Abstract 11

Opportunities and challenges of automated hate speech detection

Dr. Semire Yekta[1]

[1] Universität Duisburg-Essen

The internet and the advanced digital technologies have generated a form of connectivity whereby people interact with one another in cyberspace and share content while being physically at remote locations. The physical distance between the internet users alongside the level of anonymity provided are frequently abused for cyber hate crimes, in particular through the spread of hateful and offensive speech with real-world consequences. While hate speech per se is not a new phenomenon, the development of new technologies and the usage of social media platforms as new digital public spaces led to a rapid increase of hate speech posts as well as revolutionised its global reach. The far-reaching impact of online hate speech calls for action. There is an urgent need for combating hate speech, however the response needs to be technology-driven due the limitations of the manual recognition through humans.

This research examines algorithmic hate speech detection and provides insights into how the technology operates. An overview on the methods of recognition to differentiate between hate speech and other textual data will be given. Further, two key implications for law enforcement and the society will be addressed. First, the computer-generated automatization in hate speech detection could lead to the recognition of a large number of hate speech posts that are punishable by law. Currently, it is unclear how law enforcement can handle this outcome. While the enforcement will be expected to examine and prosecute online hate posters, there are still open questions on how such resources can be provided. Second, the other implication of the automated detection is that the technology will ultimately affect the hate speech landscape and will possibly shift the way hate speech posts are written to avoid detection. Both implications will be discussed from a socio-technical point of view.