Chapter 13
Techniques for Detecting Hate Speech in AI and Human-Generated Content

Chandrashekhar Goswami
https://orcid.org/0000-0002-9404-9352
Faculty of Computing and Informatics, Sir Padampat Singhania University, Udaipur, India

Ansar Sheikh
St. Vincent Pallotti College of Engineering and Technology, Nagpur, India

Anand Bhaskar
Sir Padampat Singhania University, Udaipur, India

Dipesh Vaya
Sir Padampat Singhania University, Udaipur, India

ABSTRACT

Hate speech detection is critical due to the rising risk of online and offline offensive material. This chapter reviews methods for detecting hate speech in human-generated and AI-generated content, addressing differences, obstacles, and potential solutions. Various methods detect explicit and implicit hate speech, including rule-based systems, traditional machine learning models, and deep learning. Key issues include biases in model outputs and challenges in recognising AI-generated hate messages, with discussions focusing on fine-tuning pretrained models and multimodal approaches to improve AI-generated content detection. The chapter also examines

DOI: 10.4018/979-8-3373-3063-1.ch013

ethical implications, such as balancing free speech and censorship, privacy issues, accountability in AI algorithms, and fair decision-making. It identifies future trends, including natural language processing (NLP) advancements, real-time detection systems, and evolving regulatory frameworks for AI content moderation.

INTRODUCTION

Generally, hate speech is defined as any speech, gesture, or display that incites violence or prejudicial action against or by a particular individual or group, or because it disparages or intimidates a particular individual or group, with the specific hate speech definition varying by jurisdiction (Schmidt & Wiegand, 2017). The rise of the internet and digital platforms, especially social media, has vastly expanded the reach and speed of hate speech, giving individuals the tools to broadcast their words to large audiences instantly. Although most human societies have understood the harm of hate speech for centuries, the steep surge in online user interaction has made identifying and regulating hate speech an unprecedented challenge (Davidson et al., 2017).

Hate speech is not just a matter of verbal insult in modern-day society. It has real-world consequences, from psychological damage to physical violence, and leads to the fragmentation of society and the loss of social cohesion and trust (Fortuna & Nunes, 2019). For instance, hate speech not only benefits the perpetrators of the crimes but also harms the victims because hate or victimisation can arise because of hate speech, encouraging children and weaker people to isolate themselves. Moreover, it also serves to normalise harmful online rhetoric, which turn, results in the distribution of extremist ideologies and the marginalisation of communities already vulnerable (MacAvaney et al., 2019).

Now we must deal with AI-generated content, whether in deepfakes or text from large language models. The hate speech produced by machines can be valuable as that written by any human, which results in a serious challenge for us because it is hard to identify whether a human or a machine writes the particular content (Bolukbasi et al., 2016). As AI models used for content generation become increasingly sophisticated, they have also enabled hate speech to manifest in more subtle forms that may not immediately be identified as harmful (Burnap & Williams, 2015). New developments must consider existing hate speech and continue to develop and adopt new detection tools and mechanisms for hate speech originating with technology.

Such content can have a large social and ethical impact, which makes the design of effective hate speech detection systems essential. As just written, hate speech is harmful because it harms individuals and also creates division between communities based on fear, prejudice, and tensions between communities (Davidson et