New machine learning models can detect hate speech and violence from texts

April 13 2017

The words we use and our writing styles can reveal information about our preferences, thoughts, emotions and behaviours. Using this information, a new study from the University of Eastern Finland has developed machine learning models that can detect antisocial behaviours, such as hate speech and indications of violence, from texts.

Historically, most attempts to address antisocial behaviour have been done from education, social and psychological points of view. This new study has, however, demonstrated the potential of using natural language processing techniques to develop state-of-the-art solutions to combat antisocial behaviour in written communication.

The study created solutions that can be integrated in web forums or social media websites to automatically or semi-automatically detect potential incidences of antisocial behaviour with high accuracies, allowing for fast and reliable warnings and interventions to be made before the possible acts of violence are committed.

One of the great challenges in detecting antisocial behaviour is first defining what precisely counts as antisocial behaviour and then determining how to detect such phenomena. Thus, using an exploratory and interdisciplinary approach, the study applied natural language processing techniques to identify, extract and utilise the linguistic features, including emotional features, pertaining to antisocial behaviour.

The study investigated emotions and their role or presence in antisocial

behaviour. Literature in the fields of psychology and cognitive science shows that emotions have a direct or indirect role in instigating antisocial behaviour. Thus, for the analysis of emotions in written language, the study created a novel resource for analysing emotions. This resource further contributes to subfields of natural language processing, such as emotion and sentiment analysis. The study also created a novel corpus of antisocial behaviour texts, allowing for a deeper insight into and understanding of how antisocial behaviour is expressed in written language.

The study shows that natural language processing techniques can help detect antisocial behaviour, which is a step towards its prevention in society. With continued research on the relationships between natural language and societal concerns and with a multidisciplinary effort in building automated means to assess the probability of harmful behaviour, much progress can be made.

More information: Leveraging Emotion and Word-Based Features for Antisocial Behavior Detection in User-Generated Content: epublications.uef.fi/pub/urn_i … 78-952-61-2464-3.pdf

Provided by University of Eastern Finland