Chapter 14
Machine Learning Models for Automated Hate Speech Detection in Synthetic Content

Tarni Khatri
https://orcid.org/0009-0008-4269-7949
Manipal University Jaipur, India

Vibhakar Pathak
https://orcid.org/0000-0002-0916-9326
Arya College of Engineering and Information Technology, India

Rohit Mittal
https://orcid.org/0000-0001-7688-5421
Manipal University Jaipur, India

Manish Mittal
https://orcid.org/0000-0002-7029-9576
Brainware University, Kolkata, India

ABSTRACT

This chapter focuses on how Machine Learning (ML) can help identify and reduce hate speech on the internet. With the growing use of Artificial Intelligence (AI), there is a risk that these systems can unintentionally spread or support hate speech. This often happens because the data used to train these models may contain biased or harmful language, and the models may not fully understand the context in which words are used. The chapter explains how different ML techniques—such as supervised learning (where the model learns from labelled examples), unsupervised learning

DOI: 10.4018/979-8-3373-3063-1.ch014

(where the model finds patterns in data without labels) can be used to detect hate speech in text. The chapter also discusses how we can measure the performance of hate speech detection models using metrics such as precision (how many detected hate speech examples were actually hate speech), recall (how many real hate speech examples were correctly found), F1-score (a balance between precision and recall), and ROC-AUC.

1. INTRODUCTION

Hate speech refers to any kind of communication—spoken, written, or behavioural—that attacks, threatens, or discriminates against a person or group based on attributes like race, religion, ethnicity, gender, sexual orientation, disability, or nationality. Hate speech targets marginalized or disadvantaged social groups in ways that can be harmful to them (Jacobs and Potter, 2000; Walker, 1994).

As we step deeper into the industry 4.0, Artificial Intelligence (AI) is no longer a futuristic concept—it's quickly becoming part of our everyday lives. From smart assistants to personalized ads, AI is transforming how businesses operate, how services are delivered, and even how we interact with each other. Machine Learning (ML), promise faster growth and better customer experiences. It has challenge like AI systems mimicking human behaviour, they also risk spreading the same biases and harmful patterns found in the real world—including hate speech. Algorithms can amplify harmful content, when trained on large language models (LLMs) like BERT, GPT (Mijwil et al., 2024; Garg et al., 2025).

However, as we rely more on these models, especially in hate speech detection, new challenges have surfaced. While LLMs can recognize patterns and language, these can reflect biases found in the data they're trained on. As AI continues to shape the digital spaces we inhabit, it's crucial to understand both the strengths and blind spots of these models.

1.1. Sources of AI-Generated Hate Speech

a) Biased Training Data

One of the main reasons AI systems sometimes generate hate speech is because of the data they're trained on. LLM's learn from massive collections of text gathered from the internet places like forums, social media, and online articles. These training datasets often contain harmful stereotypes, offensive language, and biased

Analysing Deepfake-Related Scandals in Higher Education: Case Study Insights

J. Rahila, Shakhriyor Kholbayev, Renu Jahagirdar, Sujit Kumar Acharya, R. Reginand A. Thenmozhi (2026). *Safeguarding Educational Integrity Through Deepfake Face Detection (pp. 145-168).*

www.irma-international.org/chapter/analysing-deepfake-related-scandals-in-higher-education/398643