Chapter 14
Recent Trends on Artificial Intelligence in Automated Hate Speech Detection

Nishant Goyal
https://orcid.org/0009-0007-6199-217X
VIT Bhopal University, India

Aarul Kumar
https://orcid.org/0009-0007-2477-523X
VIT Bhopal University, India

Aarushi Chaddha
https://orcid.org/0009-0000-9743-0717
VIT Bhopal University, India

D. Lakshmi
https://orcid.org/0000-0003-4018-1208
VIT Bhopal University, India

ABSTRACT

This study investigates the performance of AI in detecting HS in diverse cultural and contextual settings. Existing AI models, trained primarily on English datasets, struggle with regional dialects, idiomatic phrases, and cultural nuances. A systematic review of NLP techniques, including traditional methods (n-grams, Bag of Words) and advanced architectures (BERT, GPT, RoBERTa, CNNs, LSTMs), evaluates their effectiveness. Multilingual models like mBERT and XLM-R are assessed for low-resource scenarios while emerging trends like multimodal learning (CLIP)

DOI: 10.4018/979-8-3693-9904-0.ch014

and adversarial training (GANs) are explored for robustness. Challenges such as data bias, false positives, and cultural insensitivity are addressed through contextual embeddings, data augmentation, and Pragmatics-oriented NLP. Metrics like precision, recall, and F1-score reveal significant accuracy drops in non-English contexts. The study emphasizes culturally aware datasets, Explainable AI (LIME, SHAP), and hybrid AI-human moderation to ensure ethical, inclusive online spaces.

1. INTRODUCTION

1.1 Background and Motivation

Hate speech (HS) has been defined as speech inciting violence, discrimination, or hostility against others because of their characteristics such as race, religion, gender, sexual orientation, or nationality. HS has long proved effective for tearing people apart and creating conflict. It has found fertile ground with the developments in social media and other internet-based technologies, whereby HS easily gains global dimensions and escalates the struggle of societies. However, new developments in communication technologies have amplified the level and intensity of HS so that it now overshadows any other time in history. HS in a physical space has always been a problem, but its digital version is particularly worrisome because of digital platforms' speed, anonymity, and scalability (Liu et al., 2023).

The social media includes some that are ever present like Facebook, Twitter, YouTube, and others recently appearing on the scene, such as TikTok. All of them have opened up ways for people to form online communities, engage in debates, and express opinions. However, the same platforms make it very easy to share toxic and more harmful content without the person having an immediate risk of it causing any backlash. According to Liu et al., HS fuels online polarization, intensifies social division, and leads to offline violence. In response to these challenges, there have been efforts to mitigate HS through the employment of technologies that can detect HS, especially AI-based detection systems.

HS detection stands to be revolutionized by AI-based detection systems that advance scalable, efficient, and automated approaches for the identification of harmful content. The old ways of manual moderation and keyword filtering are not equipped to handle the enormous volumes generated daily on social media. Besides, AI models, especially the NLP systems, have early promises that show much promise in the ability to detect and classify instances of offensive language usage across various linguistic and cultural settings. However, despite all of these recent achievements, massive issues persist, ranging from cultural sensitivity and