DARING, SHARING, CARING

photo by AvoinGLAM / CC BY 2.0
We facilitate the development of a fair and open digital society by means of open knowledge and open collaboration.
LESS HATE IN POLITICS!

Machine learning and interventions as tools to mitigate online hate speech in political campaigns

Teemu Ropponen
Executive Director, Open Knowledge Finland
teemu.ropponen@okf.fi, @troppone

Salla-Maarja Laaksonen, Reeta Pöyhtäri, Matti Nelimarkka et al.
PROJECT BACKGROUND

● NGO-company-government-university collaboration

● Goals:
  ○ To promote campaigning without hate speech in the Finnish municipal elections 2017 (33 000 candidates)
  ○ To create tools of automated detection of hate speech and test them out
PROJECT BACKGROUND

- Prior to the elections, all political parties were asked to sign a commitment of zero-tolerance towards hate speech, and were notified of monitoring
  - *Charter of European Political Parties for a Non-Racist Society* signed by all parties in the Parliament (8)
- Automated streaming data collection (1 month):
  - 6400 Facebook pages
  - 1308 Twitter profiles
- Handles and urls extracted from YLE data
  - “Election advice” information
WHAT IS HATESPEECH

The Council of Europe’s Committee of Ministers’ Recommendation 97(20) on hate speech:

“Hate speech covers all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance.”

Hate Speech definition from Ethical Journalism Network

1. The Position or Status of the Speaker
2. The Reach of the Speech
3. The Objectives of the Speech
4. The Content and Form of Speech
5. The Economic, Social and Political Climate

and Article19

- Context of the expression
- The speaker
- Intent
- Content of the expression
- Extent and magnitude of the expression
- Likelihood of harm occurring
What is hate speech?

The Council of Europe’s Committee of Ministers Recommendation 97(20) on hate speech:

“Hate speech covers all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance.”

Features of hate speech:

- A call to violent action
- A call to discriminate or to promote discrimination
- Degrade human dignity based on their characteristics
- Threat of violence of promotion of violent action
- Contempt, solicitation, namecalling, slandering

+ a list of 14 different target groups we aimed to cover with the training data.
A 5 POINT TEST FOR JOURNALISTS

4 THE CONTENT ITSELF

Is the speech dangerous?
Could it incite violence towards others?

(3) Ethical Journalism Network
1562 unique anonymized messages from various online forums annotated by four trained human classifiers (Krippendorff's alpha = 0.791 for a subset of 100 messages)

3: clearly hate speech
2: disturbing angry speech
1: normal discussion with a critical tone
0: neutral
HOW DID IT GO?

- Data in total 26,618 posts
- ML system classified 205 messages as hate speech
- Manual screening done by Non-discrimination Ombudsman
- Final counts:
  - Level 2: 43
  - Level 3: 5
  - Two party letters and a few requests for police investigation
- The number of predicted false positives decreased - feedback loop worked
- Code released with MIT licence, data copyrighted
“What will the government do to make sure the non-discrimination ombudsman focuses on the non-discrimination law and not monitoring of the writings of candidates”

“Mitä hallitus aikoo tehdä huolehtiakseen siitä, että yhdenvertaisuusvaltuutetun toiminta keskittyvyy yhdenvertaisuuslain noudattamisen valvontaan eikä vaaliehdokkaiden kirjoitusten tarkkailemiseen?”
THANKS!

Teemu Ropponen - @troppone
Open Knowledge Finland - @okffi

Together with:
Salla-Maarja Laaksonen, Reeta Pöyhtäri, Matti Nelimarkka, Teemu Kinnunen, Kimmo Kärkkäinen et al.
What is hate speech?
Hate speech – what are we talking about?

AoIR pre-conference – Less hate in politics!
18.10.2017, Tartu Estonia
Reeta Pöyhtäri, Postdoctoral research fellow
Tampere Research Centre for Journalism, Media and Communication COMET
University of Tampere
Freedom of expression
Free flow of ideas, right to express and publish them (ideas of Enlightenment, e.g. John Stuart Mill 1859, ‘On Liberty’)

Universal Declaration of Human Rights, Article 19 (1948)
Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers. → PRINCIPLES OF EQUALITY, DIGNITY & NON-DISCRIMINATION in enjoying human rights, and protection by law

International Covenant on Civil and Political Rights, Article 19 (1966)
1. Everyone shall have the right to hold opinions without interference.
2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.
3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:
   (a) For respect of the rights or reputations of others;
   (b) For the protection of national security or of public order (ordre public), or of public health or morals.
Freedom of expression

● Free flow of ideas, right to express and publish them
● Guaranteed by law but NOT without restrictions; abuse of free speech rights limited
● Should the right to freedom of speech be restricted, if it is abused and to what extent?
● Laws regulate, complemented by e.g. ethics of journalism and self-regulatory systems (e.g. Journalistic Codes of Conduct)
● E.g. Journalists using the publisher’s rights for free speech in media outlets; this right extended to public in comment fields, thus in a space owned by the news organisation
● On-going discussion about the rights and responsibilities of Internet intermediaries (e.g. obligation to abide HR laws or state legislation)
Freedom of expression

Hate speech and other forms of abusive online practices endanger the aims of public engagement and principles of free discussion.
Hate speech and cyberhate
The ‘Hate Speech Pyramid’ (by Article 19, 2015)

**Must be prohibited:** Incitement to genocide and other violations of International Law → Genocide Convention + Rome Statute
Genocide, mass destruction; their promotion

**Must be prohibited:** Advocacy of discriminatory hatred constituting incitement to hostility, discrimination or violence → Article 20(2) ICCPR
All propaganda based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination → The International Convention on the Elimination of all Forms of Racial Discrimination (the ICERD) Article 4

**May be prohibited:** Hate speech which may be restricted to protect the rights or reputations of others, or for the protection of national security or of public order, or of public health or morals → Article 19(3) ICCPR
Necessary restrictions to free speech
But: restrictions provided by law; in pursuit of a legitimate aim (such as reputation of others); necessary in democratic society

**Free speech to be protected:** Lawful “hate speech” raising concerns in terms of intolerance → Article 19 ICCPR
Everyone has right to free speech, but this comes with responsibilities
Hate speech: varying definitions

Council of Europe’s Committee of Ministers’ Recommendation 1997(20) on “hate speech”:

“the term ‘hate speech’ shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.”

• In most countries ”hate speech” is not defined by an explicit law, but crimes against freedom of speech, containing hate speech, include e.g. incitement to hatred, blasphemy, defamation, libel, illegal threat, harassment, assault

(N.B. Due to differences in legal implications e.g. group vs. individual)
Hate speech: varying definitions

Hate: **the intense and irrational emotion of opprobrium, enmity and detestation towards an individual or group**, targeted because of their having certain - actual or perceived – **protected characteristics** (recognised under international law).

“Hate” is more than mere bias, and must be **discriminatory**. Hate is an indication of an emotional state or opinion, and therefore distinct from any manifested action.

Speech: **any expression imparting opinions or ideas** – bringing an internal opinion or idea to an external audience. It can take many forms: written, non-verbal, visual or artistic, and can be disseminated through any media, including internet, print, radio, or television. (Article 19)
Objects of hate speech

race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth, indigenous origin or identity, disability, migrant or refugee status, sexual orientation, gender identity or intersex status
International freedom of expression standards protect expression that is offensive, disturbing or shocking, and do not permit limitations premised solely on the basis of “offence” caused to an individual or group.

International human rights law provides no right to individuals to be free from offence, but it does unequivocally protect their right to counter such offence and speak out against proponents of that speech.

European Court, Handyside v. UK, App/ No. 5493/72, 7 December 1976
Cyberhate

• Internet and social media especially are forums of hate speech and cyberhate → broader concept than hate speech

• ICCA report (Inter-parliamentary Coalition for Combating Anti-Semitism, 2013) defines as cyberhate at least: Racism, anti-Semitism, religious bigotry, homophobia, bigotry aimed at the disabled, political hatred, rumor-mongering, misogyny and violent pornography, promotion of terrorism, cyberbullying, harassment and stalking, speech that silences counter-speech such as slurs, insults and epithets, speech that defames an entire group; also e.g. trolling, revenge porn

• Cyberviolence: “to advocate violence, separation from, defamation of, deception about or hostility towards others” through using ICTs (Franklin 2010, 2, www.hatedirectory.com/hatedir.pdf)
Questions concerning hate speech

What constitutes a **protected characteristic** for identifying an individual or group that is the targets of ‘hate speech’?

The degree of focus given to the **content** and **tone** of the expression?

The degree of focus given to **harm caused**; whether the expression is considered to be **harmful in itself** for being degrading or dehumanising or is considered to have a potential or actual harmful consequence, such as:
- inciting a manifested **action** against the target by a third person or group of people, such as violence
- causing an **emotional response** in the target, such as insult or distress; or
- **negatively affecting societal attitudes**, by “spreading” or “stirring up” hatred?

The need for **causation** to be proven between the expression and the specified harm?

The need for any harm to be **likely** or **imminent**?

The need to **advocate** harm, implying that the speaker has an intent for harm to occur, and public dissemination of the expression? (Article 19)
FOUNDED 2012

Now:
~ 4000+ PEOPLE IN NETWORK
~ 400 MEMBERS
~ 9 WORKING GROUPS
~ 20 PROJECTS
~ 10 EMPLOYEES
~ 500 000€ BUDGET
(including MyData 2016 budget)

MODES OF ACTIVITY

● PROJECT HOME
● EVENTS & COMMUNITY
● ADVOCACY
● EDUCATION & SERVICES
● OPENNESS LABORATORY
● GLOBAL NETWORK
A world where knowledge creates power for the many, not the few.