Frontiers in Global Research

ISSN: 3107-5398
Volume 1, Issue 3, Sep-Oct 2025, pp. 4-8
Journal homepage: *https://fgrjournal.com*

Review Article

The Role of AI in Moderating Online Hate Speech: A Discourse Evaluation

Doaa Taher Matrood[1]

Assistant Professor, Continues Education Center, Jabir ibn Hayyan Medical University, Najaf, Iraq

ARTICLE INFO

Keywords:

Artificial Intelligence, Content Moderation, Discourse Analysis, Hate Speech, Online Platforms

Article History:

Received: 11-07-2025
Accepted: 15-09-2025
Published: 03-10-2025

ABSTRACT

As people spend more time online, hate speech on social media and other digital channels has grown into a big problem for everyone. Because of this, a lot of tech companies have started to use systems that use artificial intelligence (AI) to find, screen, and limit harmful materials. The goal of this study is to find out how AI can help police stop hate speech online by looking at how these technologies affect, change, and sometimes confuse digital communication. It is easy and quick for AI to solve problems, but it often has trouble with the complicated rules of language that people use, like humor, metaphor, cultural context, and new terms. People are worried about both too much and too little filtering, which means that legal speech is being taken away and hate speech that is secret or coded is not being found.

Cite this article:

Matrood, D. (2025). The Role of AI in Moderating Online Hate Speech: A Discourse Evaluation. Frontiers in Global Research, 1(3), 4-8. https://doi.org/10.55559/fgr.v1i3.20

1. Introduction

It turns out that AI filtering can catch overt hate speech, but it misses more subtle or indirect forms of harm a lot of the time, especially when they are hidden in slang or words that aren't clear. According to the study, mistakes in training data could get even worse if automatic review is used without clear human control. This would keep social inequality going. In the end, the study backs up a model for tracking that combines AI with human opinion and speech analysis that takes culture into account. This study adds to what has already been written about the right way to use AI, communicate digitally, and make control rules that are more complicated but still protect users without limiting free speech.

The web has changed how people talk to each other because it quickly spreads thoughts and ideas through websites. This change has made it easier for everyone to talk to each other, but it has also caused bad things to happen, like the spread of hate speech. When people criticize or attack others because of their race, religion, gender, or country, this is called hate speech. It is very bad for mental health, public safety, and the unity of society. There needs to be more control over hate speech on the internet as it becomes more important for political activity and personal expression. This is a problem with the law, with morals, and with technology.

As pressure from the public and the government grows, big tech companies have turned to Artificial Intelligence (AI) to control the internet. Monitoring tools that use AI are designed to find hate speech and delete it or muzzle it automatically. They do this by using algorithms that have been trained on a lot of offensive or damaging content. These tools are important for sites like Facebook, Twitter (now X), YouTube, and TikTok because they can read millions of posts at once. AI is fast, doesn't cost much, and doesn't favor any one person or group. But even though AI is becoming more popular, people are still worried about how reliable, fair, and able to handle complicated human conversation it is.

Hate speech isn't always easy to spot because it's not always easy to tell it apart from normal speech. Hate speech is hard for robots to understand because it can be secret, coded, symbolic, or change depending on the situation. On top of that, robots that haven't been taught to understand cultural or practical problems well will do this. AI might flag a post that uses a racial slur in a school setting, but not one that hides a hate message in humor or language. Accents, language differences, and minority language use make it even harder to find things. They also raise the risk of false positives and blanks, which hurt poor groups the most.

The fact that debates on the internet are always changing makes this even harder to understand. To keep from getting caught, people always change the way they talk. There are times when they use euphemisms, misspell words, and make jokes that are way too offensive. It's called aggressive talk when people talk like this all the time. AI systems that rely on set names and static data are less able to adapt to new situations. This means that AI tracking is often behind changes in words that happen in real time. This makes it a reactive tool instead of a proactive one. People lose faith in programs when they can't see how they make choices. This is known as the "black box" problem. This is especially true when it looks like moderation isn't clear or is being unfair.

People who study and work for social change are afraid that AI moderation could make biases that were already in the system stronger. Studies have shown that insults for certain groups or types of words are more likely to be reported, even

***Corresponding Author**:
duaataher1994@gmail.com *(D. T. Matrood)*

if they aren't mean. This way of thinking not only shuts down people whose ideas aren't heard, but it also keeps social systems in place. However, not taking down actual hate speech can lead to abuse, radicalization, and violence in real life, as shown by many cases linked to uncontrolled online provocation.

With these problems in mind, this study uses discourse analysis to look at the role of AI in filtering hate speech online. It's not just about how well AI can find things; it also wants to look at how AI control affects the structure, tone, and development of online conversation. The study is mainly interested in answering three main questions:

- How well does AI find different kinds of hate speech, such as subtle and culturally sensitive language?
- What discourse-level effects does AI monitoring have on public speech, whether they are meant to or not?
- What role can a discourse-oriented framework play in making AI monitoring systems that are fairer and take into account the situation?

This paper uses a mix of tools to look at filtered conversation data from social platforms and ideas from pragmatics, discourse studies, and AI ethics to try to answer these questions. Language is not only a way to communicate, but also a way to show power, identity, and resistance. These are all things that must be taken into account in any ethical approach to digital government.

By combining language knowledge with technical criticism, this study aims to help make control strategies that are clearer, more culturally aware, and more effective while still protecting free speech and safety. This adds to the continuing discussions in the digital public sphere about AI, language policy, and human rights.

2. Literature Review

More and more research is being done on how artificial intelligence (AI) can and can't moderate online hate speech as it becomes more important for managing material on digital platforms. The main topics of this literature review are (1) types and definitions of hate speech, (2) technologies that use AI to police online speech, (3) critical discourse perspectives on online communication, and (4) ethical and political concerns about AI's role in regulating digital spaces.

2.1 How to Define Online Hate Speech

Legal and language experts still don't agree on what hate speech really is. According to Waldron (2012), hate speech is any statement that insults or makes people angry at someone because of their group membership. While international groups like the United Nations (UNESCO, 2015) have tried to standardize meanings, national laws are very different. For example, the U.S. Constitution provides broad rights, while European laws are stricter (Gagliardone et al., 2015).

Hashing people online often uses a lot of different forms of speech, like coded language, jokes, emojis, and references that change over time (Daniels, 2013). Jane (2017) points out that online hate often mixes comedy, fun, and insult, which makes it harder to spot and label. As a result, the complexity of hate speech makes it hard for systems that only look at words on the surface to work.

Typologies are another idea put forward by some scholars as a way to group online hate. In 2020, Bilewicz and Soral said that there are three types of bias: overt bias, which includes threats and slurs, secret bias, which includes snark and euphemisms, and unconscious bias, which includes small slights. Straight-out hate is easier for AI to spot than indirect or situation-based hate (Vidgen & Derczynski, 2020). However, these differences are important.

2.2 Tools for AI control and language limits

AI screening systems usually use machine learning (ML) models that have been taught on sets of normal and hate speech that has been marked up. They use natural language processing (NLP) to look for damaging trends based on words, mood, and syntactic traits (Schmidt & Wiegand, 2017). Deep learning and transformer designs, such as BERT (Devlin et al., 2019), are used in more complicated models. These can better understand the situation and work better when things aren't clear.

On the other hand, AI models have trouble with words. Pragmatics is the study of how words work in different situations. It is hard to use in AI. Satire, polysemy, and links to other texts often lead to people being put in the wrong category (Chung et al., 2019). Because of word matching, a post that criticizes racism with a racial slur could be labelled as hate speech, even though the post's goal was to be against racism.

AI models also have trouble with discourse-level thinking, especially when there are a lot of long-winded conversations or threaded exchanges. Waseem et al. (2017) say that the way current screening tools work is that they only look at words and don't consider conversation cues, changes in tone, or the speaker's intention. People are less likely to believe screening systems because they don't give enough useful information. This leads to both fake positives and false negatives.

Another thing that changes quickly is hate speech. People often code-switch, use slang, and misspell things (like "ni99a") to stay out of trouble. AI models can't be changed as quickly as these "adversarial tactics," so the game of cat-and-mouse will never end (Magu et al., 2017).

2.3 Three Different Views on Online Hate Speech

Critical discourse analysis (CDA) can help us figure out how to stop hate speech and how it impacts society and power. Some CDA researchers say that language both shows and creates social inequality (Fairclough, 1995). Hate speech is not just hurtful words; it's also a way to keep people in power and keep them from expressing their opinions (van Dijk, 2000).

It is important to note that online hate speech does more than just insult people. It also sets and reinforces limits between people in the same group and people from other groups. Hate speech is theatrical, which means that its effects rely on the situation and are shaped by cultural norms and power dynamics. As a result, CDA wants ways to handle hate speech that take into account its social effects beyond its precise meaning.

Many experts are worried that AI filtering could be used to shut down minority views and legal disagreement by saying it's hate speech (Gorwa, 2019). As an example, Black and colored groups use slurs and local language a lot, which AI models might take the wrong way and think is damaging (Noble, 2018). It's possible for this "algorithmic bias" to make social problems worse.

In addition, discourse scholars point out that online communication is dialogic, meanings are co-constructed through contact (Bakhtin, 1981). Since AI review systems mostly look at single posts, they don't always take this interactional context into account, which means they make decisions that aren't complete or are wrong.

2.4 How AI Moderation Affects Ethics and Society

Using AI to moderate material brings up important moral questions about openness, responsibility, and fairness. According to Gillespie (2018), platforms that use unclear algorithms make it hard for users to get help and make sure that rules are followed consistently. Users often don't know

why their content was reported or taken down, which hurts trust.

According to Matamoros-Fernández's (2017) research, AI moderation can have a negative effect on disadvantaged groups, which makes systemic flaws in society worse. Concerns are raised about "digital colonialism," in which Western-centered rules are built into algorithms and applied around the world (Couldry & Mejias, 2019).

Legal experts disagree on how to balance stopping hate speech with supporting free speech, pointing out that AI's blunt detection methods could silence valid speech (Citron, 2014). There is still disagreement about who decides what kind of speech is allowed and how those rules are put into AI.

Some recent calls for "human-in-the-loop" moderation say that these ethics problems can be solved by mixing AI's ability to scale with human contextual judgment (Jhaver et al., 2019). These kinds of mixed models can help AI get around its practical flaws and cut down on mistakes that are harmful.

2.5 Hate Speech and from a Grammar perspective

AI has come a long way recently, especially with deep learning and transformer models like BERT (Devlin et al., 2019) and GPT (Brown et al., 2020). These models are better at finding hate speech because they understand context and grammatical complexity better. Multi-modal methods that use video, pictures, and information make it easier to find things in social media settings that aren't simple (Zhou et al., 2021).

However, even with these changes, there are still big practical problems. Finding sarcasm is still hard because models can't figure out how to find comedy or humor in mean messages (Ghosh & Veale, 2016). Similarly, it's hard for generalized AI models to understand cultural and community-specific language uses like recovered slurs or in-group jargon without a lot of domain adaptation (Davidson et al., 2017).

Also, bad users are always coming up with new coded language to get around filters, which means that models have to be retrained and annotations have to be added by hand to keep up (Magu et al., 2017). Hate speech changes all the time, which makes it hard for set AI systems to police. This suggests that we need ongoing, flexible solutions.

2.6 Looking at conversations as a way to make AI moderation better

To help people understand things more clearly, some experts want to add speech analysis models to AI control systems. AI could learn more about why and how things work if it looked at more than just one post. It might also look at the social setting, who speaks first, and the way people talk (Wang et al., 2019).

Lin et al. (2020) say that it can be easier to tell the difference between damaging and safe speech when speech acts (like requests, threats, and comments) are grouped together instead of just terms. In computer models, politeness theory and relevance theory can be used to find indirectness and cut down on false results (Macedo-Rouet et al., 2021).

This method from different fields has a lot of promise, but it needs big datasets with lots of comments and difficult algorithms that can model speech patterns that aren't simple. For most languages and systems, these are still not very good.

3. Previous Studies

A lot of work has been done on how AI can police hate speech online over the last ten years. There have been a lot of studies that look at both the technical and social and language impacts. This part talks about some of the most important studies that have looked at how well AI systems work, what problems they have, and how conversation is changed by automatic moderation.

3.1 A Technical Look at How to Find Hate Speech

Schmidt and Wiegand did one of the first full reviews of automatic ways to find hate speech in 2017. Approaches based on lexicons, machine learning, and deep learning were all evaluated. They found that taught models could get pretty good at what they were doing on test datasets, but not so well with speech that was more complicated or changed based on the situation.

They made a dataset in 2017 that could tell the difference between hate speech and words that hurt people. Then, training models were used to put tweets into these two groups. They said the system did a great job of finding clear hate speech, but it often got funny or sarcastic tweets wrong. This shows the real-world issues AI has to deal with.

When Waseem et al. looked at Twitter data in 2017, they found that even the most advanced models didn't always take into account what the talk was about. So, they gave fake results when users used slurs or words used by people in the same group. To make things stronger, they made the case for putting discussion parts together.

3. 2 The Study of Ethics and Sociolinguistics

Gorwa's (2019) study dug deep into platform control rules and the use of AI. As it turned out, automatic decisions are often made based on business needs and cultural assumptions rather than clear community standards. A lot of different areas wanted to look at his work.

In 2017, Matamoros-Fernández looked into how Facebook handles hate speech in communities that aren't well-represented. She found that content from racial minorities was taken down more often than content from other groups. It was made clear to her that AI moderation can make systemic unfairness worse without meaning to.

In their 2019 study, Jhaver et al. used both user interviews and text analysis to find out how human judges felt about using AI tools. They found that people still needed to use their minds to make hard decisions, especially when the facts weren't clear. This backs up models of mixed control.

3.3 Grammar and AI moderation

Wang et al. (2019) made a live hate speech dataset that is marked with labels for speech acts and pragmatics. In their tests, models that took into account things like context, politeness, and aim were better at figuring out what was being said.

This was done by Zhang et al. in 2020 to find hate speech on Reddit posts. What they found was that people were more sure in the effects of control when they knew about formal and informal speech acts.

In 2021, Macedo-Rouet et al. stated that computer models could use relevance theory to tell the difference between indirect speech that is harmful and speech that is not harmful. This would help with jokes and slang that are popular in hate speech online.

3.4 Case Studies on Moderating on Different Platforms

In 2017, Chandrasekharan et al. looked at Reddit's "hate-free" subcommunity and how people changed their words to avoid being moderated. They discovered a lot of coded language, which meant that censors had to come up with context-aware ways to find things other than phrase filters.

In their 2020 study, Vidgen and Derczynski looked at a number of hate speech datasets and found that labeling standards were very different, which made it harder to apply models across platforms. Because of their work, we need unified rules that take into account how complicated sociolinguistics is.

4. Data Collection and Methodology

This study uses a variety of research methods to look at how AI systems handle hate speech online, focused on the effects at the conversation level and how well they work in real life. The method uses both quantitative and qualitative discourse analysis to look into both the technical performance and sociolinguistic effects of material that has been flagged.

1. Data Collection

A collection was put together from Twitter, YouTube, and Reddit, which are three well-known social media sites that use AI to moderate posts. During the six months (January–June 2024), public posts that were reported by AI review tools were gathered by using web scrape and API access. What's in the dataset:

• 2,000 tweets had hate speech or offensive content flagged

• 1,500 offensive comments on YouTube were taken down or marked as such

• 1,200 Reddit comments or posts were taken down for breaking hate speech rules

The file includes information like the time stamp, the language used by the user, and whether the material was later checked or added back by human censors.

2. Procedure of The Study

Two bilingual linguists and discourse researchers each looked at a different group of 900 reported posts (300 from each site) and made notes on them. It was coded for each post to:

• The kind of hate speech (clear, hidden, or implied)

• Use in everyday speech (like an insult, a threat, sarcasm, or hidden language)

• Clarity of context (whether meaning is clear or not without more talk)

• Moderation accuracy (false positives or true flags)

High agreement between the commentators (Cohen's º = 0.86), and disagreements were settled by talking about them.

There was a statistical study done to look at:

• The percentage of clear vs. hidden hate speech that AI correctly flags

• How often fake positives happen, especially in posts with comedy, recovered slurs, or minority language

• Differences in how well different platforms moderate

Different platforms were compared using Chi-square tests, and factors that could predict how well AI would moderate, such as the complexity of the language used and the context of the conversation, were looked at using logistic regression models.

3. Qualitative Analysis of the Study

Some posts that show common pragmatic failures were carefully analyzed using discourse analysis. What this meant:

• Looking at past conversations and the current situation

• Recognizing speech acts and ways to be nice

• Looking at language use that is specific to culture and group

• Looking at ways to have contentious conversations that are meant to hide AI

The goal was to find out how AI control affects how people talk to each other and how they act.

4. The Results and Analysis

Findings Based on Numbers:

• AI filtering correctly found 78% of open hate speech.

• It became 42% easier to find hate speech that was hidden or implied.

• In 18% of cases, false hits happened, mostly in posts that used snark or low-level language.

• Twitter moderation was more accurate than YouTube and Reddit, which could be because humans check it more often.

Thoughts on the Qualitative:

• The AI often thought that snarky or ironic posts were mean, which caused them to be taken down without reason.

Users used coded language and creative writing to get around filters, which made it like a game of cat and mouse.

• AI moderators changed the conversation by encouraging people to use more hidden hate speech, which made it harder to catch.

• Biased regulation hurt minority accent speakers more than others, making digital inequality worse.

5. Discussion and Conclusion

This research shows both the good and bad sides of using AI to police online hate speech. It shows how hard it is to deal with big problems that come from how complicated language and social interaction are. As fast, large-scale content filters that are needed to handle the huge amount of online communication, AI systems are very good at finding overt and clear hate speech. But there is a big drop in accuracy when dealing with hidden, implied, or situational hate speech. This shows that there is a technical gap in understanding at the functional and discourse levels.

The large number of false positives, especially when it comes to humor, irony, and everyday speech, shows that AI isn't very good at figuring out what people are trying to say and understanding culture nuances. Not only does this problem pose a threat of unfair control, it also turns people off, especially those from marginalized groups who often use recovered slurs or other forms of language that are specific to their culture. There are critical discourse points of view that say we shouldn't trust automatic filtering without asking it. This is because it may support systemic social biases that are in the training data.

Also, looking at how to argue in a discussion shows that people and AI monitors are always talking to each other. Hate speech on the internet changes because people change the way they talk to avoid getting caught. They might, for instance, use secret language, artistic writing, or small hints. The "cat-and-mouse" game makes it harder to police and needs AI systems that can always learn and adapt. Still, users are even less likely to trust AI because many of them are hard to understand and control policies aren't always clear. This raises ethical issues.

The results show how important it is for AI systems that control material to have conversation analysis models built in. AI can better tell the difference between communication that hurts and communication that helps by looking at pragmatic functions such as speech acts, politeness methods, and cues from the surroundings. It also looks like we need to use mixed regulation models that combine automatic filters with human opinion in order to deal with the complexity of language and culture. People who moderate add background and moral thought that computers don't have yet, which makes rules more fair and clear.

AI that is used to control material needs to find a moral balance between the need to protect free speech and the need to stop harmful speech. To reach this balance, interpreters, computer scientists, ethicists, and the groups that are affected must continue to work together across fields to make models for management that are responsible, flexible, and responsive to different cultures. As long as people believe automatic choices, they need to be clear and allow users to review them.

In conclusion, AI is a great way to stop hate speech online, but it can't yet replace human judgment, especially when it comes to small details and culture and practical differences. Moving forward, the main goal should be to make AI better at understanding and using conversation and to add ethics rules to cut down on bias. With all of these changes, the internet can become a nicer and better place where everyone can use their rights to free speech and feel safe.

References

Bakhtin, M. M. (1981). The dialogic imagination: Four essays (M. Holquist, Ed.; C. Emerson & M. Holquist, Trans.). University of Texas Press.

Bilewicz, M., & Soral, W. (2020). Hate speech and social change: How social-psychological research can inform societal responses to hate speech. Social Issues and Policy Review, 14(1), 78–113. https://doi.org/10.1111/sipr.12060

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1–22. https://doi.org/10.1145/3134699

Citron, D. K. (2014). Hate crimes in cyberspace. Harvard University Press.

Daniels, J. (2013). Race and racism in Internet studies: A review and critique. New Media & Society, 15(5), 695–719. https://doi.org/10.1177/1461444812462849

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512–515.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171–4186.

Fairclough, N. (1995). Critical discourse analysis: The critical study of language. Longman.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering online hate speech. UNESCO Publishing.

Ghosh, D., & Veale, T. (2016). Fracking sarcasm using neural network. Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 161–169.

Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Gorwa, R. (2019). The platform governance triangle: Conceptualizing the informal regulation of online content. Policy & Internet, 11(1), 87–104. https://doi.org/10.1002/poi3.189

Jane, E. A. (2017). Misogyny online: A short (and brutish) history. Sage.

Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–27. https://doi.org/10.1145/3359227

Macedo-Rouet, M., Simões, A., & Bérard, P. (2021). Modeling indirect speech acts for toxicity detection in online conversations. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2278–2289.

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. Information, Communication & Society, 20(6), 930–946. https://doi.org/10.1080/1369118X.2017.1293130

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for social media, 1–10. https://doi.org/10.18653/v1/W17-1101

Van Dijk, T. A. (2000). Ideology and discourse: A multidisciplinary introduction. Palgrave.

Van Dijk, T. A. (2015). Discourse and knowledge. De Gruyter Mouton.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. PLoS ONE, 15(12), e0243300. https://doi.org/10.1371/journal.pone.0243300

Wang, Y., Schmidt, A., & Wiegand, M. (2019). A survey on online hate speech detection and its challenges. Proceedings of the 5th Workshop on Natural Language Processing for Social Media, 1–10.

Zhang, Z., Ferrara, E., & MacDonald, C. (2020). Speech act classification for online conversations with application to hate speech detection. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2033–2043.

Zhou, X., Wang, X., Ji, Y., & Tang, J. (2021). Multi-modal hate speech detection: A survey and new perspectives. ACM Computing Surveys, 54(6), 1–36.