INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

M.Tech Thesis Defense

Analysing and Detecting Hate Speech on TwitterProcess

ON

19th August @ 3:30 PM IST

Chhavi Jain

Advisors: Dr. Tanmoy Chakraborty

Examiners:

1. **Dr. Mukesh Mohania, Professor, IIIT Delhi, India.**
2. **Dr. Koteswar Rao Jerripothula, Assistant Prof. IIIT Delhi, India.**

Please visit https://cse.iiitd.ac.in/events-seminars/ for more details

Abstract

Due to the proliferation of harmful content and its ability to spread quickly and bring forth strong reactions, hate speech detection has emerged as a significant area of research today. However, most of the efforts regarding the variety and volume of hate speech are limited to English. Additionally, with social media platforms having a semi-automatic setup of flagging and removing explicit hate speech, it is challenging to build a dataset containing a large amount of hate. In addition to direct hate speech made by a user, there are also instances of instigation/provocation that tread that line of what is hateful or not. Such content usually contains implicit/covert language that is not easily detected as hate while still managing to convey a hateful sentiment and/or stir a hateful response from other members of the user's network. In this work, we combine to tackle three challenges -- a) developing a large-scale dataset for multilingual hate speech in India setting, b) identifying provocative content from Twitter, c) incorporating network signals for hate speech detection. We curate and annotate a dataset of 66,493 tweets focused majorly around national and international socio-political issues. In addition to the tweets, we also look at the social network to understand the influence of endogenous signals, banking on the follower-followee relations to help determine hateful content. Finally, we present HADES, a deep learning model that incorporates tweets as well as the activity history of the first level network to interpret signals from the exposure of the users. Through topic specific results, we show that the inference of signals for unknown data can be improved by including data that is topologically or contextually relevant. We also study the impact of various factors that relate to the task , finally preferring the tweet features, hashtag signals and network activity history.