New artificial intelligence tools could help tackle online abuse

February 24 2020, by Sean Barton

Credit: CC0 Public Domain

New tools which could be used to automatically detect and counter abuse on social media, are being developed by researchers at the University of Sheffield.

The research, led by Professor Kalina Bontcheva from the University's Department of Computer Science, is developing new artificial

intelligence (AI) and natural language processing (NLP) methods that could be used to responsibly tackle abuse and hate speech online.

Launched in collaboration with Wendy Hui Kyong Chun of Simon Fraser University in Canada, the project is looking at AI methods that are currently being used to detect online abuse and hate speech within two areas; the gaming industry and messages directed at politicians on social media.

Researchers in the study intend to use their findings to develop new AI algorithms that are effective, fair and unbiased. The systems will be context-aware and respectful of language differences within communities based on race, ethnicity, gender and sexuality.

Researchers will examine the biases embedded within current content moderation systems that often use rigid definitions or determinations of abusive language. These current systems often paradoxically create new forms of discrimination or bias based on identity, including sex, gender, ethnicity, culture, religion, political affiliation or others.

The research team is aiming to address these effects by producing more context-aware, dynamic systems of detection.

Furthermore, the project is looking to empower users by making the new tools open source, so they can be embedded within new strategies to democratically tackle abuse and hate speech online. They could also be used as part of community-based care and response measures.

Professor Kalina Bontcheva said: "There has been a huge increase in the level of abuse and hate speech online in recent years and this has left governments and social media platforms struggling to deal with the consequences.

"This large rise in abuse and hate speech online has sparked public outrage with people demanding governments and social media companies do more to tackle the problem, but there are currently no effective or technical processes that can tackle the problem in a responsible or democratic manner.

"We are developing novel AI and NLP methods to address the problem while also developing a substantial programme of training for academics and early career researchers to build capacity and expertise in this key area of research."

The project, Responsible AI for Inclusive, Democratic Societies: A cross-disciplinary approach to detecting and countering abusive language online, is led by the University of Sheffield in collaboration with Simon Fraser University. It is funded by UK Research and Innovation (UKRI) as one of 10 UK-Canada projects to support the responsible development of AI, including ensuring all members of society trust AI and benefit from it.

More information: Online Abuse toward Candidates during the UK General Election 2019: Working Paper. arXiv:2001.08686 [cs.CY] arxiv.org/abs/2001.08686

Provided by University of Sheffield