New technology to monitor anti-Polish hate online

August 6 2019

Credit: CC0 Public Domain

Artificial intelligence is being used to tackle anti-Polish hate crime in the run up to Brexit.

Researchers at HateLab, based at Cardiff University, are working with Samurai Labs, a Polish Artificial Intelligence laboratory, to monitor aggressive social media content and pinpoint any connections to offline events.

The year following the Brexit vote saw the largest spike in police recorded hate crime—up 57 percent on the previous year. In 2017/18, there were 94,098 hate crime offences recorded by the police in England and Wales—an increase of 17 percent compared to the year before.

More than 900,000 Polish people live in the UK, making them the largest national minority. Pilot studies conducted by Samurai Labs show that up to 5 percent of material published about this group on UK social media has a negative or offensive connotation.

HateLab is a global hub for data and insight into hate speech and crime. Using data science methods, including ethical forms of AI, the initiative was set up to measure and counter the problem of hate both online and offline. The Online Hate Speech Dashboard has been developed by academics with policy partners to pre-empt outbreaks of hate crime on the streets.

Professor Matthew Williams, Director of HateLab at Cardiff University, said: "We know that the 2016 EU Referendum prompted a surge in online hate speech and coincided with significant increases in hate crimes offline. As the United Kingdom prepares to leave the European Union, using the most advanced methods of artificial intelligence is going to be vital in helping the authorities to quickly recognise warning signs and provide reassurance and security to the Polish community living here."

Artificial Intelligence algorithms developed by Samurai Labs can accurately differentiate between web aggression and harmless

comments. They are also capable of pinpointing the precise type of aggression. These features are particularly important when there is a need to identify peaks in such content around offline events.

Michał Wroczyński, CEO of Samurai Labs, said: "We first distinguished eight categories of verbal aggression targeted at UK residents of Polish descent, starting with insults and requests for them to go home, to threats to life and limb.

"Verbal aggression does not necessarily refer to the origin. Most of it is simply offensive words that could be targeted at any nation. Such events constitute almost 40 percent of the entire total of cases.

Gniewosz Leliwa, Director for Artificial Intelligence Research at Samurai Labs, said: "We have to be very precise and work with a deep understanding of the language. Keywords are certainly insufficient. It is possible for someone to make a punishable threat without using any of the characteristic words. "Similarly, the use of vulgarity is not always necessarily associated with an aggressive message. Our system is able to understand the context and correctly separate a real threat from an ordinary conversation."

HateLab, part of the Social Data Science Lab based between the University's School of Social Sciences and School of Computer Science and Informatics, has been established with funding from the Economic and Social Research Council (ESRC) as well as the US Department of Justice. It has received a total of £1,726,841 in funding over five ongoing projects.

Professor Williams added: "We are very excited to be working with Samurai Labs. Their innovative approach has proved to be very accurate in detecting cyberbullying and internet aggression. We are very much looking for similar successes in detecting aggression against minorities.

"In view of Samurai Labs' Polish roots, we hope that we will be able to gain new knowledge and better understand the nuances associated with this problem. The technology that they are developing will form an important part of our Online Dashboard."

Provided by Cardiff University