Chapter 3
Harnessing Ethical AI: Measures of Redressing Hate Speech in the Social Media Context

Sheikh Inam Ul Mansoor
https://orcid.org/0000-0001-8636-4769
Dayananda Sagar University, India

Showkat Ahamd Wani
https://orcid.org/0009-0009-0240-9073
Alliance University, Bangalore, India

ABSTRACT

As communication technology advanced Social media came in as means and platform of deliberation and connection wherein at the same time it opened doors for hate speech and toxic content. In this chapter, the author discusses moderation of ethical AI and hate speech in today's social media context that is increasingly becoming more polarized. Hate speech is not only an issue of some concern to users of technology but it is also a global concern in the current society especially since it affects social harmony, human rights, and freedom of speech and press thus the need to have an anti-hate speech technology with strong ethical backing. This Chapter focuses on the concept of Hate Speech from the legal, ethical and social aspects within and its effects on society and individuals. It considers how hate speech circulates, how it works by analysing sophisticated methods that include, the functions and procedure performed by algorithms, processes such as echo chambers, and the principles of things going viral.

DOI: 10.4018/979-8-3693-9904-0.ch003

1. INTRODUCTION

The social media have in the course of the last few decades became the most effective means of communication, interaction and sharing information. Social Media platforms such as Facebook, twitter, Instagram, and You tube has insisted or provided ways through which individuals or group of people can reach so many people within no time. With an Internet connection it has been made possible by social media that everyone can post, share or comment on any issue including International affairs. The change has been dominated by enlightening consequences such as: Encouragement of free speech, Encouragement of diversity of thought, Empowerment of minority opinions. But negative trends have created on social media some of which include hatred speeches, fake news, and social division. People tend to produce aggressive content on social networks thus extended the realm of hate speech to digital world. (Di Domenico et al., 1988) Hate speech mean any spoken or written words that incite violence or prejudice against people or organization based on the colours, religion or sex or color or ethnicity, sexual orientation or gender or anything like that, and it has stimulated highly over the last couple of years in these platforms. (Matamoros-Fernández & Farkas, 2021) Social media as its characteristic has become the middleware in spreading such content. By its design it promotes the sharing, encouraging a greater number of people to get engaged and share and very often the same content is inflammatory or at least sensational. (Shahbaznezhad et al., 2021) Also, social media plays a role of echo chambers, where algorithms bring like-minded individuals or users into' one space, therefore creating a space where hate speech could also be encouraged. (Cinelli et al., 2021)

The use of hate speech has some serious effect on society. In fact, social media may have communication opportunities yet threats may exist such as "radicalisation and political instability and violence." It has been linked to actual violence like few shooting incidents and ethnic conflicts in the real world. (Singh et al., 2024) It worsens the segregation of societies and brings rivalry between users of the two groups. Besides, hate speech is dangerous to the victims as they suffer psychological and negative consequences for the rest of their lives. However, given these great risks, moderating hate speech on social media has remains inconsequential and controversial because of the problems of moderating content from such big platforms. (Matamoros-Fernández & Farkas, 2021)

Social media platforms have been for a long time in the receive end of criticism over how they contain some of these negative posts. Although all the platforms have rules against hate speech, the scale at which people post content makes it almost impossible to moderate such content manually. It is there where artificial intelligence (AI) has rose as a possible solution. (Wilson & Land, 2021) Machine learning and NLP have been used in a large scale to detect and mitigate cases of