Chapter 6
AI-Generated Hate Speech Detection

Neha Yadav
https://orcid.org/0000-0002-7793-2063
G.L. Bajaj Institute of Technology and Management, Greater Noida, India

Mayank Singh
https://orcid.org/0000-0002-8393-3652
SS University, Noida, India

Vipin Tyagi
https://orcid.org/0000-0003-4994-3686
Jaypee University of Engineering and Technology, Guna, India

ABSTRACT

The advent of large language models (LLMs) in generative artificial intelligence (AI) has fundamentally changed the scope of hate speech online. Where hate speech was traditionally unambiguously, human using AI can now produce aggressive, degrading, or even coded language with incredible fluency, speed, and variation. This chapter investigates the emergent technical, social, and ethical issues of hate speech produced by AI. In addition, the chapter includes a broad survey of the history of hate speech in the internet era, highlights the subtle distinctions between offensive language and hate speech, and provides an analysis of the problematic nature of AI-generated threats, including linguistic variability, scale, and semantic ambiguity. The chapter concludes with thoughts on what it will take to protect digital public spaces in the age of generative AI, including continued technical advances, collective social advancements, and ethical collaborative work across disciplines.

DOI: 10.4018/979-8-3373-3063-1.ch006

1. INTRODUCTION

The rapid development in the area of artificial intelligence (AI), including natural language processing techniques, has changed the way of online interaction of individuals. These days with the help of models like OpenAI's GPT, Meta's Llama, and other large language models (LLMs), the world is now able to produce human-like text on a scale never seen before is possible to generate human like text. These models provide a number of positive possibilities, such as automating conventional communications, assisting with accessibility tools, enhancing creativity etc. However, along with these advantages, risks in online interactions, generation and distribution of hate speech is also possible. Earlier hate speech was manually posting of abusive content by an individual or group of users, but these days AI is used to create and generate hate messages quickly and efficiently (Boutadjine, Harrag, & Shaalan, 2025). Thus, identifying hate speech from any source, must be an essential part of protecting individuals and communities online, maintaining trust in digital spaces, and meeting social and legal expectations.

Users cannot always distinguish the intent behind AI programs that produce hate speech or content promoting hate. It may be either intentionally generated or inadvertently generated as a result of training on biased datasets. In addition, many of the existing LLMs can produce hate speech in unique and subtle, contextually related ways such as rephrasing, or employing culturally specific references. In such cased detection is difficult using traditional keyword or rule-based detection systems. Those promoting hate speech can also repeatedly adjust their input or change the model prompting to prompt the output making it difficult to recognize as hate speech (Wu, Wang, Zhang, Wang, & Pang, 2025). The type of hate speech generated by AI programs requires that researchers, reevaluate their approaches and adopt new detection and mitigation strategies. Hate speech is a language that diminishes, threatens, and marginalizes individuals or groups based on characteristics such as race, religion, gender, sexual orientation, nationality, or disability. Use of AI into generation of hate speech makes it even more challenging to differentiate between accidental and purposeful harm, as well as the distinction between direct slurs versus implicit and coded language.

1.1. Context and Importance

Over the last decade, advancements in artificial intelligence, particularly in the area of natural language processing (NLP). Large Language Models (LLMs) like OpenAI's GPT-T series, Google's PaLM, Meta's Llam are able to produce human-like text (Mijwil, Naji, Doshi, Hiran, Bala, & Ali, 2024) (Vetagiri, Mogha, & Pakray, 2024). While these advancements have created multiple opportunities for beneficial