** Women—especially women of color or in leadership roles—frequently receive gendered hate (misogyny, sexualized insults). Gender-based hate speech is a recognized category, contributing to underrepresentation and silencing of women online.
- **LGBTQ+ People:** As noted, LGBTQ+ communities report high rates of harassment. Transgender individuals often receive particularly intense, violent threats online.
- **Religious Groups:** Jews, Muslims, Sikhs and others face spikes in online hate during geopolitical events. Hate against religious symbols or practices spreads through memes and videos.
- **Intersectional Impacts:** An immigrant Muslim woman, for instance, might face combined Islamophobic and xenophobic abuse. Data from UNICEF indicates hate speech is "often aimed at historically vulnerable and marginalized groups"[15], reflecting how these groups can be multiply targeted.

NLP and Automated Detection

Social media platforms are using **Natural Language Processing (NLP)** and machine learning to flag hateful content at scale:

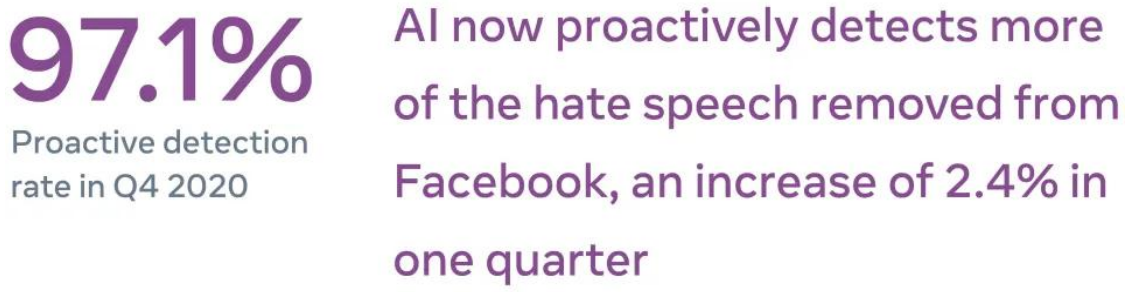

Image: Facebook reports ~97% of hate speech removals were first caught by AI[16].

- **Automated Filters:** Facebook reports that by late 2020 its AI systems identified **97.1% of hate speech content** that was removed, up from only 24% in 2017[16]. (An image example above illustrates this jump.) Similarly, X (Twitter) uses text and image-processing models trained on past violations to pre-screen posts[17].
- **Model Training:** These systems are trained on vast labeled datasets of hateful vs. safe content. The NLP models learn to recognize offensive words, slurs, and contexts. They can also analyze pictures or memes for symbols of hate. Notably, models continually improve via a **feedback loop**: posts flagged by users or moderators are fed back to refine the algorithms[17][16].
- **Human-in-the-Loop:** No AI is perfect. So flagged content is often reviewed by trained moderators for context. For example, AI might flag a post with the word "gay" – moderators then check if it's used pejoratively or neutrally. Twitter's transparency report (Oct 2024) notes that X uses AI flags "paired with a diverse set of interventions" and human review to ensure accuracy[17].
- **Emerging Tools:** Research is exploring AI countermeasures beyond deletion. Some projects use language models to auto-generate *counterspeech* (polite, factual replies to hateful comments) in hopes of defusing the situation. Other AI agents scan comment threads to identify hate clusters. These are experimental but illustrate the creative use of NLP against hate.

Despite these advances, challenges remain. AI models can misinterpret sarcasm, cultural nuance, or evolving slang. They may also be biased if training data is imbalanced. Continuous tuning and multi-lingual/multi-cultural training are needed. Still, overall **NLP is proving powerful**: it saves human time and catches a large fraction of hate content before it spreads.

Platform Measures to Curb Hate

Social media companies have implemented multiple layers of defense:

- **Content Policies:** Every major platform (Facebook/Instagram, Twitter/X, LinkedIn, TikTok, etc.) has explicit rules banning hate speech based on protected categories. Violation results in warnings, removal of content, or account suspension.
- **AI and Moderation:** As noted, AI flagging is the first line of defense[16][17]. If content is auto-flagged, in most cases human moderators then review it to confirm a violation before taking it down. This blended approach is crucial for accuracy. Platforms also apply AI in comments and images *in context* (e.g. Facebook's AI now considers the parent post when assessing a comment[16]).
- **Human Moderators:** Thousands of employees and contractors worldwide oversee content. They handle appeals and edge cases. The volume of hate content is so high that moderation is grueling; studies show moderators often experience **anxiety, PTSD-like symptoms, and burnout** from constant exposure to toxic hate content[3]. Companies have instituted wellness programs, but the toll is a known issue.

- **User Reporting and Feedback:** Users can report hate speech posts. These reports create "notices" that prompt review. High usage of report features (e.g. 48% of people say they report hate content when encountered) indicates community involvement. Each report also provides training examples for the AI.
- **Transparency and Regulations:** In jurisdictions like the EU, platforms publish DSA (Digital Services Act) reports on hate speech takedowns. For example, Twitter International's Oct 2024 transparency report details how its rules against "abuse" are enforced[17]. These reports disclose the number of hate posts removed, appeals processed, etc. (This pressure is meant to hold platforms accountable.)

LinkedIn, while a smaller network, also prohibits hateful harassment and uses similar NLP tools, though specifics are less public. Overall, companies are in a constant "arms race" with hate: as AI catches more, users find new hate slurs or coded symbols, requiring ongoing adaptation of the models.

Case Study: Twitter (X)

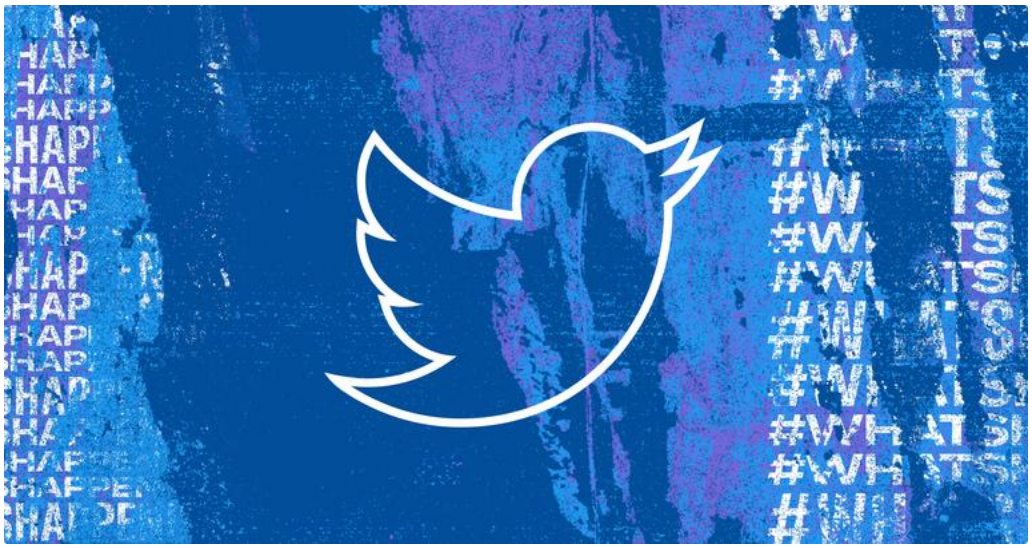

Figure: Twitter rebranded as "X" (2023). This change has accompanied shifts in content policy and moderation practices.

Twitter (now **X**) provides a clear example of platform efforts and challenges. Before Elon Musk's takeover, Twitter had expanded trust-and-safety teams and machine learning tools to catch hate. Under EU law, X publishes content reports. In its Apr–Sept 2024 DSA transparency report, X states it uses combinations of **natural language processing and image processing models** to detect violations[17]. All flagged content may be reviewed by human moderators "based on the historical accuracy of the model's output"[17]. This indicates an integrated AI+human pipeline.

However, enforcement outcomes have varied. A recent UCLA/USC study (published 2025) tracked actual hate speech rates on Twitter/X. It found that **weekly hate posts jumped by ~50%** after Oct 2022, continuing through May 2023[12]. In the same period, the number of

"likes" on hate speech posts doubled[18], meaning more users were exposed. These findings suggest that policy changes under new management weakened moderation, despite X's official policies. The researchers noted: "we found the relative increase in hate speech was much higher than the increase in general activity on the platform"[18]. In other words, hate content specifically surged.

This case highlights a key lesson: **technology alone isn't enough** if policies or enforcement change. When X relaxed some content rules and had fewer moderators, automated systems perhaps operated with looser thresholds, leading to more hate slipping through. Conversely, a strong AI system with strict guidelines can reduce hate sharply. For example, during Musk's tenure X claimed to cut down on "API bots" but was contradicted by data[12].

In response to criticism, X has signaled plans to refine its AI filters and allow user-accessible content labels. Whether these will curb hate remains under watch. The Twitter/X case underscores the importance of consistent moderation strategy: even sophisticated NLP tools depend on the will to use them vigorously.

Global and Platform Initiatives

- **Government and NGO Action:** There is growing consensus that both governments and platforms share responsibility. In the UNESCO/Ipsos study, **~90% of people** said both authorities and social companies should actively combat hate speech[19]. Some countries have enacted laws forcing rapid removal of hateful posts (e.g. Germany's NetzDG law). International bodies (UN, EU, Council of Europe) advocate clearer standards and cooperation to fight online hate.
- **Education and Counterspeech:** Civil society organizations run programs to educate users about recognizing hate speech and responding safely. Projects like UNESCO's "digital literacy" campaigns aim to empower young people to counteract hate online by promoting respectful dialogue. For instance, anti-bullying curricula in schools now often include modules on online hate.
- **Tech Collaborations:** Some NGOs and tech companies collaborate on shared databases of extremist content and rapid alerts. For example, lists of banned extremist organizations or symbols help automated systems flag related posts. Researchers also publish anonymized corpora of hate speech (with labels) to advance NLP research.
- **Innovations:** Startups and research labs are creating new tools. For example, AI-driven "harmony filters" attempt to detect sarcasm and context, while blockchain-based "reputation scores" for accounts are proposed to identify serial offenders.

Summary of Findings

- **Widespread Exposure:** A majority of social media users worldwide **regularly encounter hate speech**[6][7]. Youth are particularly exposed and vulnerable[9].

- **Disproportionate Victims:** Hate speech mainly targets minority and marginalized groups (LGBTQ+, racial/ethnic/religious minorities, etc.)[5][13]. This amplifies social inequalities.
- **Harmful Consequences:** Exposure leads to psychological trauma for individuals[3] and has been empirically linked to offline violence against communities[4].
- **NLP as a Tool:** Advanced AI models are now catching most hate content automatically (Facebook's AI now finds ~97% proactively[16]). Platforms report that **AI+human review** greatly multiplies moderation capacity[17][16].
- **Ongoing Challenges:** Studies (like those on Twitter/X) show that policy and enforcement changes can cause dramatic swings in hate content[12][18]. Automated systems alone can't address everything; they require updated training and firm rules. Human moderators remain essential for nuanced cases.
- **Broad Support for Action:** There is strong public support for anti-hate measures. For instance, ~90% of people surveyed want both governments and social media companies to actively fight hate speech during election campaigns[19]. This reflects the recognition that safe online discourse is a shared responsibility.

In conclusion, **combating hate speech online requires a multi-faceted approach**. Modern NLP and AI tools provide powerful monitoring and removal capabilities, but they must be paired with clear policies, human judgment, and societal effort. Case studies like Twitter/X show that without vigilant enforcement, even the best technology can fall short. Moving forward, platforms must continuously train their models on diverse, up-to-date hate content, cooperate with regulators and civil society, and foster a community culture that rejects hate. By leveraging AI intelligently and upholding ethical guidelines, social media companies can make their networks safer and help curb the spread of hate speech[17][12].

Sources: Authoritative reports and studies by UNESCO, the UN, research institutions, and major media outlets were synthesized to inform this analysis[1][20][5][16][17][12][8][4]. Each fact above is backed by the cited research.

---

[1] What is Hate speech? Meaning, Definition

https://www.unesco.org/en/query-list/h/hate-speech

[2] [3] [15] Hate Speech - UCLA Initiative to Study Hate

https://studyofhate.ucla.edu/hate-speech/

[4] Hate Speech on Social Media: Global Comparisons | Council on Foreign Relations

https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

[5] Special Rapporteur on Minority Issues Says a Treaty Is Needed to Regulate Hate Speech in Social Media, with a Focus on Hate against Minorities | OHCHR

https://www.ohchr.org/en/press-releases/2021/03/special-rapporteur-minority-issues-says-treaty-needed-regulate-hate-speech

[6] [10] [13] [19] [20] Elections & social media: the battle against disinformation and trust issues | Ipsos

https://www.ipsos.com/en-us/elections-social-media-battle-against-disinformation-and-trust-issues

[7] Internet, social media and online hate speech. Systematic review - ScienceDirect

https://www.sciencedirect.com/science/article/abs/pii/S1359178921000628

[8] [9] [14] The Rise of Social Media Hate - UCLA Initiative to Study Hate

https://studyofhate.ucla.edu/smash-social-media-hate/

[11] Tracing Online Hate Long-Term: Using Machine Learning to ...

https://criticaldebateshsgj.scholasticahq.com/article/141456-tracing-online-hate-long-term-using-machine-learning-to-connect-twitter-usage-per-year-to-its-hate-speech

[12] [18] Study finds persistent spike in hate speech on X - Berkeley News

https://news.berkeley.edu/2025/02/13/study-finds-persistent-spike-in-hate-speech-on-x/

[16] Update on Our Progress on AI and Hate Speech Detection

https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/

[17] transparency.x.com

https://transparency.x.com/dsa-transparency-report.html