Chapter 4
Algorithms Evaluation and Challenges in Automated Hate Speech Detection for Generative AI

Rituraj Jain
https://orcid.org/0000-0002-5532-1245
Marwadi University, India

Ashish Sharna
https://orcid.org/0009-0007-6644-6362
Jodhpur Institute of Engineering and Technology, India

Venkateswararao Pulipati
Koneru Lakshmaiah Education Foundation, Bowrampet, India

Nausheen Khilji
JIET University, India

Rakesh Saxena
Jai Narain Vyas University, India

ABSTRACT

The research investigates automatic hate speech identification tools meant for detecting text generated by advanced generative AI systems including GPT and BERT. The text provides details about traditional machine learning as well as contemporary

DOI: 10.4018/979-8-3373-3063-1.ch004

deep learning methodologies while putting highlight on transformer models that excel at detecting subtle and context-driven hate speech. The analysis is focused on essential evaluation metrics together with benchmark datasets but also explains implicit toxicity as well as biases through datasets and cultural interpretations challenges. This paper explores both ethical aspects and mitigation techniques alongside real-time moderation system integration practices. The section ends by providing predictive perspectives about multimodal detection methods as well as zero-shot learning and responsible AI deployment frameworks.

1. INTRODUCTION

Automated hate speech detection is an important field of study that gains research momentum with the emergence of generative AI. With AI growing and deployed in many different models, it has become increasingly clear that there is a desperate demand to detect and stop the spread of harmful content. Linguistic and semantic analysis combined with state-of-the-art machine learning techniques are employed in the detection of hate speech, and barriers still exist. Hate speech is associated with sarcasm, slang that changes with time, cultural allusions, and nuances often hard to make sense of by a machine. The people who commit these crimes are constantly modifying language to avoid being caught, making any anti-detective system go out of date all the time. Cultivation of good, multi-lingual and culturally aware data is a crucial yet complicated task with ethical challenges, such as biasness, annotator trauma and censorship. Suppression of online freedom of expression with the safety of online users is a very important legal and social concern. However, automated detection is crucial in regards to ensuring responsible usage of generative AI services across every platform, the development of healthy online culture, and the development of AI ethics by promoting transparency, accountability, and socially worthwhile development.

1.1 Evolution of Generative AI and its capabilities

In the past few decades Generative Artificial Intelligence (AI) has grown to a large extent from the simplest rule-based systems to complex models which create human like content in many different domains. In early logic-based systems like the Logic Theorist, the journey began from there, through statistics, and to the deep learning (Radanliev, 2024). This was a rung in this evolution, GANs and VAEs being a revolution in generating very realistic content (Kar et al., 2023). These models would form the base for more complex architectures, which include the GPT series from Generative Pretrained Transformer and Google's BERT series with great suc-