SafeNet: Challenges and Strategies in Combatting Online Hate Speech

Peter Trolle, Luca Rossi, Christian Hardmeier
Department of Computer Science, IT-University of Copenhagen

The 24-month project SafeNet, uniting 21 partners in 18 European countries, including members of the International Network against Cyber Hate (INACH), aims to combat online intolerance, racism, and xenophobia. To do this, the project seeks to report up to 20,000 cases of illegal hate speech, monitor IT companies' responsiveness, provide multilingual info sheets, and host advocacy roundtables to engage IT companies, authorities, and media to expedite the removal of illegal hate speech and foster safer online environments.

Initial training and following meetings ensured understanding of the hate speech definition and the data collection processes among all partners. Hate speech instances are collected, excluding personal identifiers, and accompanied by English translations and screenshots. The subsequent info sheets reported response times, removal rates, and the distributions of reported hate types.

With around 4000 cases reported so far, the project appears to be a success. The need for a project like this is emphasized by the fact that only 42% of reported cases were removed and 26% were left unanswered.

However, seeing the project as a way of creating a large multilingual hate speech dataset, it is important that all partners apply the same definition of hate speech so that there is coherence between all cases reported. Yet, being a large project with partners in several different countries makes this difficult. In our view, the primary issues regarding coherence are the following:

- **Definition variation**: In many cases it is not clear whether something is hate speech or not, and the project's guidelines cannot encompass all corner cases of this problem.
- **Influence of hate type**: The threshold for labelling content as hateful can be influenced by the historical marginalization of targeted groups and the frequency of targeting.
- **Geographical and cultural factors** across countries may influence definitions of hate speech and which groups are marginalized or frequently targeted.
- **Specialization bias**: Some partners focus exclusively on specific types of hate, thus neglecting other types.
- **Sampling strategy**: Frequent reporting requirements may have resulted in oversampling of prevalent hate types, neglecting rarer forms.

Looking at the types of hate reported for 6 partners, we see a pattern where partners specialize in a particular type while reporting few instances of other types (see Figure 1). We suspect that this mostly is the result of a combination of the last three items listed above.

Addressing these factors in the beginning of a multi-stakeholder project would minimize their impact on dataset coherence. This could be done by including more examples of hate speech as well as tests in the training materials so that all partners agree on what defines illegal

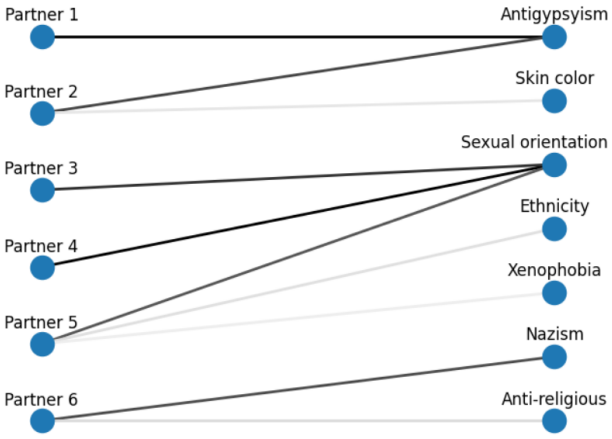

Figure 1: The types of hate accounting for 80% of all cases for 6 partners, starting with the most prevalent. Darker edges represent higher percentages.

hate speech regardless of its type, where they should look for it, and which geographic and cultural factors to consider. Inclusion of legal experts might also help define illegal language.

Moving forward, SafeNet will continue to monitor IT companies, working towards a safer online environment.