Evaluating, Monitoring and Regulating the Identification of Offensive Content

Thomas Mandl, Prasenjit Majumder, Sandip Modha, Mohana Dave
Universität Hildesheim, Germany & DA-IICT, Gandhinagar, India

Hate Speech and Offensive Content Identification in Multiple Languages

- Much aggression and hate online
- Need for content moderation systems
- Supervised machine learning systems
- Scientific benchmarks necessary
- Testbeds and evaluation resources for content in multiple languages
- HASOC offers English, Hindi and German

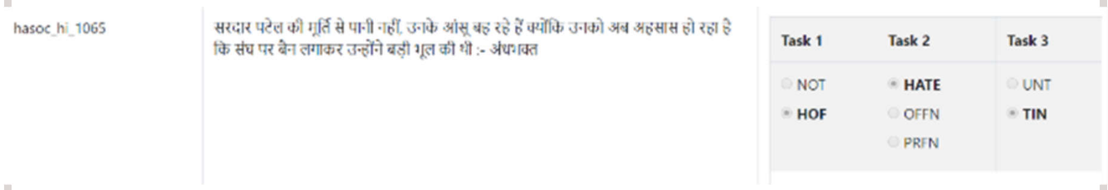
Labeling of Posts in Twitter and FB

NOT

In case you missed this! You're uniting the country alright Ireland

HOF

When are these douche bags going to realize that most of the economic gains are just continuing trends from Obama's administration? Also, these don't negate him being a racist asshole. It's like the battered wife making excuses for staying in the relationship. #FuckTrump

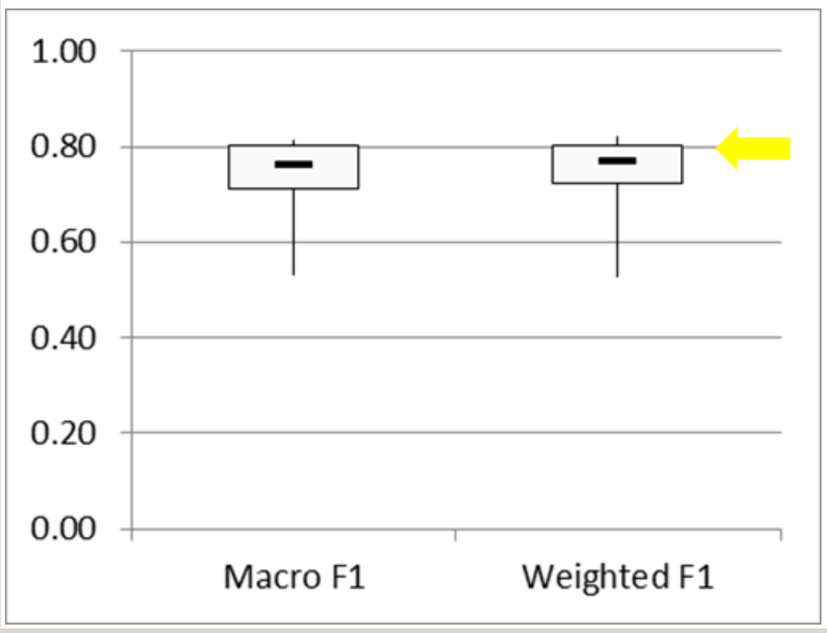


Evaluate Algorithms for Hate Speech Detection

Ensembles Runs

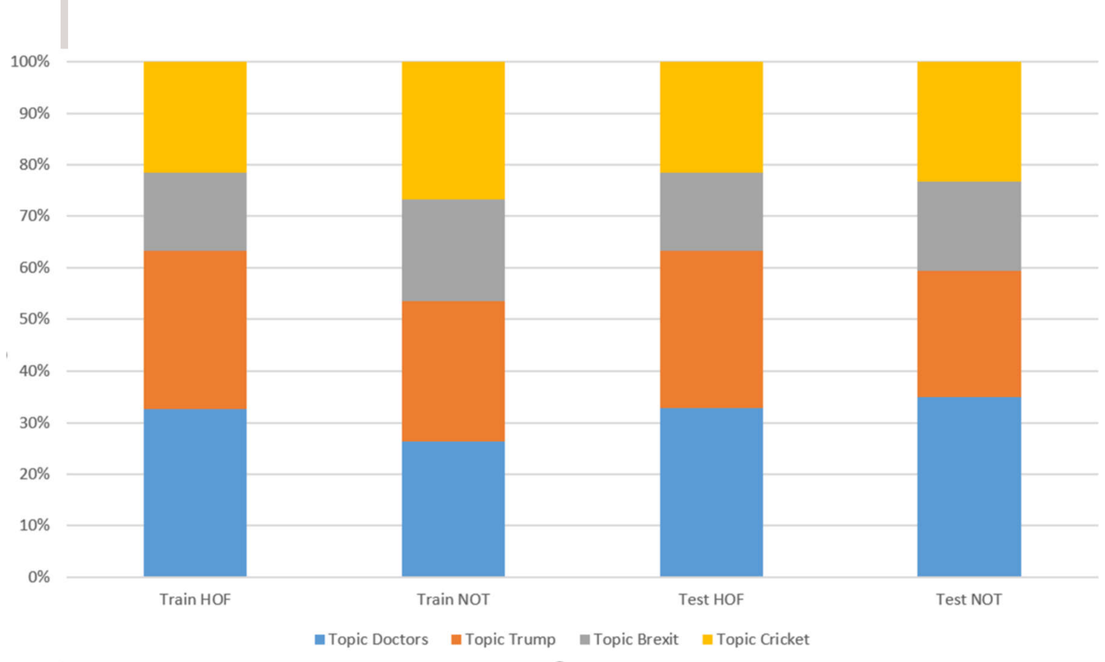
	MajVote3 Top	MajVote2 Top	MajVote3 0All	MajVote3 5All	MajVote 40All
Recall HOF	0.746	0.645	0.709	0.789	**0.853**
Precision HOF	0.613	0.707	0.639	0.587	0.507
Recall NOT	0.835	0.906	0.859	0.806	0.71
Precision NOT	0.904	0.88	0.894	0.916	**0.932**
Weight. F1	0.768	0.856	0.798	0.733	0.620

Performance for Hindi



Most often Deep Learning systems, in particular BERT perform best

Topic Model Analysis



Need for Regulation

- How can systems be monitored by authorities?
- Are there biases in the algorithms?

Future Work

- HASOC 2020 with more languages
- New ways to create a Hate Speech evaluation resource

For further information contact: mandl@uni-hildesheim.de
For participating in HASOC 2020: check hasocfire.github.io