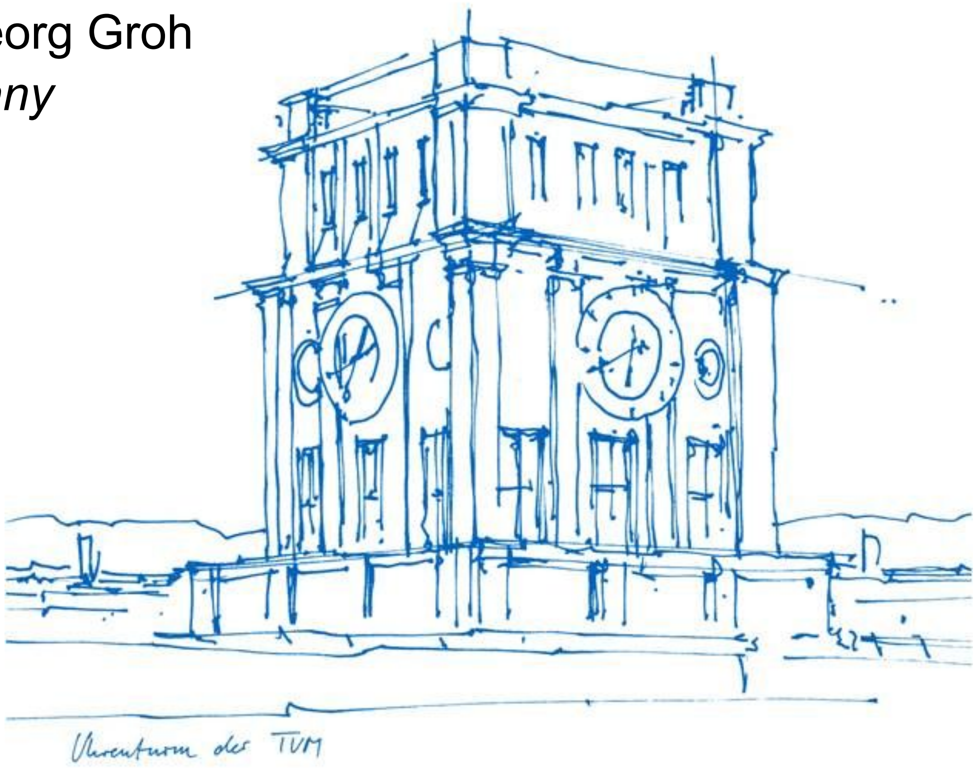
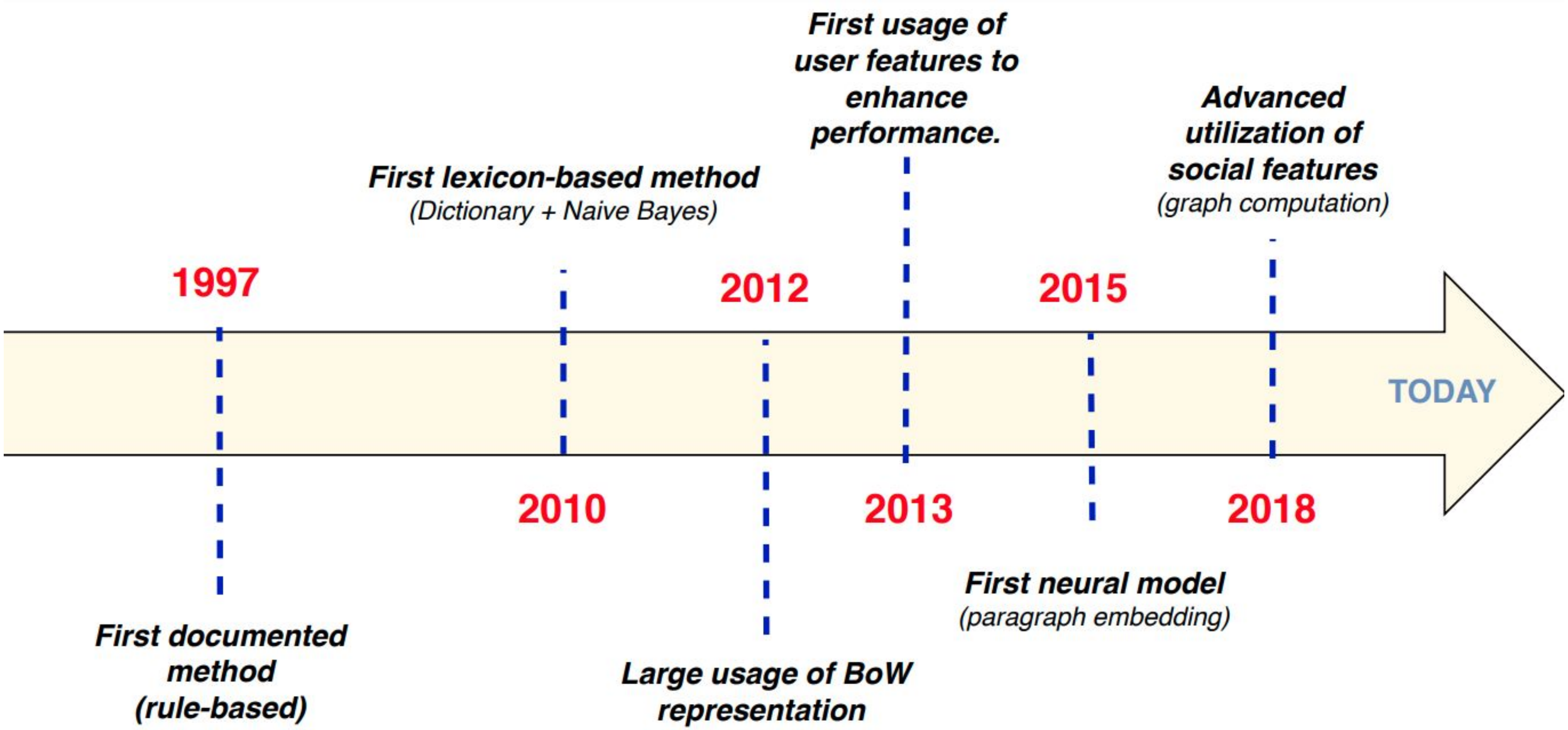
Understanding and Interpreting the Impact of User Context in Hate Speech Detection

Edoardo Mosca, Maximilian Wich, Georg Groh
Technical University of Munich, Germany



SocialNLP @ NAACL 2021
6-11th June

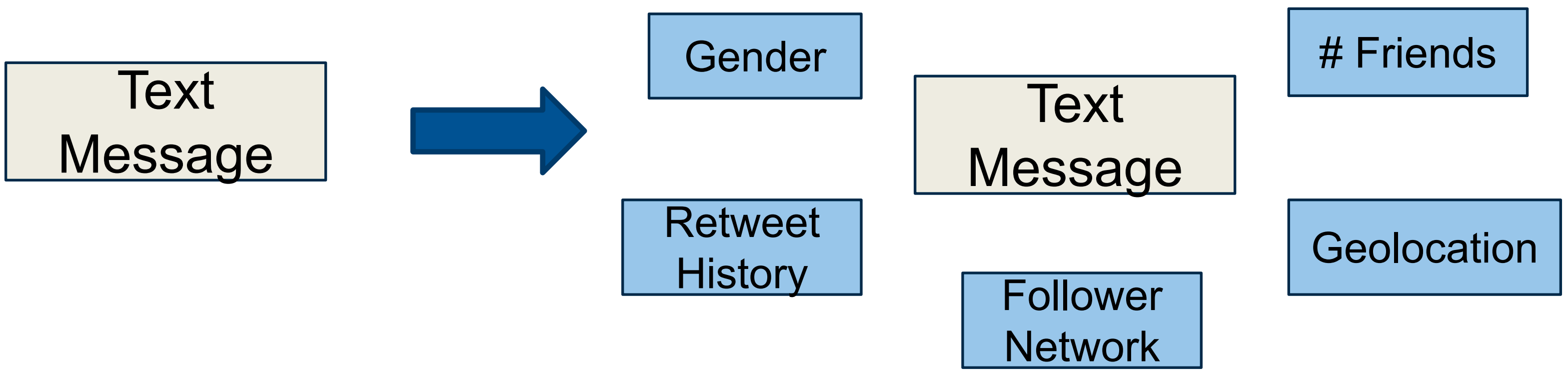
Current State in Hate Speech Detection



Detection models have evolved over time. The current SOTA, substantially relying on DNNs, still faces **limitations in accuracy and interpretability**.

Using Social Features

Several works leverage **user context features** found on social media.



Better Performance!

This Work

- What is the **impact** of including user features?
- Unlike previous work, model comparison **beyond performance** metrics.

Datasets

We test on two popular twitter hate speech detection benchmarks:

Waseem & Hovy [1]

Class	# Tweets
Racism	3,378
Sexism	1,970
Neither	11,501
Authors	2,024
Connections	9,955

Original Benchmarks

User Context

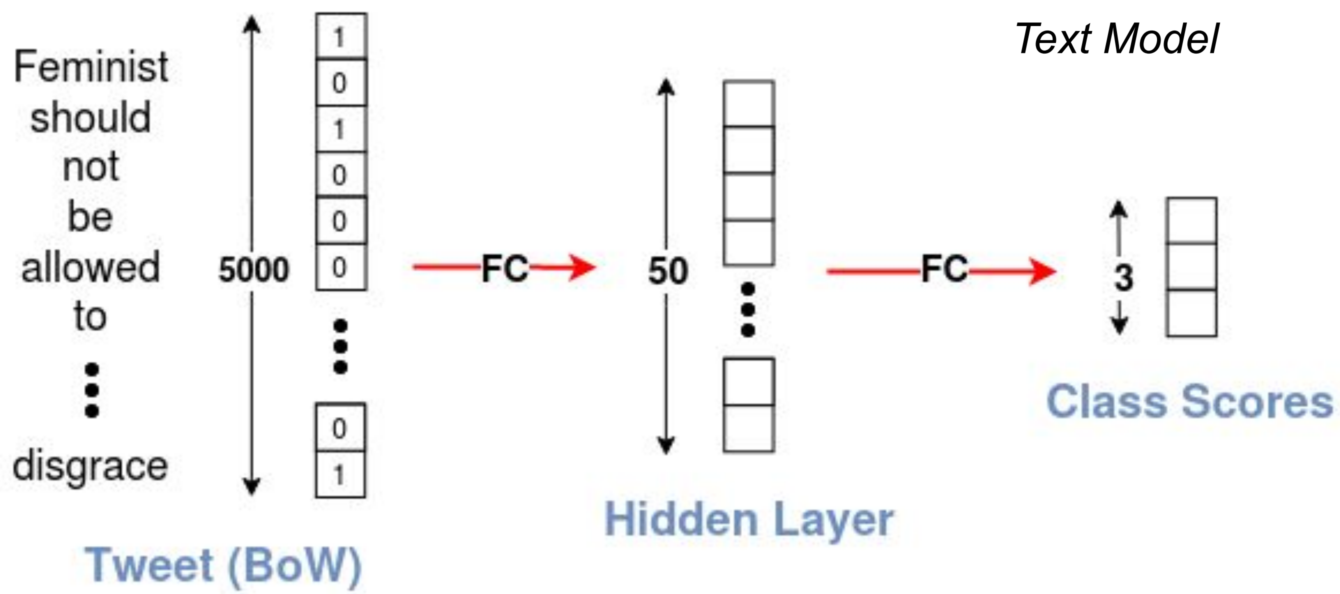
Davidson [2]

Class	# Tweets
Hateful	1,430
Offensive	19,190
Neither	4,163
Authors	6,725
Connections	19,597

[1] Z. Waseem and D. Hovy. 2016. *Hateful symbols or hateful people? predictive features for hate speech detection on twitter*
[2] T. Davidson et al. 2017*. Automated hate speech detection and the problem of offensive language*

Utilized Models

Compare two models: one based only on text **(text model)**, and one that also leverages context **(social model).**

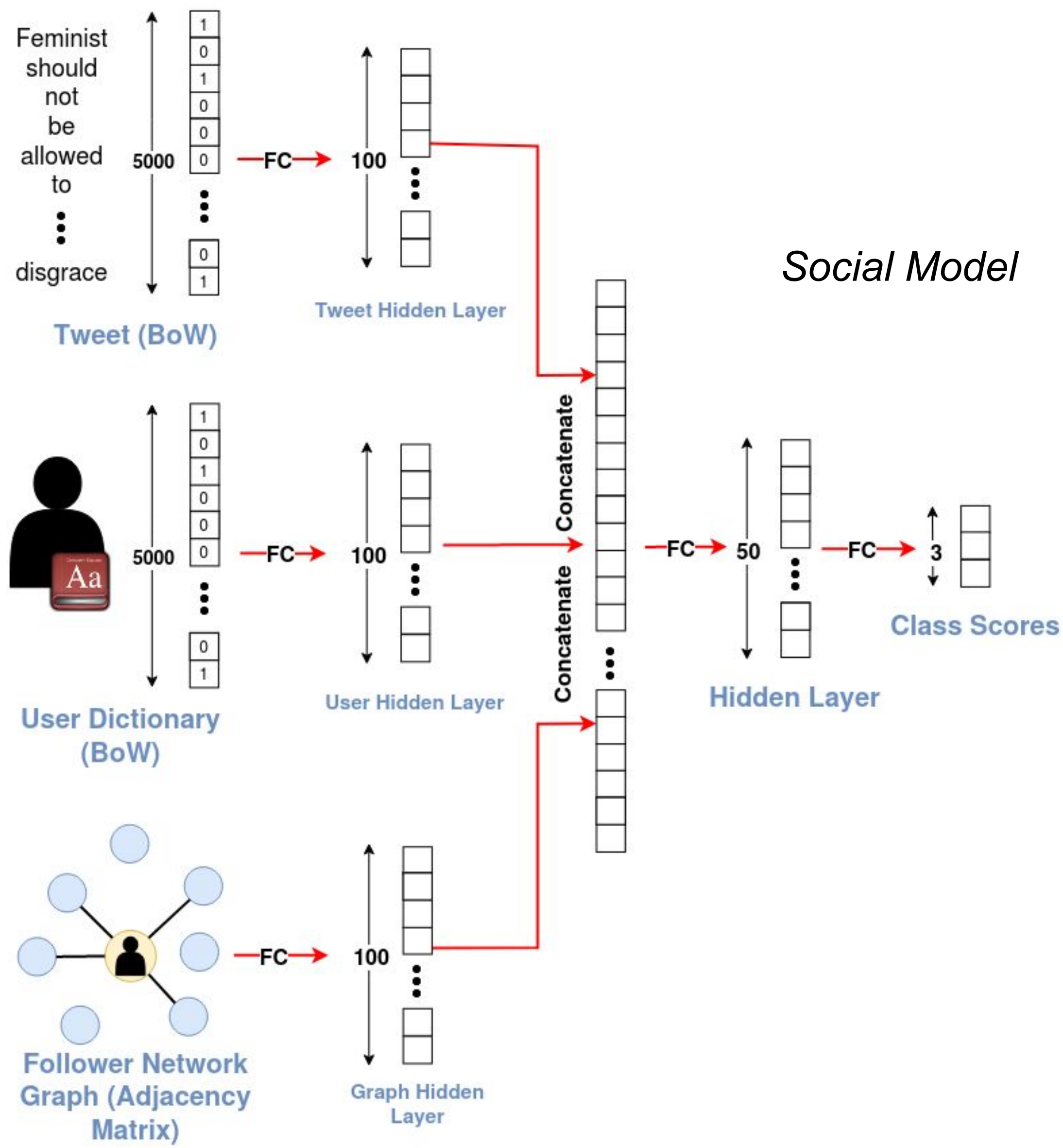


Utilized Models

Same input as text model.

Merge of all the tweets of the user. Represents overall writing style/content.

Two users are connected if one of them follows the other (very sparse graph).



Comparison: Performance

F1- Scores on
Waseem & Hovy

Class	Text Model	Social Model
Racism	71.1	73.5
Sexism	70.3	83.2
Neither	88.1	90.7
Overall	82.9	87.2

Considerable improvement (+4,3%), visible in every single class.

F1-Scores on
Davidson

Class	Text Model	Social Model
Hate	15.4	34.7
Offensive	93.9	93.9
Neither	80.9	81.5
Overall	87.6	88.6

Minor improvement (+1%), mostly on the hate class.

- Major impact on Waseem & Hovy, we focus on this dataset.
- Is context actually improving the model or is it only due to the architecture?

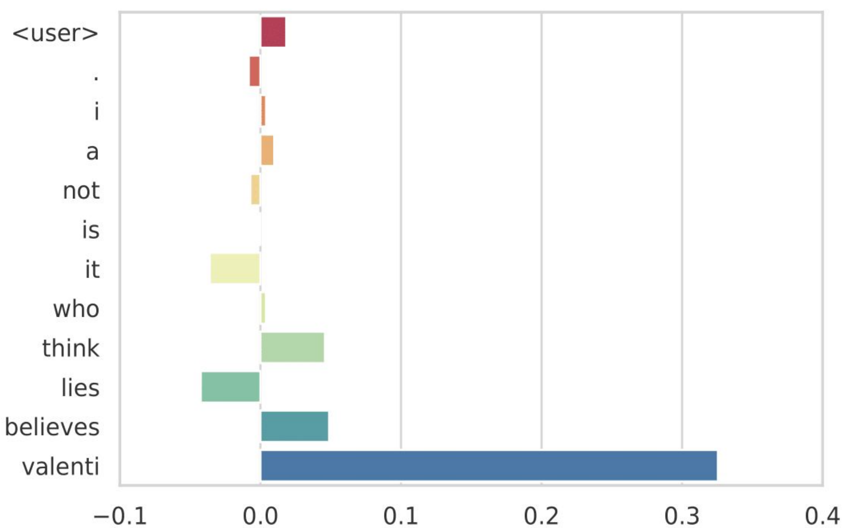
Comparison: Shapley Values Approximation

Tweet: *”<user> I think Arquette is a dummy who believes it. Not a Valenti who knowingly lies.”*

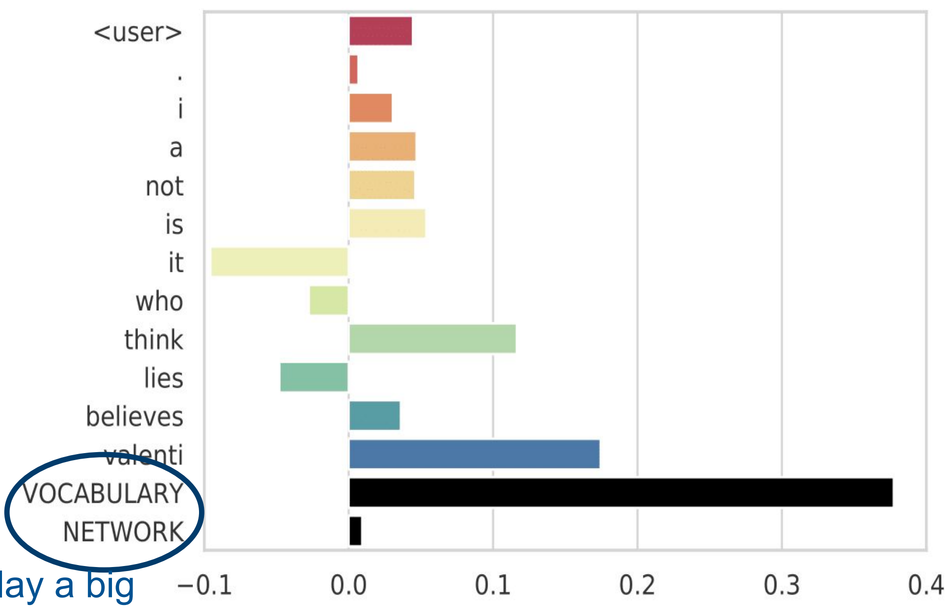
Predicted as sexist

Contribution (Shapley value) of each feature to the sexist class.

Text Model



Social Model

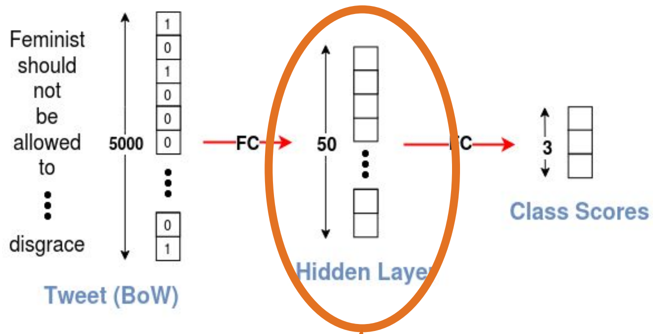


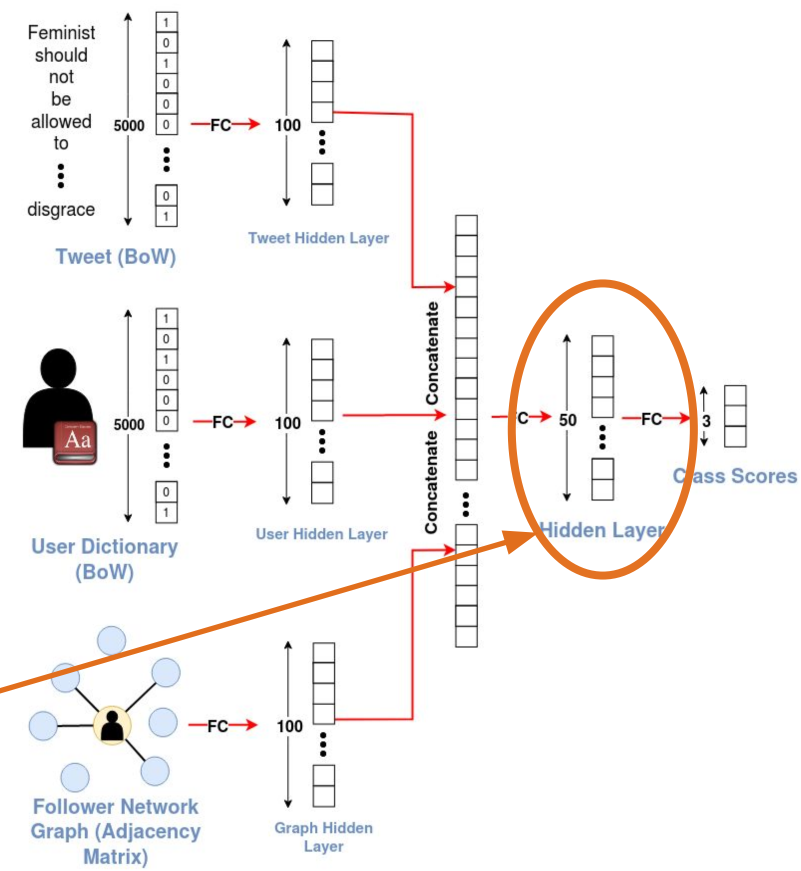
Context can play a big role in the decision.

User context is the reason for performance gains, but **why**?

Comparison: Feature Space Analysis

We can visualize the feature space learned by both models.

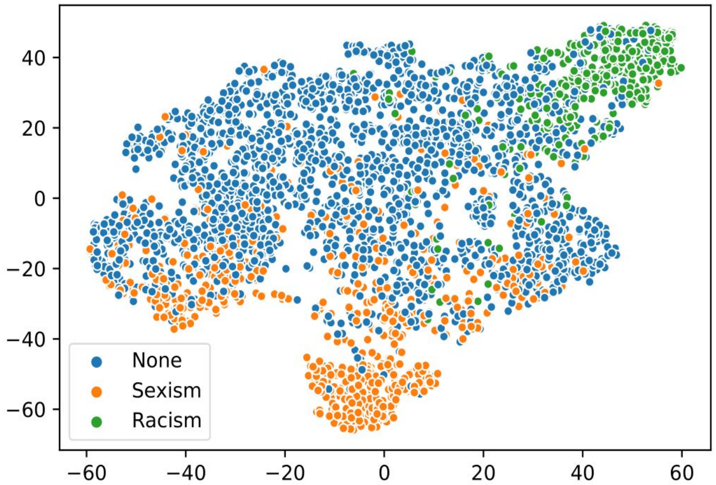




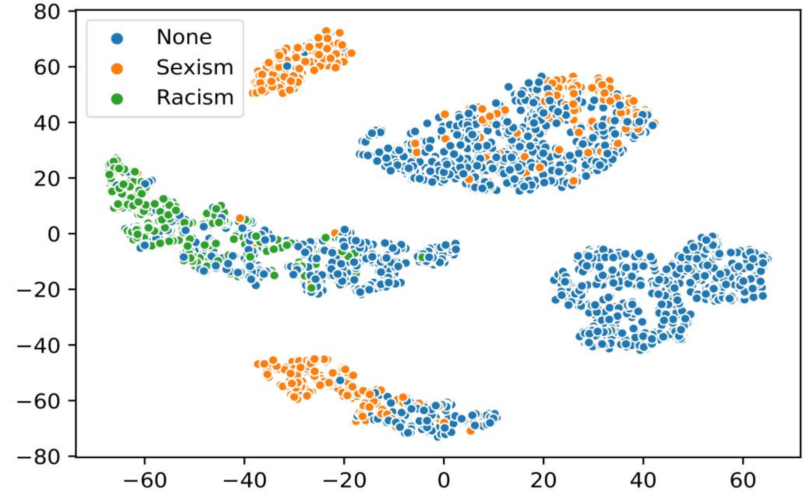
1. Throw the last layer and forward-pass all samples
2. t-SNE to reduce the 50-dimensional outcome and visualize.

Comparison: Feature Space Analysis

Text Model



Social Model

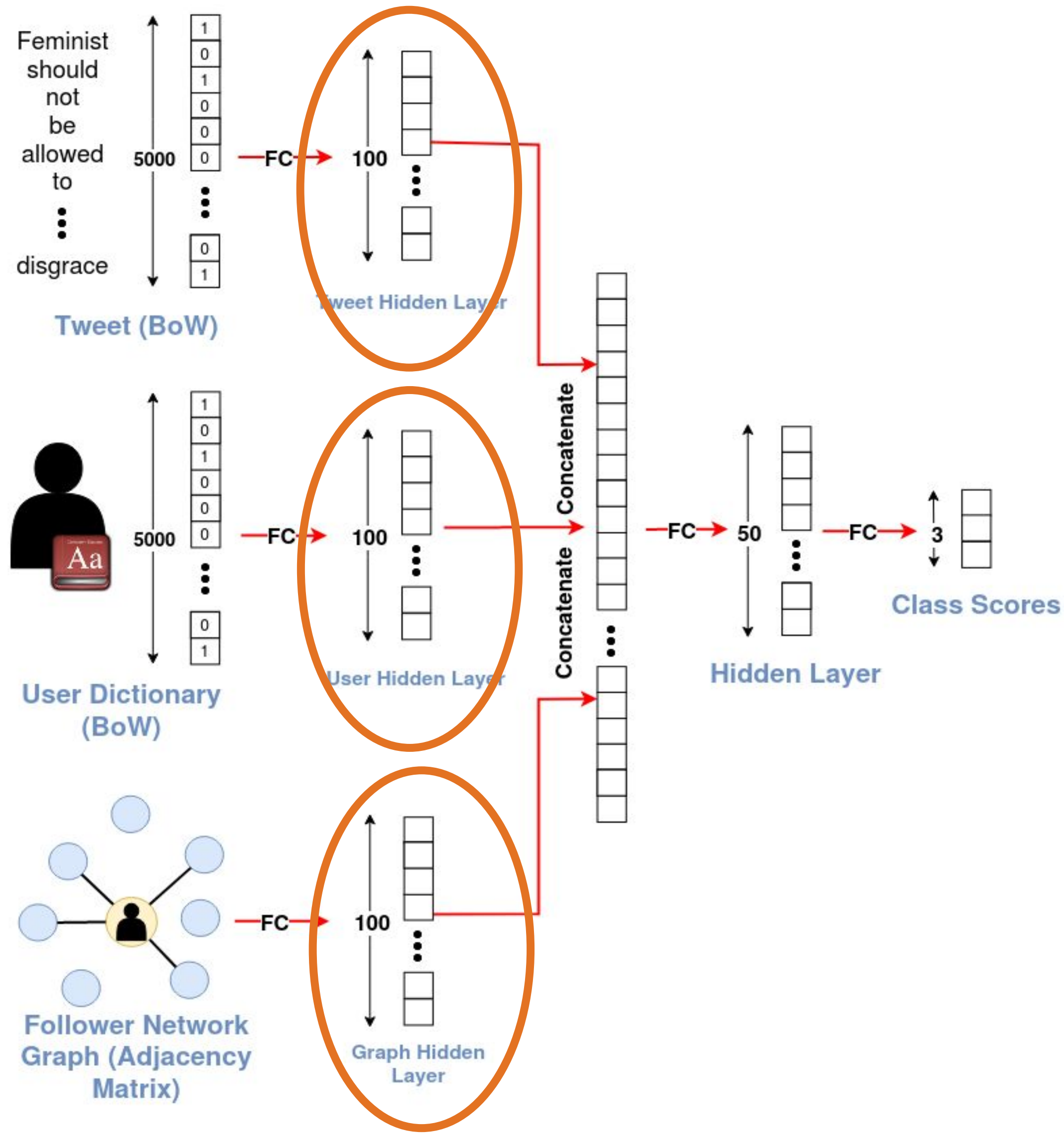


Tweets appear all in one single cluster. Racism is concentrated in one area, sexism is more sparse and hidden among normal tweets.

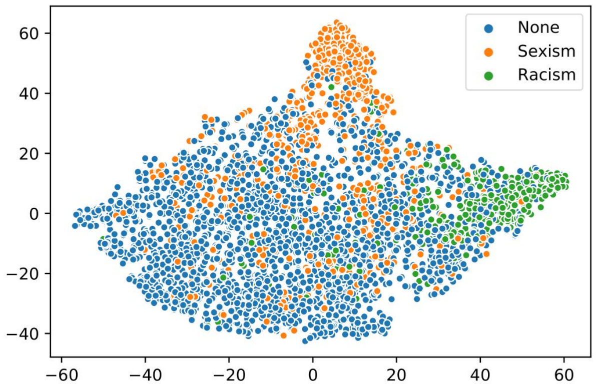
Tweets are separated in clusters. Racism is only found in one of them. Sexism, once again, shows a more sparse and hidden distribution.

What part of the social model is responsible? Repeat the procedure for the single branches!

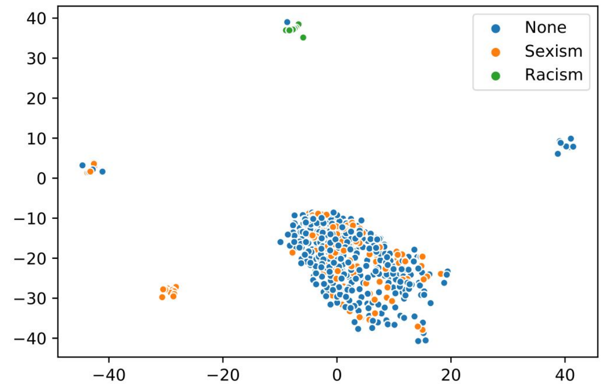
METHODOLOGY



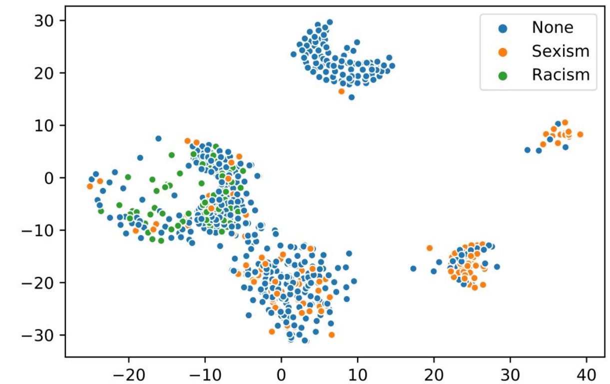
Feature Space Analysis: Social Model Branches



Tweet Branch



User Vocabulary Branch



Follower Network Branch

- The tweet branch looks similar to what learned by our text model.
- The other branches present separated clusters. Racism is always concentrated in small areas. We also observe (almost) pure clusters.
- Intuitively, being able to separate samples in clusters should be useful for classification at later layers (deciding within a small cluster is easier).
  □ That seems to be why the social model is better.

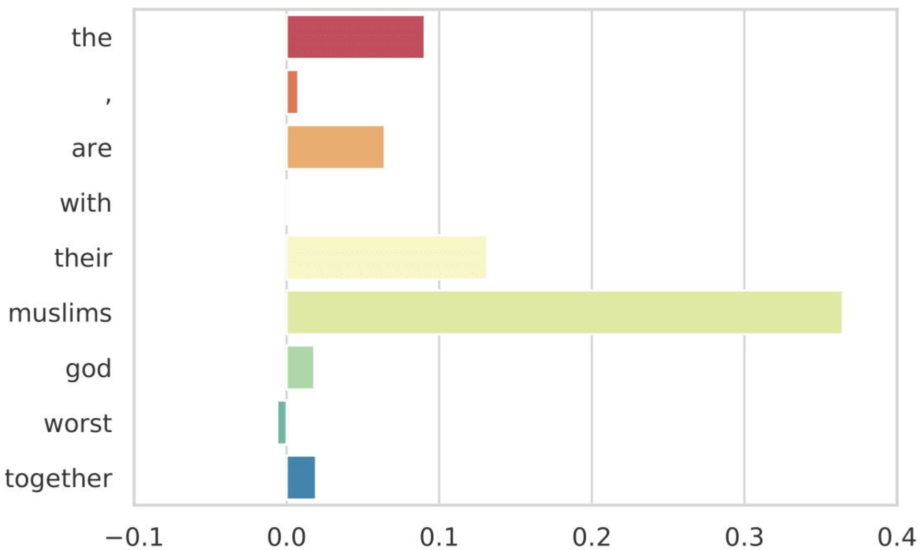
Practical Application beyond the Dataset

- If we check the models' responses to an **artificially crafted tweet**, we could also check their **behaviour in specific scenarios**.

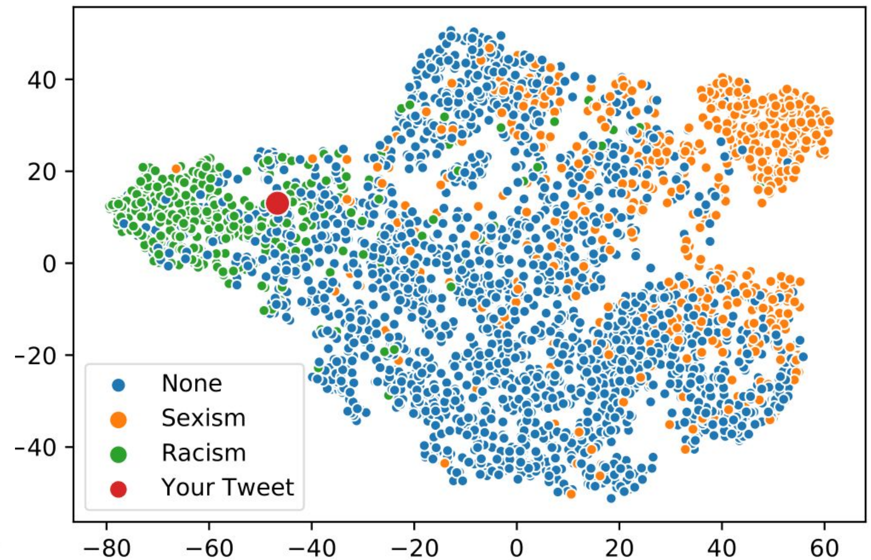
- Besides using Shapley values, we can **project where the new tweet would be positioned by the models** w.r.t. the rest of the dataset.

Artificially Crafted Tweets: Text Model 1

Artificial Tweet: *”muslims are the worst, together with their god”*
Predicted as racist (75%)



Text Model, Shapley values

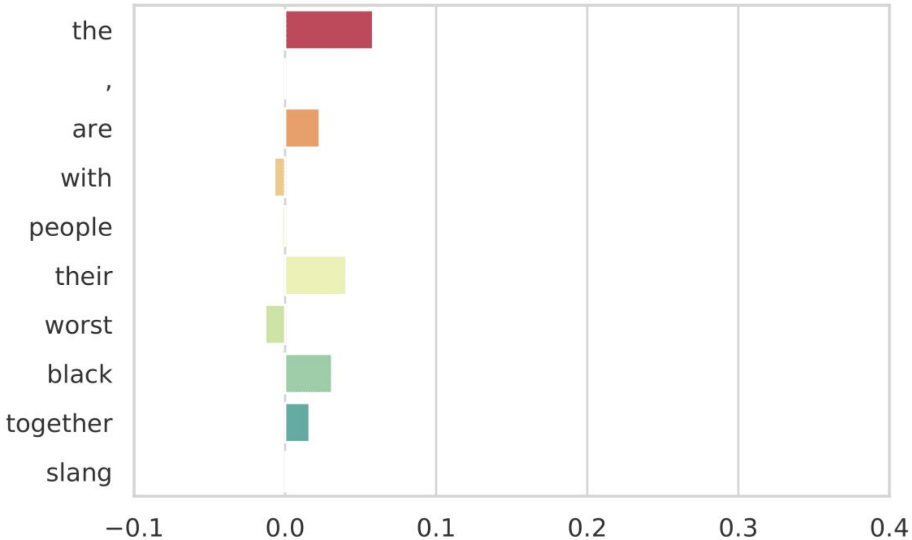


Text Model, Projection onto Feature Space

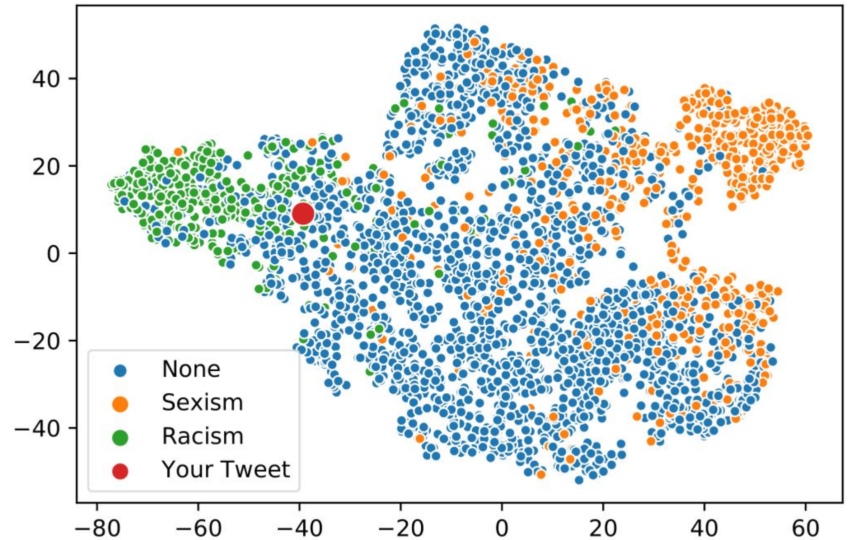
What happens if we change the target of the hate?

Artificially Crafted Tweets: Text Model 2

Artificial Tweet: *"black people are the worst, together with their slang"*
Not predicted as racist (24%, neither 73%)



Text Model, Shapley values

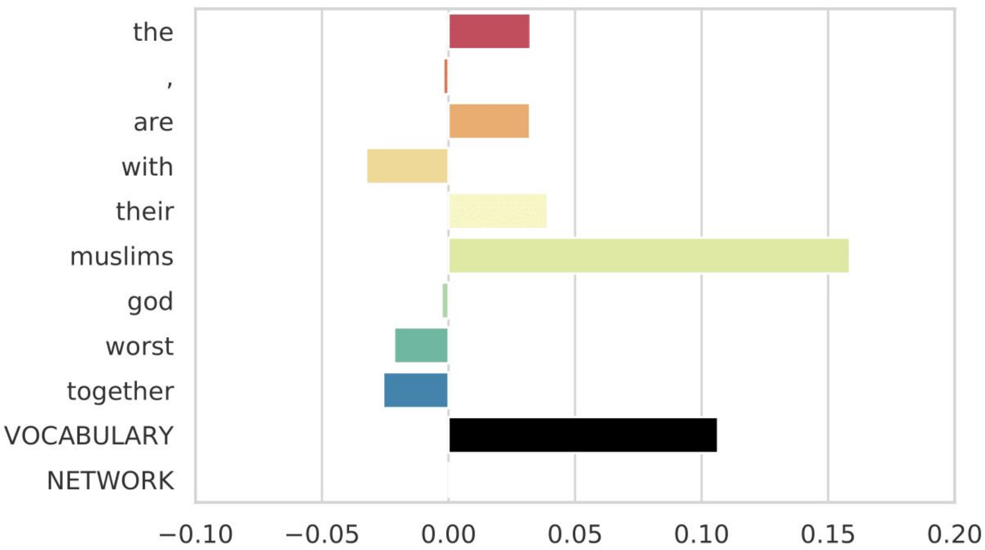


Text Model, Projection onto Feature Space

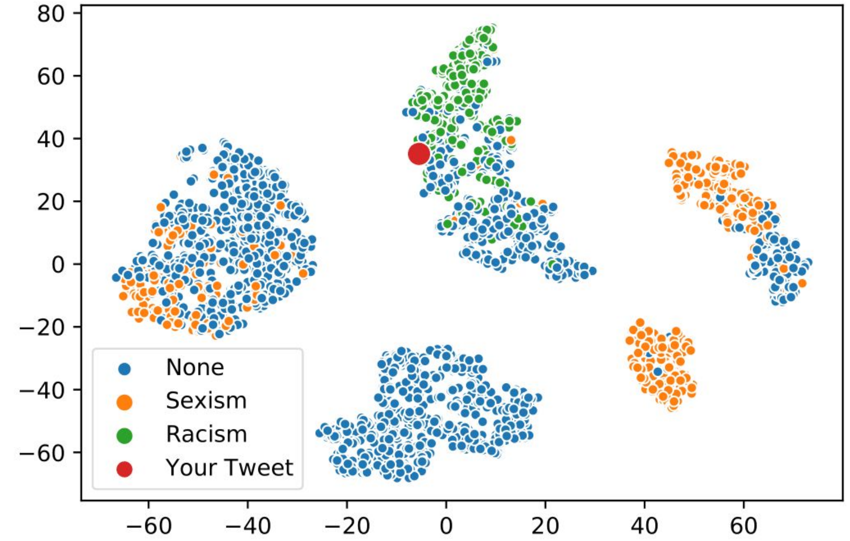
The text model suffers from bias in the text!

Artificially Crafted Tweets: Social Model 1

Artificial Tweet: *"muslims are the worst, together with their god"* User: Racist.
Predicted as racist (64%)



Social Model, Shapley values

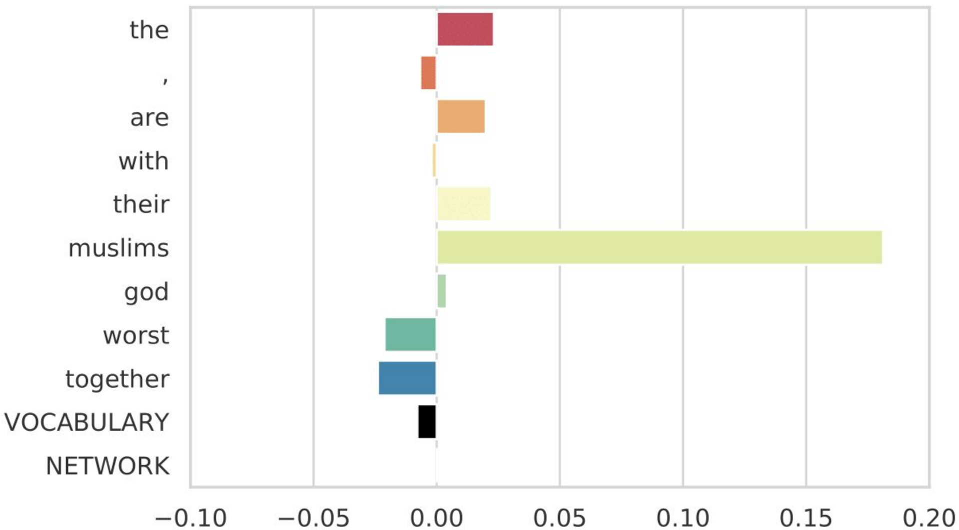


Social Model, Projection onto Feature Space

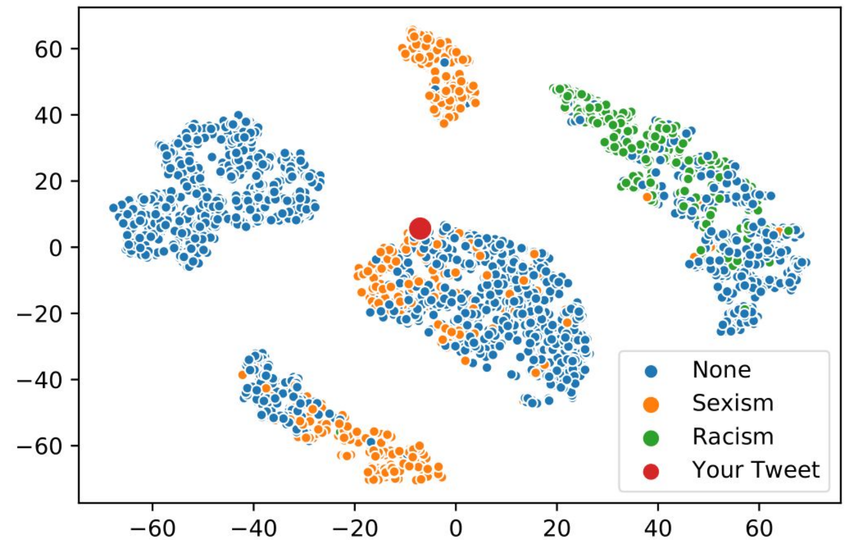
What happens if we change the tweet's author?

Artificially Crafted Tweets: Social Model 2

Artificial Tweet: *"muslims are the worst, together with their god"* User: Neither.
Not predicted as racist (19%)



Social Model, Shapley values



Social Model, Projection onto Feature Space

Even if the social model can be more resilient to bias in the text, it suffers from bias in the user context.

Conclusion and Takeaways

- Performance is not enough: compare using XAI

- Shapley values **user and social context** are the reason for **performance gains.**

- Models' feature space how such features are leveraged for detection.

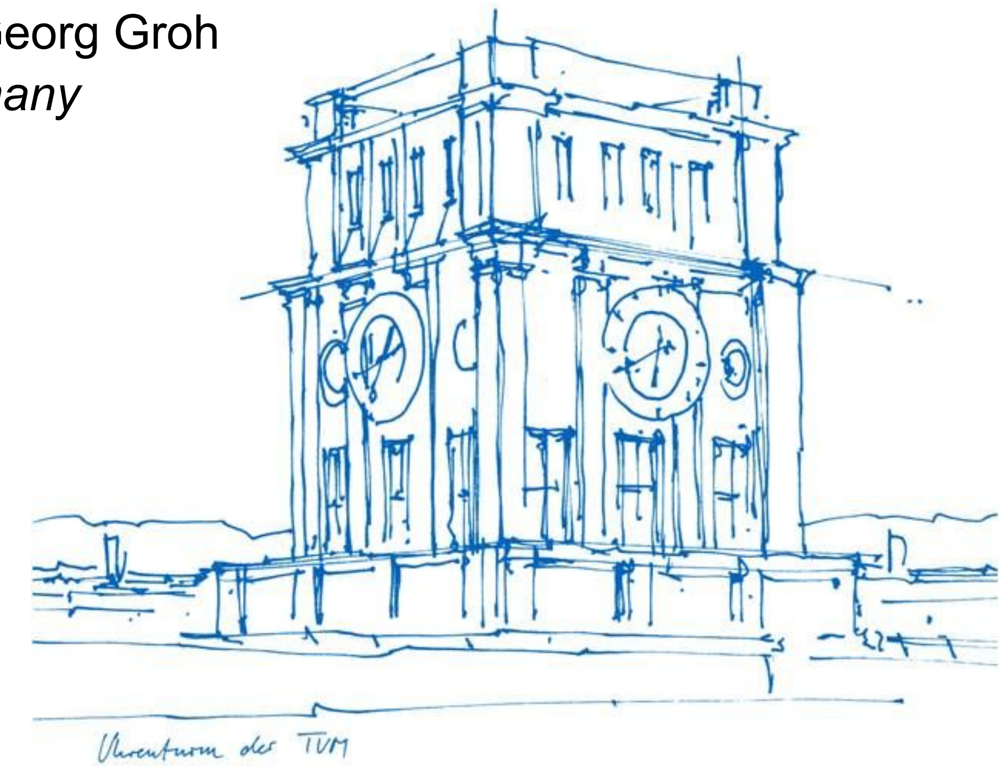
- Incorporating context **suffer less** from **bias in the text.**
  ..but **new type of bias** originating form user information.

Thanks for your Attention!

Questions?

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

Edoardo Mosca, Maximilian Wich, Georg Groh
Technical University of Munich, Germany



SocialNLP @ NAACL 2021
6-11th June