Chapter 25
Comprehensive Analysis of the Artificial Intelligence Approaches for Detecting Misogynistic Mixed-Code Online Content in South Asian Countries: A Review

Sargam Yadav
https://orcid.org/0000-0001-8115-6741
Dundalk Institute of Technology, Ireland

Abhishek Kaushik
https://orcid.org/0000-0002-3329-1807
Dundalk Institute of Technology, Ireland

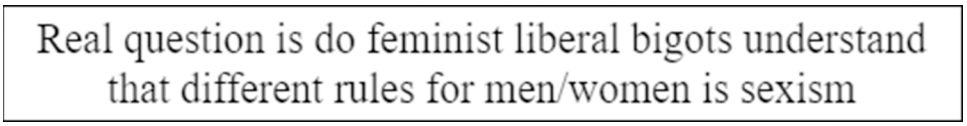
Surbhi Sharma
Institute of Dental Science, Jammu, India

ABSTRACT

The rise of social media has drastically altered several aspects of daily life and businesses. With all its advantages, the anonymity and lack of accountability social media provides encourages unsavoury individuals to spread hate. Hate targeted towards a particular group, such as women, can have a silencing effect and discourage them from participating in online discourse. In this chapter, the authors review recent studies and toolkits that attempt to tackle the issue of hate speech on online platforms using natural language processing (NLP) techniques. Challenges and shared tasks that are regularly conducted to advance the current state-of-the-art in hate speech detection in English and other under-resourced languages are also reviewed. The comprehensive survey suggests that despite the recent increase in interest in the problem of filtering online hate speech, the field is still in its infancy, specifically the problem of misogyny identification in under-resourced languages.

DOI: 10.4018/978-1-6684-8893-5.ch025

Figure 1. Example of a hateful statement



INTRODUCTION

Hate speech refers to disparaging speech that targets people based on factors such as race, religion, gender, sexual orientation, etc. Although social media has greatly boosted connectivity and communication, it has also provided a new medium for individuals who want to spread hate (Djuric et al., 2015) and misinformation, and facilitate trolling (Cheng et al., 2017) and cyber-bullying (Moreno et al., 2019). The volume of content posted online daily makes it difficult to manually moderate and remove such content. There has been an increasing amount of interest for automatic hate speech detection amongst NLP researchers. Social media platforms have become the epicenter of communication, networking, and gaining visibility. As most businesses have migrated online and political discourses are being conducted on social media, targeting hate towards a group of individuals can have a devastating impact on their continued participation. The United Nations Human Rights Council states that human rights in the offline world also apply online (*Amnesty Decoders - Troll Patrol India*, 2021), and social media companies must respect human rights where they operate (*Guiding Principles on Business and Human Rights Implementing the United Nations "Protect, Respect and Remedy" Framework*, 2011). Thus, the responsibility of ensuring a safe environment for all users falls on the platforms.

One generally agreed upon definition of hate speech is "any communication that disparages a target group of people based on some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic" (Nockleyby, 2000). Figure 1 shows an example of hate speech on the basis of gender.

According to (Feinberg & Robey, 2009), cyberbullying can be defined as "sending or posting harmful or cruel text or images using the internet or other digital communication devices. It can involve stalking, threats, harassment, impersonation, humiliation, trickery, and exclusion." Thus, cyberbullying is generalized abuse as opposed to hate speech, which is abuse directed towards a unique, non-controllable attribute of a group of people such as race and gender (*Ditch The Label 2016 | Brandwatch*, 2016). Hate speech detection is a challenging task, as it is highly contextual and subjective. The research in the field is still very nascent. Very few studies have tackled misogyny identification solely (Fersini, Rosso, et al., 2018), with the number being even lower for under-resourced and code-switched languages (Mandl et al., 2021).

India serves one of the largest user bases of social media platforms, with an approximate 450 million YouTube (*YouTube Users by Country 2022 | Statista*, 2022) users and 23 million Twitter (*Countries with Most Twitter Users 2022 | Statista*, 2022) users as of 2022. A large portion of content posted online in India combines some features of the native languages with English through code-mixing. For example, 'Hinglish', a portmanteau of English and Hindi, consists of English words, as well as Hindi words written in the Roman script rather than Devanagari script. For example, the hateful statement "Look ye politicians suvar jaise baithe rahte hain sirf money ke liye kaam karte hain. They don't care about public" in Hinglish translates to "Look these politicians are sitting like pigs and they only work for money. They don't care about public." in English (Sreelakshmi et al., 2020). Currently, Twitter's hateful conduct