Chapter 12
Deep Learning and NLP Methods for Automated Hate Speech Detection Across Human and Machine-Generated Content

Anushree Kangoo
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

Ahmed Antwi-Boampong
Department of Information Technology, Ghana Communication Technology University, Accra, Ghana

ABSTRACT

Hate speech is a widespread and ever-changing issue that poses a serious threat to the safety and dignity of individuals and communities around the globe. This chapter offers a thorough look at hate speech, starting with what it is and how it can show

DOI: 10.4018/979-8-3373-3063-1.ch012

up in different forms—whether in private conversations, public settings, or online spaces. It also addresses tactics like denial and trivialization. A significant focus is given to the alarming increase of hate speech on social media platforms, breaking down key areas such as online harassment, hurtful comments, unfair treatment, and trolling. The chapter discuss challenges such as the ambiguity of language, cultural contexts, and the specific nuances of different platforms. It explores modern detection techniques, particularly highlighting the transformative impact of Natural Language Processing (NLP). The chapter examines the stages of NLP and how it can be applied to understand and identify hate speech, showcasing how linguistic and computational tools can effectively work together to tackle digital toxicity.

1. INTRODUCTION

Hate speech can be defined as speech containing or promoting hatred towards a person or group of people due to their attributes like race, ethnicity, religion, gender, nationality or sexual orientation. It manifests in numerous ways including the damage it causes, e.g. by creating a hostile environment; its content, e.g. by expressing or inciting hatred; its linguistic nature, e.g. by using slurs words; or its effects on human dignity, e.g. by demeaning a person about his or her basic social status. (Shridhara et al., 2023). Hate speech can be looked at in different ways. Content-based and intrinsic property is an example; it attempts to define what is said, or how it is said, whereas harm-based definitions are concerned with the actual effect of speech. Dignity concepts highlight both moral and social outcomes of speech that deprives certain groups of their equal status in society (Schmidt and Wiegand, 2017; Mittal et al., 2024; Choubisa et al., 2022; Kesarwani et al., 2023). Hate speech and its legal implications are listed below.

2. TYPES OF HATE SPEECH

1. **Privately expressed hate speech**: It happens when people speak privately as shown in figure 1. No man is usually answerable to be punished by the law, but this speech can be offensive as jokes are among friends, people do not want to make people hate anyone after listening to this type of speech. Uncontrolled hate speech on the other hand might affect mass media and spread a whole culture of abuse and discrimination, should it not be controlled.