BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Freie Universität Berlin

Digging deeper

Using ‘information rich’ data to extend the error analysis of a hate speech algorithm

Paasch-Colberg, Sünje[1], van Aken, Betty[2], Strippel, Christian[1], Laugwitz, Laura[1], Löser, Alexander[2], Trebbe, Joachim[1], & Emmer, Martin[1]
([1] Freie Universität Berlin, Germany; [2] Beuth University of Applied Sciences Berlin, Germany)

In order to address hate speech in online discussions more efficiently, the field of automatic hate speech detection has expanded in recent years. The heterogeneity of hateful content, the difficulties defining hate speech, and its consistent distinction from offensive language are challenging (Davidson et al., 2017; Fortuna & Nunes, 2018). Our aim is to contribute to this research by providing:

- an error analysis of a deep learning model for hate speech detection using theory-based and manually multi-labeled data,
- insights into the ability of our model for different forms of expressions.

Data and algorithm implementation

We developed a theory-based coding scheme to account for different forms of hate speech and offensive language (Table 1). In a case study on the topic of migration, 11,263 German user comments posted in 2018, 2019 and 2020 to a sample of news sites, right-wing blogs, and social media were manually classified. Judgements of individuals or groups in the comments were identified and qualified using the coding scheme. To compensate for the high variance in our dataset, we used transfer learning and large corpora of unlabeled text to pretrain a deep learning model on the basic concepts of the German language. We then fine-tuned the model on our specific task using the training data with binary labels for hate/no hate (‘hate’ applies if at least one hate speech element has been identified).

Label and codes	Description
Target of judgement (1-7)	Social group that is judged (e.g. refugees, politicians)
Subject of judgement (1-9)	Subject of judgement (e.g. appearance, behavior, sexuality)
Hate speech elements	
Negative stereotype (0/1)	Attribution of negatively connotated characteristic, role, behavior to a group
Dehumanization (0/1)	Judgement that portrays individual/group as inferior, less than human
Physical violence (0/1)	Judgement that legitimizes, incites or threatens physical violence
Killing (0/1)	Judgement that legitimizes, incites or threatens killing
Offensive language	
Insults and slurs (0/1)	Judgement that contains swear words or insulting group names
Metaphors (0/1)	Judgement that contains derogatory metaphors or comparisons
Word creations (0/1)	Judgement that contains derogatory word creations (e.g. rapefugees)

Table 1. Coding scheme

Results and Discussion

The trained model was run on the test-set with 11,263 samples. Evaluation scores were low, especially the precision and F1-score for the ‘hate’ samples. The high information density of our data set allowed an extended error analysis to better understand which forms of comments cause the misclassifications. Results show that the model often misclassifies comments containing elements of offensive language as hate speech. The ability of the model also differs with each target group; e.g., comments mentioning migrants/refugees are often misclassified for hate speech. With respect to hate speech forms, the results show that comments dehumanizing the target or calling for violence or death are often misclassified.

Thus, the error analysis reveals bias in our model that need careful consideration. The study also confirms that the trade off between hand-coded data and balanced data is difficult to solve for a rare task such as (extreme forms of) hate speech.

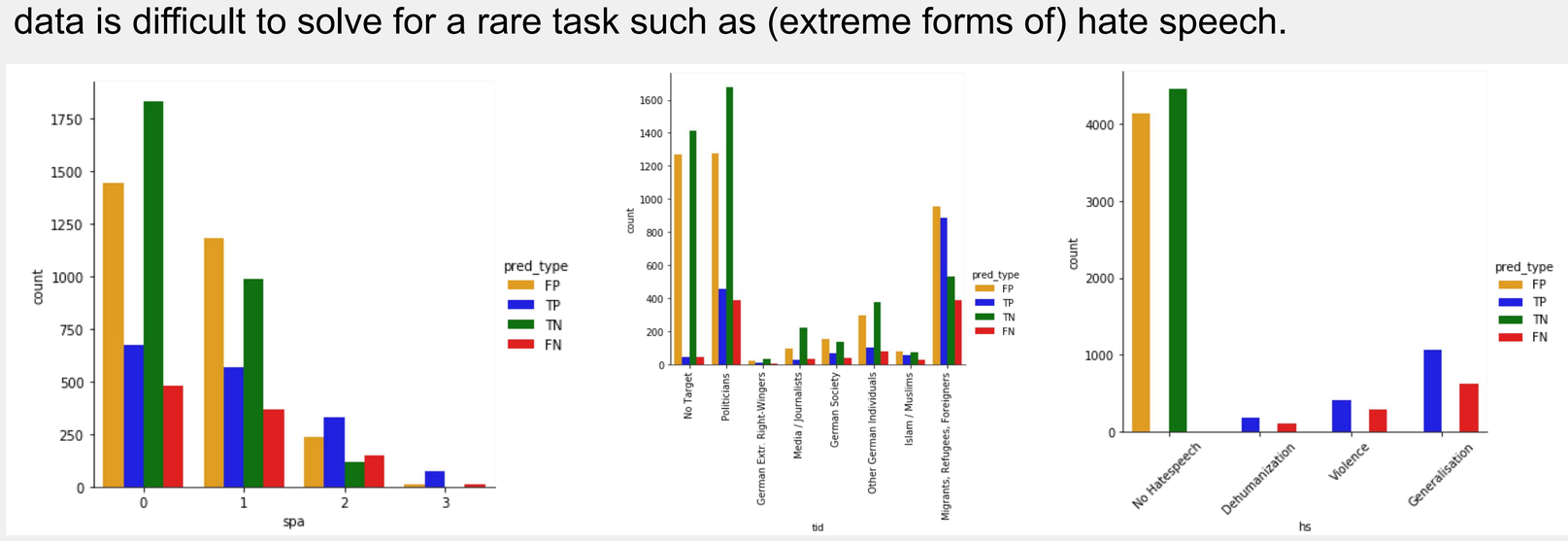

Figure 1: Number of false/true negatives and positives on different levels of verbal aggression (left), different hate target groups (middle), and different hate speech forms (right).

References:
Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. arXiv:1703.04009.
Fortuna, P, & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys 51(4):1-30. DOI: 10.1145/3232676

NOHATE

IC²S²

SPONSORED BY THE
Federal Ministry of Education and Research