Freie Universität Berlin

Modularized Hate Speech Annotation

A Human-Labeled Dataset of German User Comments on Flight and Migration

Christian Strippel, Sünje Paasch-Colberg, Laura Laugwitz, Martin Emmer & Joachim Trebbe
— Institute for Media and Communication Studies, Freie Universität Berlin

Human-labeled datasets are very valuable for the scientific community since they can be used for algorithmic classification. However, such datasets are scarce. This applies in particular to the field of analyzing 'hate speech' in user comments on news sites and social media (Fortuna & Nunes 2018). Despite the high relevance of the topic, only a few datasets are available (e.g., Waseem 2016; Wiegand et al. 2018) and many have serious limitations regarding the theoretical background of the constructs, the documentation of the labeling procedure, and data reliability (Ross et al. 2016). Low reliability results primarily from three factors:

– Labeling is often carried out by non-experts (Snow et al. 2008).
– Binary labels for complex constructs (e.g., hate/no hate) lead to problems with ambivalent cases.
– Conflicting training data when only part of a comment is hate speech.

Data collection

To address such limitations, we present a human-labeled dataset consisting of more than 9,000 German user comments on flight and migration. One portion was collected in August 2018 from various sources including 8 news sites, 7 Facebook pages, 31 YouTube channels, 3 right-wing blogs and 1 Q&A site. A second round of collection in March 2019 focused only on sources that contained a high amount of hate speech.

The collection of comments was carried out in two steps: First, the sources were searched for articles or postings on flight and migration using relevant terms. Then, the first 50 user comments on each of these articles and postings were considered for annotation.

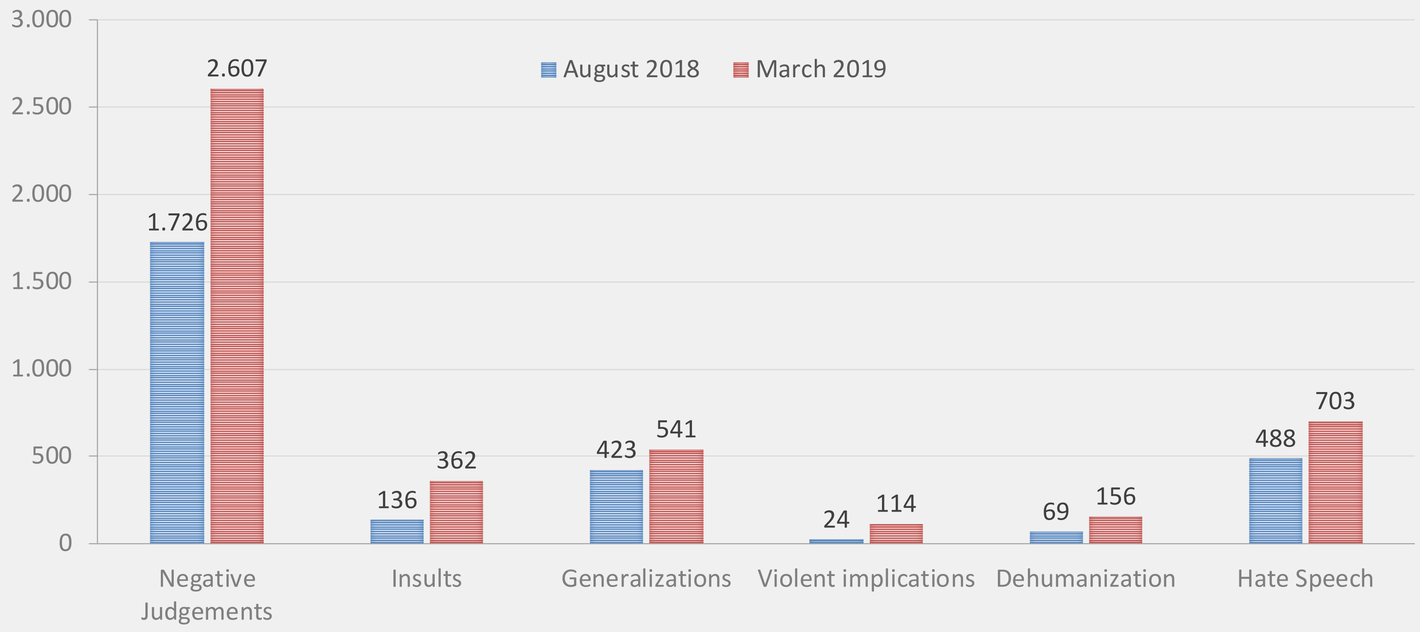
Annotation

Using a detailed theory-based manual, five students were trained to label eight categories using BRAT (Stenetorp et al. 2012). Hate speech was identified indirectly through a modularized approach:

– First, all negative judgments of individuals or groups within the comments were annotated as 'entities'.
– Second, these 'entities' were then further qualified by attributing 7 labels, including targets of judgement (e.g., refugees, migrants, politicians), subject of judgement (culture, sexuality, character/behavior), violent implications (e.g., call for violence), generalizations (by ethnicity, gender, religion etc.), as well as insults, derogatory comparisons, and forms of dehumanization.

Figure 1 gives an overview of the amount of negative judgments identified within the user comments as well as the frequency of insults, generalizations, violent implications and dehumanization associated with these judgments. Using this method, it is possible to analyze the user comments in our dataset against the background of various definitions of hate speech or offensive language. Based on a common understanding of hate speech, we considered all negative judgments that include generalizations, violent implications and/or dehumanization as hate speech. Our dataset contains a total of 1,191 of such judgments.

Figure 1. Frequency of annotated judgments and labeled categories within the dataset



NOHATE
IC²S²
SPONSORED BY THE
Federal Ministry of Education and Research