Chapter 2
From Reactive to Proactive:
AI's Role in Promoting Civility and Respect in Online Discussions

Swati Chakraborty
https://orcid.org/0000-0003-0799-1954
Concordia University, Canada

ABSTRACT

The proliferation of social media and online forums has democratized discourse but also led to an alarming rise in incivility, hate speech, and toxic interactions. Historically, the approach to addressing these issues has been reactive—identifying and removing harmful content after it has already caused damage. However, advances in artificial intelligence (AI) present an opportunity to shift from reactive to proactive strategies for fostering civility and respect in online discussions. This paper examines the evolution of AI-driven content moderation tools and explores how they can be leveraged to not only detect and filter inappropriate content but also encourage respectful engagement. By emphasizing preemptive interventions, AI can become a powerful force in shaping healthier online communities.

1. INTRODUCTION

Online platforms have become integral spaces for public discourse, allowing individuals from diverse backgrounds to exchange ideas. While this has led to the democratization of speech, it has also introduced new challenges, notably the rise of toxic behavior, including hate speech, trolling, and harassment. Traditionally, content moderation has relied on user reporting and human moderators to manage

DOI: 10.4018/979-8-3693-9904-0.ch002

such behavior. However, these methods are largely reactive, addressing harmful content after it has already been posted.

With the advent of AI technologies, there is an opportunity to transform this reactive approach into a proactive one. By using machine learning algorithms, natural language processing (NLP), and predictive analytics, AI can not only detect but also anticipate potentially harmful behavior. More importantly, AI can be designed to promote positive engagement by nudging users towards more respectful discourse, setting the tone for constructive conversations.

This paper explores the transition from reactive to proactive AI systems in online content moderation, focusing on the role AI can play in promoting civility and respect in online discussions (M. N. O. Sadiku, 1989).

The proliferation of social media and online forums has democratized discourse but also led to an alarming rise in incivility, hate speech, and toxic interactions. Historically, the approach to addressing these issues has been reactive—identifying and removing harmful content after it has already caused damage. However, advances in artificial intelligence (AI) present an opportunity to shift from reactive to proactive strategies for fostering civility and respect in online discussions. This paper examines the evolution of AI-driven content moderation tools and explores how they can be leveraged to not only detect and filter inappropriate content but also encourage respectful engagement. By emphasizing pre-emptive interventions, AI can become a powerful force in shaping healthier online communities.

Proactive AI moderation has demonstrated success in creating healthier online communities by preventing harmful interactions before they escalate. For example, Reddit's **AutoModerator (AutoMod)** allows subreddit moderators to establish specific rules, automatically screening posts and comments for spam, offensive language, or irrelevant content. This system reduces the burden on human moderators and ensures that inappropriate content is removed swiftly, fostering a more respectful atmosphere. By enabling customization, Reddit empowers its users to define community-specific standards, making moderation both effective and adaptable (A. Mufareh.2020).

Similarly, platforms like **YouTube** have leveraged machine learning to improve comment moderation. AI tools flag or hold comments containing harmful or inappropriate language for manual review. Features such as "hold potentially inappropriate comments for review" give creators more control over their content, allowing them to maintain a positive environment while reducing exposure to toxic comments. These tools not only promote civility but also enhance user engagement by creating safer spaces for interaction. Consistent updates to these AI systems reflect a focus on inclusivity and fairness, addressing concerns like algorithmic bias (M. N. O. Sadiku, 2018).