Chapter 2
Detecting Hate Speech:
Human and AI-Generated Content

Kashvi Chaturvedi
https://orcid.org/0009-0005-3192-3400
Ajeenkya D.Y. Patil University, India

Aditya Shrivastav
https://orcid.org/0009-0000-8461-8171
Ajeenkya D.Y. Patil University, India

Yadnyesh Khapekar
Ajeenkya D.Y. Patil University, India

Atharva Haresh Saraf
https://orcid.org/0009-0006-3842-5508
Ajeenkya D.Y. Patil University, India

Sunil Sankathala
https://orcid.org/0009-0004-7809-7647
Ajeenkya D.Y. Patil University, India

Susanta Das
https://orcid.org/0000-0002-9314-3988
Ajeenkya D.Y. Patil University, India

ABSTRACT

Hate speech refers to language that incites hostility, discrimination, or violence against individuals or groups based on attributes such as race, gender, religion, or sexual orientation. Its detection poses significant challenges due to contextual ambiguity, cultural variance, & linguistic nuance. Artificial intelligence, particularly natural language processing & deep learning models such as CNNs, RNNs, & transformer-based architectures like BERT, have emerged as a critical tool for automating hate speech detection. Word embeddings, contextual modelling, and hybrid detection frameworks that combine human oversight with algorithmic scalability are being actively developed to improve performance. Ethical issues arise around freedom of speech, data privacy, and algorithmic bias. The review highlights techniques for detection, intervention strategies, such as Reflective User Interfaces and content flagging systems, which aim to encourage positive digital behaviour while minimizing harm.

DOI: 10.4018/979-8-3373-3063-1.ch002

1. INTRODUCTION

In the digital age, social media platforms are major sources of hate speech (HS), targeting individuals or groups based on characteristics like race, religion, or gender. The anonymity of these platforms accelerates the spread of HS, making regulation challenging. Despite manual moderation, the growing scale of cyberspace demands more efficient solutions. Advances in natural language processing (NLP) and machine learning (ML) have improved automated HS detection. Competitions like SemEval and GermEval have pushed the development of more effective detection models, using ML methods like Naive Bayes, CNNs (Computational Neural Network), LSTMs (Long Short-Term Memory), and BERT (Bidirectional Encoder Representations from Transformers). However, these systems still struggle with the linguistic and contextual diversity of HS (Jahan & Oussalah, 2023). There can be violence or societal division as a result of spread of Hate Speech on social media. Neural networks, including CNNs, RNNs, and Transformer-based models like BERT and GPT, show promise, as they can process large datasets and adapt to evolving online language. Pre-trained models like BERT excel in capturing both semantic and syntactic features, but challenges remain, such as distinguishing harmful content from legitimate discourse. These technologies must be used responsibly to ensure fairness (Velickovic & Elbassuoni, 2024). The subjective nature of HS detection introduces biases. Human annotators' socio-demographic factors influence how they label offensive content, which can be mirrored by automated systems. Biases in human annotations and large language models (LLMs), like persona-based models, remain critical issues (Giorgi et al., 2024). Regulatory responses like the UK's "Online Harms White Paper" aim to curb online harms but face criticism for misunderstanding the complexities of HS detection. While AI could improve detection, concerns over privacy, bias, and accountability persist (Cortiz & Zubiaga, 2021). As social media grows, HS increases, making automated detection vital. The challenge lies in the complexity of HS, with abbreviations, coded language, and varying expressions across cultures and languages. Various ML models have been developed, but specialized types of HS (e.g., religious, racial) remain hard to detect (Saleh et al., 2023). Generative AI, like GANs (Generative Adversarial Networks) and GPT (Generative Pre-trained Transformer), aids in hate speech detection, especially for low-resource languages, but faces challenges due to cultural and linguistic diversity. Traditional moderation systems are inadequate as content volume grows. Semi-supervised learning (SS-Learning) and GANs are being used to generate data for underrepresented languages. The SS-GAN-PLM (Pre-trained Language Model) framework, using models like mBERT, has been effective for multilingual HS detection (Mnassri et al., 2024). AI's role in content creation raises ethical concerns, particularly in handling sensitive topics like school shootings and