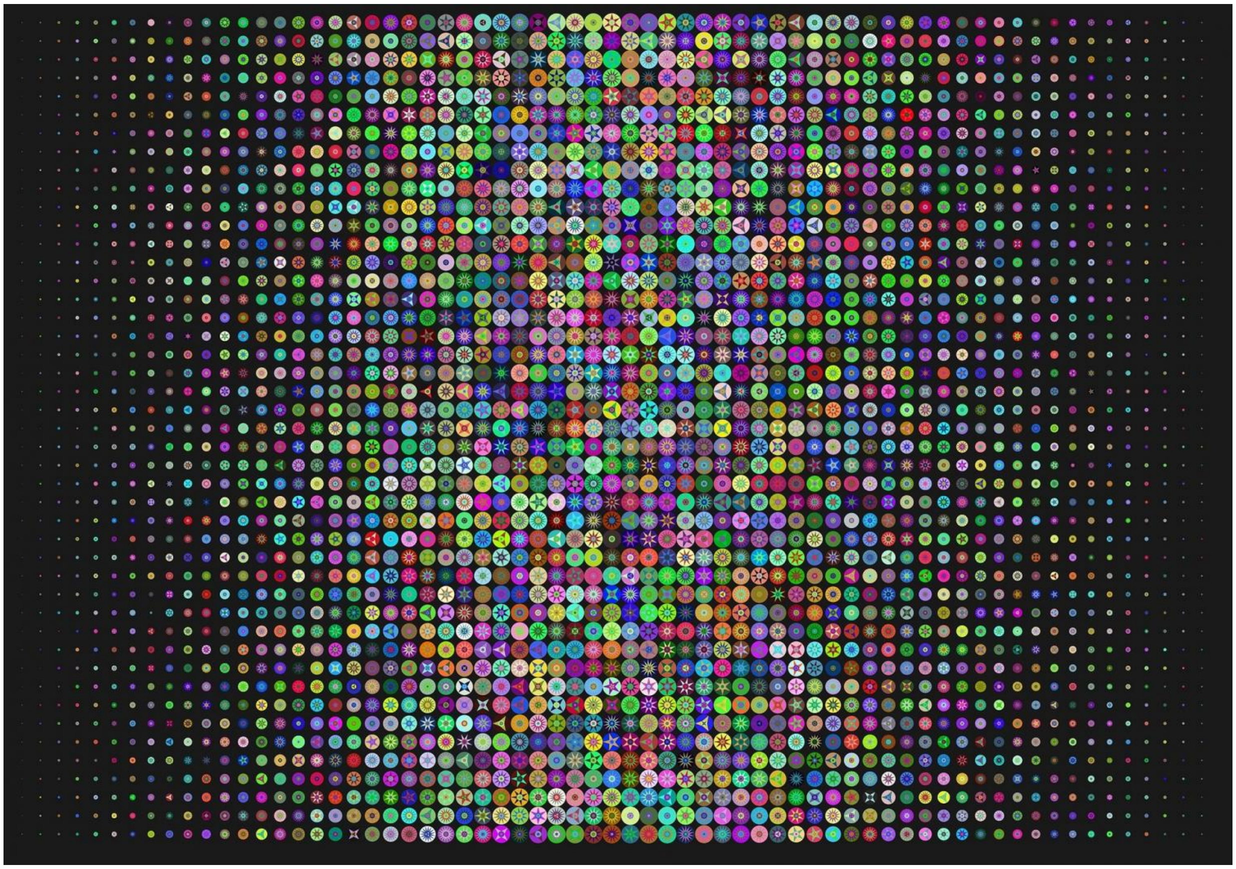
PHYS.ORG

# Internet researchers harness the power of algorithms to find hate speech

October 13 2017



Credit: CC0 Public Domain

During the municipal elections in spring 2017, a group of researchers and practitioners specialising in computer science, media and communication implemented a hate speech identification campaign with the help of an algorithm based on machine learning.

At the beginning of the campaign, the algorithm was taught to identify hate speech as diversely as possible, for example, based on the big data obtained from open chat groups. The algorithm learned to compare computationally what distinguishes a text that includes hate speech from a text that is not hate speech and to develop a categorisation system for hate speech. The algorithm was then used daily to screen all openly available content the candidates standing in the municipal elections had produced on Facebook and Twitter. The candidates' account information were gathered using the material in the election machine of the Finnish Broadcasting Company Yle.

All parties committed themselves to not accepting hate speech in their election campaigns. On the other hand, if the candidate used a personal Facebook profile instead of the page created and reported for the campaign, it was not included in the monitoring. Finnish word forms and the limited capability of the algorithm to interpret the context the same way humans do also proved to be challenging. The Perspective classifier developed by Google for the identification of hate speech has also suffered from the same problems in recognising the context and, for example, spelling mistakes.

Once the messages have been identified, it is key to define the actions that will follow.

"From the point of view of the authorities, there were no more than 20 messages that caused measures. Listing words as such is not sufficient because words get their meaning from the way they are combined. On the other hand, without the hate speech machine and researchers, we would not have the resources to do monitoring on this scale," says Non-Discrimination Ombudsman Kirsi Pimiä.

## Hate speech focuses on emotions and beliefs of inequality

To teach the algorithm, the researchers prepared material consisting of thousands of messages and cross analysed it to be able to make it scientifically valid.

"When categorising messages, the researcher has to take a stance on the language and context, and it is therefore important that several people participate in interpreting the teaching material," says Salla-Maaria Laaksonen from the University of Helsinki.

It was important that all types of hate speech could be found during the campaign. Immigration and asylum seekers are often the most prominent themes, but it is equally important to identify hate speech targeted at women, ethnic minorities or certain political opinions.

"Hate speech has always existed. It has always been produced to support the status of one's own group and to discriminate against the others, but social media has now made it more visible than before. Expression and beliefs based on emotions are emphasised and they are also circulated online. For example, if the candidate removed what he or she had written soon after it had been published during the campaign, it could still remain as a screen capture," says Reeta Pöyhtäri from the University of Tampere.

Hate speech is defined in the legislation in many European countries, whereas ordinary people use the term hate speech with very broad meanings. All angry speech is not punishable hate speech from the point of view of the law. For example, it has to be targeted at groups that are in positions that are more vulnerable, be discriminatory or contain a threat of violence. The project used the definition of hate speech drawn up by the Council of Europe and the Ethical Journalism Network.

## Hate speech also as topic in a conference

According to Salla-Maaria Laaksonen, social media services and platforms, such as Facebook and Twitter, could utilise identification of hate speech if they wanted to and that way influence the activities of internet users.

"There is no other way to extend it to the level of individual citizens."

Apart from the changes in Finnish society and culture, the economic situation is also regarded as a factor that increases xenophobia. Changing the behaviour involving hate speech therefore seems to be a challenging task in spite of the monitoring, moderation, campaigns to change attitudes and media education that are carried out.

"We should analyse the reasons behind hate speech in more detail. It would be interesting to know who are the people sending those hate messages, what motivates them and how many of them are trolls. Are there any common factors in their circumstances, such as social exclusion, and why do they have to demonstrate their hatred by despising people and by questioning other people's human dignity," Kirsi Pimiä says.

The work done during the campaign will continue in a conference organised by the Association of Internet Researchers in Tartu between 18 and 21 October. One of the workshops will discuss the state of hate speech on the internet, the possibilities and challenges in the identification of hate speech, and the ways to respond to the challenges hate speech poses online. The workshop is organised jointly by the researchers of Aalto University and the Universities of Helsinki and Tampere who participated in the campaign and Open Knowledge

Finland.

"It was important for us to reflect on how researchers could contribute to the solution of such an important societal problem. Confrontation takes place at many levels in society today and we would like to challenge the international science community to discuss this phenomenon together in our workshop," says Matti Nelimarkka who is a researcher at Aalto University and HIIT.

In addition to the three universities, the Office of the Non-Discrimination Ombudsman and the Finnish League for Human Rights together with researchers from the Advisory Board for Ethnic Relations, Open Knowledge Finland, Futurice and Rajapinta ry participated in the campaign implemented during the municipal elections. The project is linked to four research projects funded by the Academy of Finland and the Kone Foundation.

Provided by Aalto University

Citation: Internet researchers harness the power of algorithms to find hate speech (2017, October 13) retrieved 23 April 2024 from https://phys.org/news/2017-10-internet-harness-power-algorithms-speech.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.