Understanding and Interpreting the Impact of User Context in Hate Speech Detection

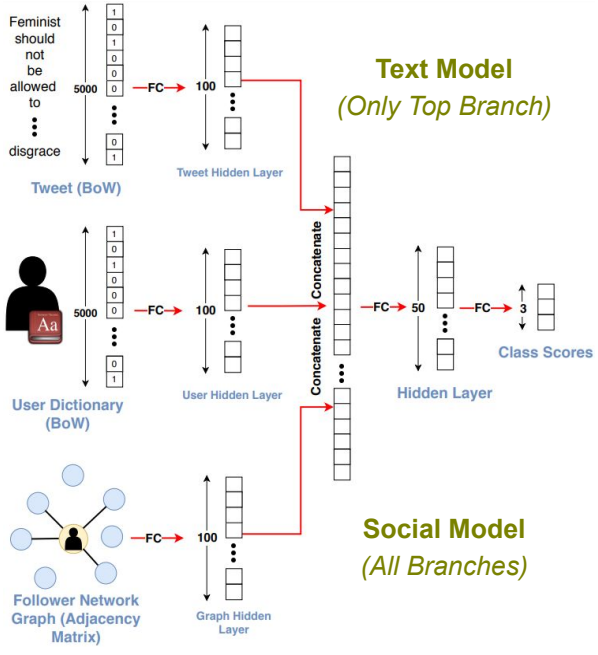
Social Computing Group, Department of Informatics
Technical University of Munich

Edoardo Mosca, Maximilian Wich, Georg Groh

Motivation and Objectives

- Detecting hate speech is **challenging** due to the complexity and variety of hate speech.
- **Leveraging user and social network data** seems promising, but their influence on the decision-making classifier is unclear.
- Our work investigates the **impact of including user and network data** into hate speech detection methods, **beyond detection performance.**

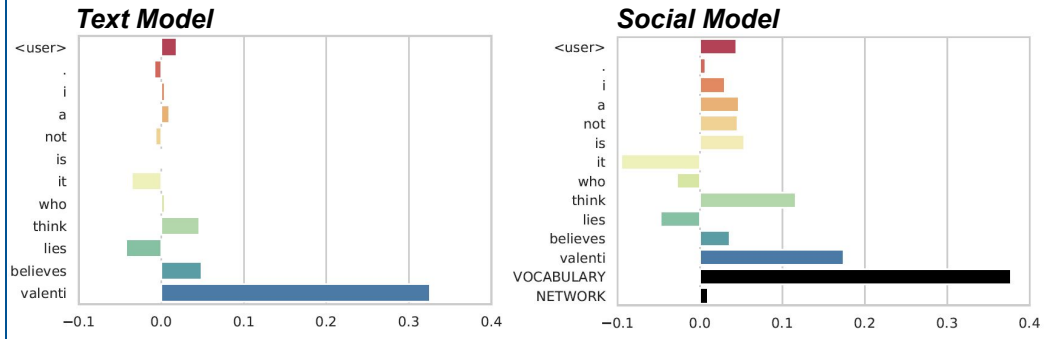
Analyzed Detectors



Methodology

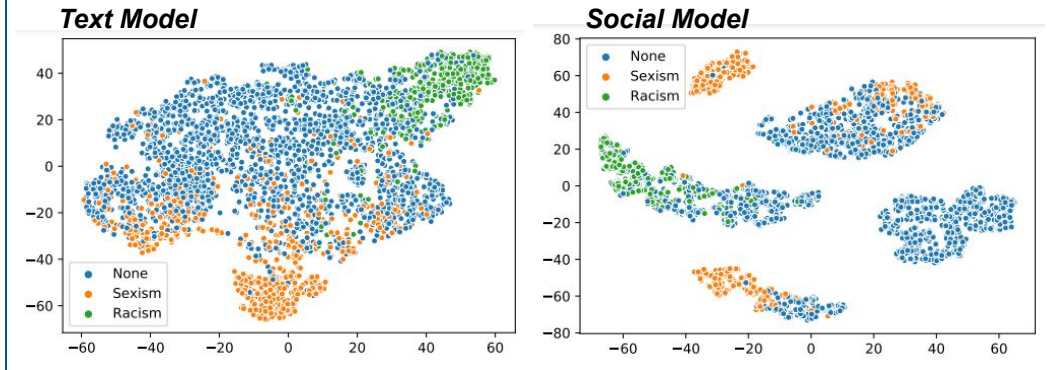
We combine **explainable Artificial Intelligence (XAI)** techniques to **compare our text- and social models.** Models only differ on the usage of user and context features.

Shapley-Values Analysis: Contribution of Each Feature



User Features play a role in the decision. *They are the reason for improvement.*

Learned Feature Space Exploration



User information used to cluster samples *That is WHY the social model is better.*
The same technique can be applied to each branch separately.

F1 Scores	Text Model	Social Model
Racism	0.711	0.735
Sexism	0.703	0.832
Neither	0.881	0.907

Performance on Waseem & Hovy. *The social model outperforms (by **4.3%**) the text model. Weaker results obtained on* Davidson *(**1%**).*

Further Experiments

- A **novel tweet can be projected** onto the feature space to see how model perceives it.
- Both techniques combined with artificially crafted tweets shows **how the model reacts** to different hate targets and message authors. This works as a powerful bias detector.

Takeaways

- Performance is not enough: compare using XAI
- Adding **user and social context** to hate speech detection models is the reason for **performance gains.**
- Model's learned features space illustrates how such features are leveraged for detection.
- Models incorporating user **features suffer less** from **bias in the text.**
- Those same models contain a **new type of bias** that originates from adding user information.