Less hate in Politics!

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

AoIR 2017 preconference, Tartu, Estonia, Oct 18 2017
Salla-Maaria Laaksonen, Reeta Pöyhtäri, Matti Nelimarkka et al.

Workshop Program

09:00 – 09:15 **Welcome**, introductions and organization of the workshop.
An overview of the Finnish case (Salla-Maaria Laaksonen et al.)

09:15 – 09:30 **Keynote 1**: What is hate speech and how does it relate to freedom of speech? (Reeta Pöyhtäri)

09:30 – 09:45 **Keynote 2**: Hate speech as a technological problem (Matti Nelimarkka)

09:45 – 10:30 **Workshop** 1: Classifying and tagging content for hate speech detection
SHORT COFFEE BREAK

11:00 – 11:30 **Demos** of automated classification with the common training data set (Matti)

11:30 – 12:30 **Workshop 2**: Discussion of the best practices: Interventions, action research and moderation as means to tackle hate speech

Say 1-3 words that summarize the most pressing questions regarding hate speech in your country?

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
- 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
- **Automated streaming data collection** (1 month):
  - 6400 Facebook pages
  - 1308 Twitter profiles
  - Handles and urls extracted from YLE VAA data

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.”

Hate Speech definition from Ethical Journalism Network

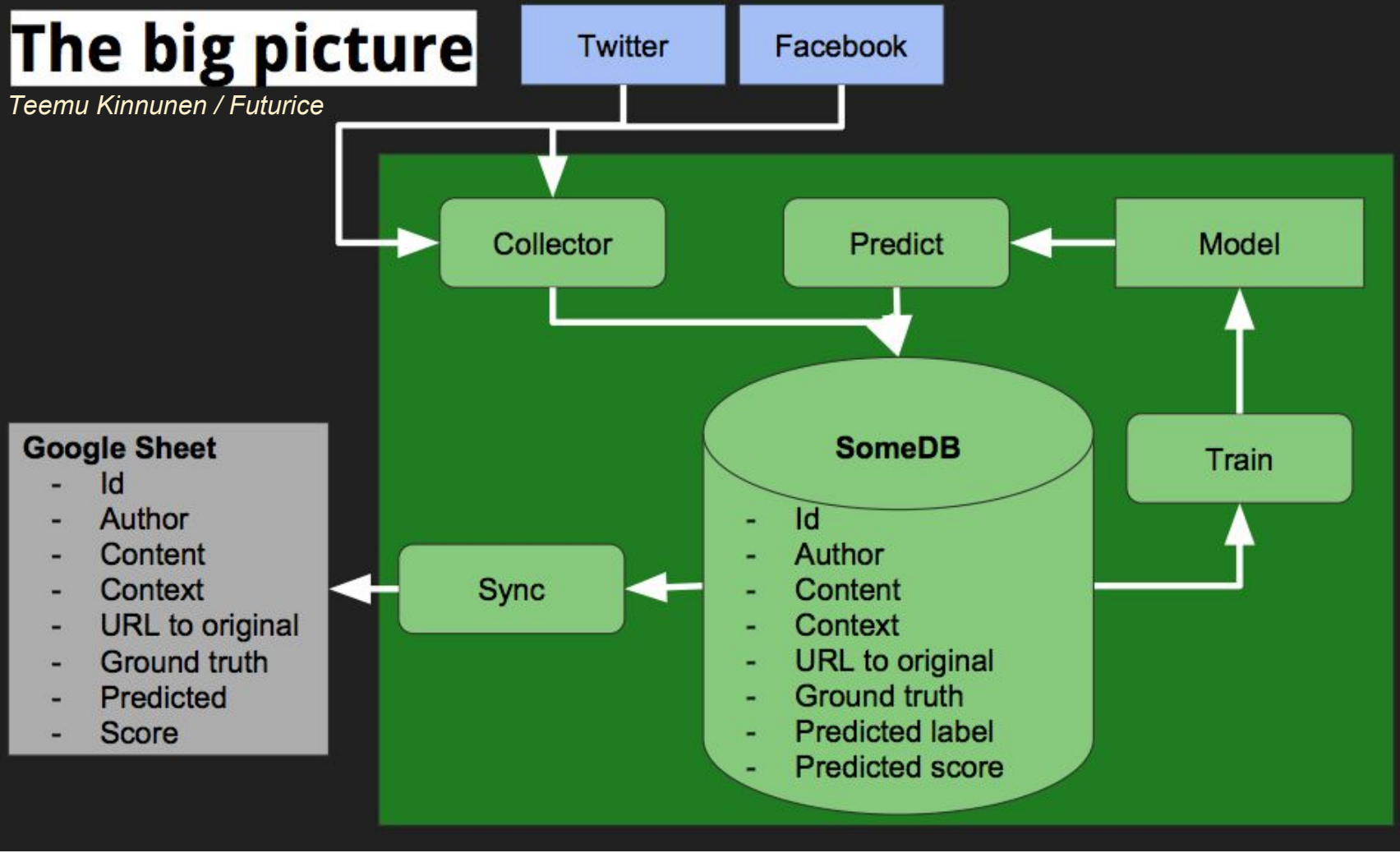
1. The Position or Status of the Speaker
2. The Reach of the Speech
3. The Objectives of the Speech
4. The Content and Form of Speech
5. The Economic, Social and Political Climate

and Article19

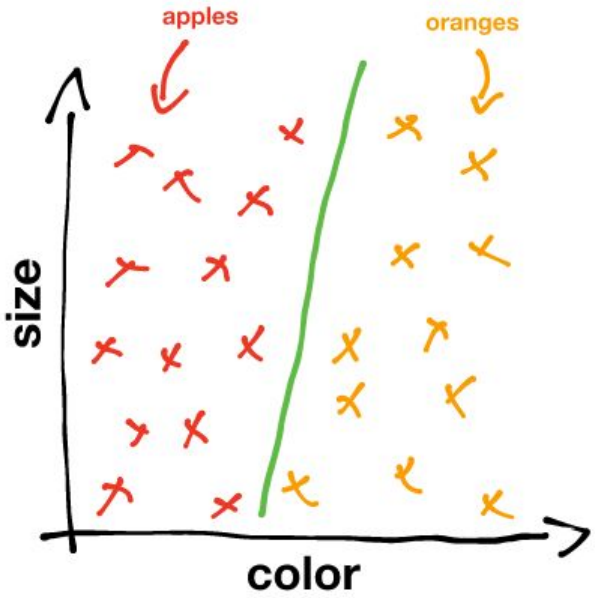
- Context of the expression
- The speaker
- Intent
- Content of the expression
- Extent and magnitude of the expression
- Likelihood of harm occurring

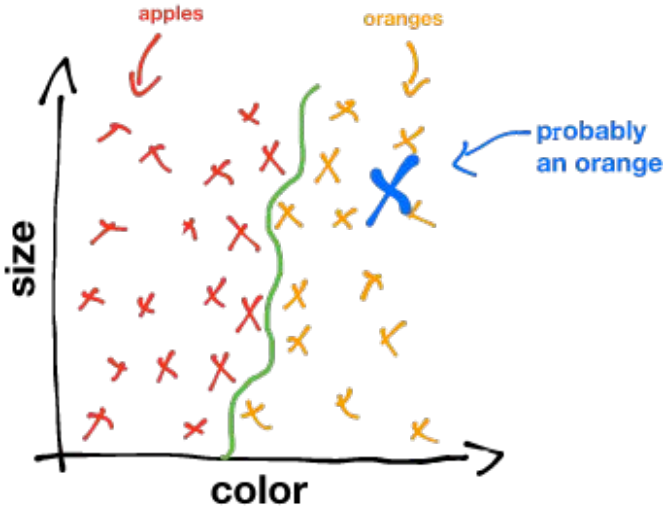
The big picture

Teemu Kinnunen / Futurice



Supervised machine learning classification





Picture: Geng & Shih 2016

Training data

	text	vihapuhe_as
2055014895	Våra medverkande stödjer kampen mot rasism	0
1143111119	Kolmen vuoden ajanjaksolla 51 alaikäistä teki itsemurhan. Nuorimmat henkensä riistäneet olivat vasta 13-vuotiaita. "Yhdenkään nuoren ei pitäisi riistää henkeään, koska tyttöystävä on jättänyt. Ketään ei myöskään pitäisi kiusata esimerkiksi seksuaalisen identiteetin takia niin paljoa, että itsemurha tuntuu ratkaisulta", Valonen sanoo. Valosen mukaan osa itsemurhista on todennäköisesti ollut hätähuutoja, joissa nuorella ei ole ollut vakaata tarkoitusta tappaa itseään, mutta vahingossa ollaan menty liian pitkälle. Kolme neljästä itsemurhan tehneestä lapsesta oli poikia.	0
1143111119	Maria "inte min talman" Lohela tietää: "Islaminuskoisissa kulttuureissa ... vääräuskoisen (ei-islaminuskoisen) naisen raiskaaminen on jopa kannustettava teko." Onko Lohela myös sitä mieltä, että kristinuskoisissa kulttuureissa pikkutyttöjen raiskaaminen, tappaminen ja uunissa polttaminen on kannustettava teko, koska Jammu Siltavuori?	3
1469548606	tuli kaveripyyntö. rupean suhtautuun näihin vähän epäluulosesti. tälläkään ei ole mitään julkaisuhistoriaa sivullaan. muutama yhteinen kamu kyllä olisi, mutta..... https://www.facebook.com/profile.php?id=100007955386682&fref=jewel	0
2619293539	Jos 30 tutkijaa tuottaa vuodessa valtiolle 80 miljoonan euron lisätulot, niin ainakin heille voisi pientä palkankorotusta ehdottaa! Lisää tällaista vai mitä? :)	0
1469548606	Tuleeko teillä ylimääräsii kaveripyyntöjä? Itsellä koko ajan Allahu Kakbar, Nazir, alibabatsäkäläkä yrittää kaveriks... Saatanan persenaamat jättäköön mut rauhaan! En puhu koraanii!	3
2245722809	Voi kaaheeta. Mamut joutuisivat jonottamaan rasistien jonoissa ja hakemaan vaatteensa kirpputoreilta. Itäkeskuksessakin joutuisi palloilemaan ryysyissä, kun ei olisi enää varaa merkkivaatteisiin.	1
	TAPAAJAVANHEMPI – UHKA UUSPERHEELLE JA YHDEN VANHEMMAN PERHEELLE? Kun lapsen vanhemmat eroavat, hajoaa lapsen niin kutsuttu alkuperäinen ydinperhe. Koska lasta ei tietenkään voi jakaa kahteen tulee	

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

Finns Party's councillor to be investigated for social media posts

FINLAND / CREATED: 16 OCTOBER 2017

POLITICS

Chairperson Jussi Halla-aho and deputy chairperson Laura Huhtasaari of the Finns Party spoke to the press in Helsinki on 4 August, 2017. Both Halla-aho and Huhtasaari were elected to key positions within the party in early June.

Kirsi Pimiä, the Non-Discrimination Ombudsman in Finland, has revealed that a pre-trial investigation is set to be opened into statements made on social media by Sebastian Tynkkynen (PS), a councillor for the City of Oulu, in the run-up to the municipal elections held on 9 April, 2017.

Pimiä estimates that the campaign ran by the outspoken councillor was partly racist and widened social divisions.

What is hate speech?

UNIVERSITY OF TAMPERE
Faculty of Communication Sciences

COMET
Journalismin, viestinnän ja median tutkimuskeskus
Tampere Research Centre for Journalism, Media and Communication

Hate speech – what are we talking about?

AoIR pre-conference – Less hate in politics!

18.10.2017, Tartu Estonia

Reeta Pöyhtäri, Postdoctoral reseach 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 Enlightment, e.g. John Stuart Mill 1859, 'On Liberty')

Universal Declation 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)

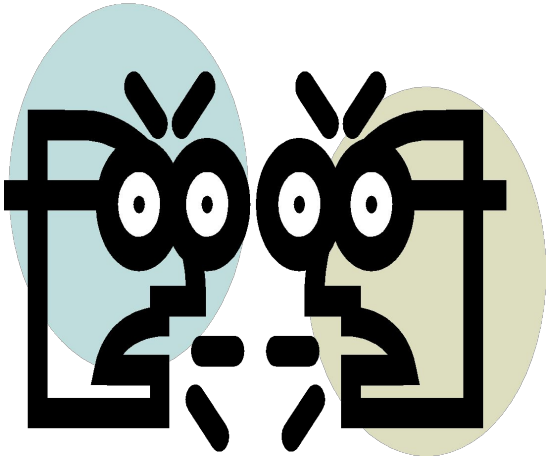
- E.g. Journalists using the publisher's rights for free speech in media outlets; this right extended to public in comment fields, thus in a space owned by the news organisation

- 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



Copyright: UNESCO/
Zemgus Zaharans

Copyright: Unknown

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

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

Hate speech to be tolerated

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

Workshop

Workshop 1: What is hate speech?

Let’s do a little classification exercise!

1. Go to: http://tinyurl.com/lesshatedata
2. Make your **own copy** of the Sheet named “Original Messages” by clicking the small arrow on the tab > “Duplicate” > **rename it with your name** or nick.
3. Classify each message to one of the following categories:
   3: clearly hate speech
   2: disturbing angry speech
   1: normal discussion with a critical tone
   0: neutral

Demos

Automated text analysis for hate speech detection

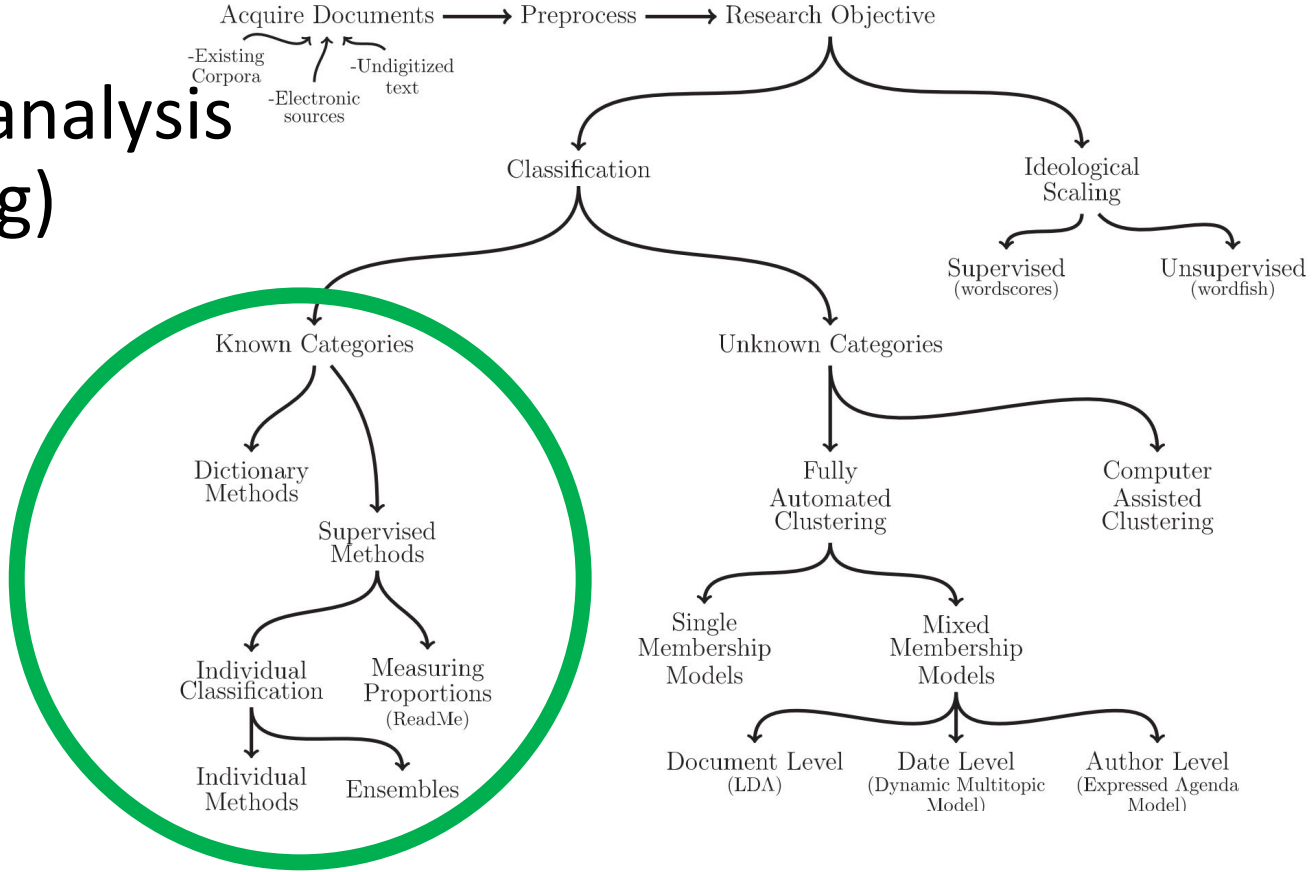
- Intro (3 mins)
- Context freeness (3)
- Different techniques: Dictonaries (3)
- Different techniques: SVMs etc. (5)
- Validity and reliability considerations (3)

Automated text analysis (machine learning)

	Human analysis	Computer-based analysis
A priori schema	Codebook based content classification	Supervised learning
Data-driven	Grounded theory Content classification without a code book	Unsupervised learning

Purhonen, S., & Toikka, A. (2016). “Big datan” haaste ja uudet laskennalliset teksiaineistojen analyysimenetelmät. *Sosiologia*, (1), 6–26.

Automated text analysis (machine learning)



Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–29.

The bag of words

Mercy. No Siberia!

=

No Mercy. Siberia!

Dictionary-based approaches

Political Behavior
September 2017, Volume 39, Issue 3, pp 629–649 | Cite as

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

Authors | Authors and affiliations

Kevin Munger

Original Paper
First Online: 11 November 2016

Dictionary-based approaches

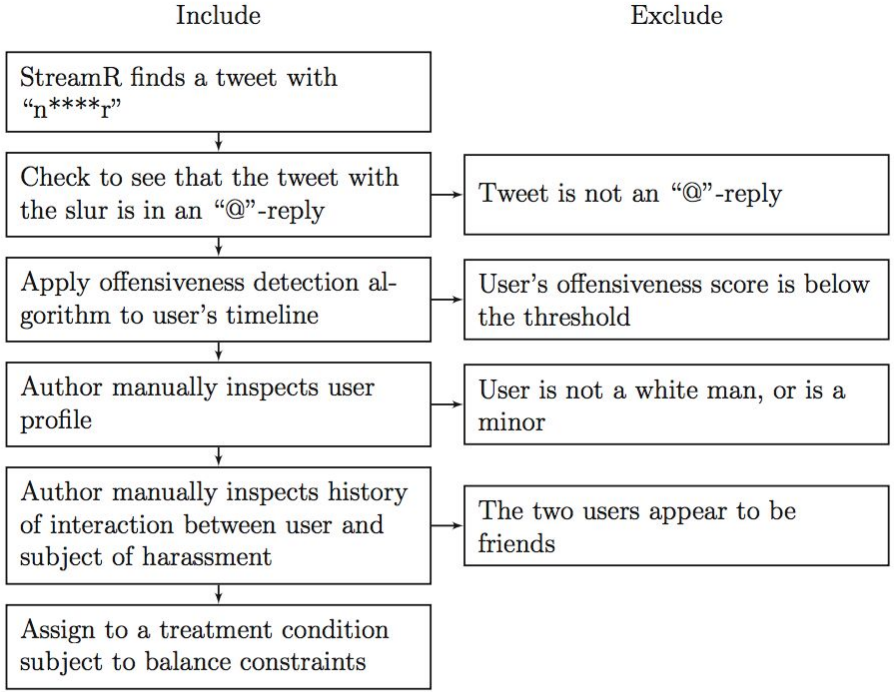


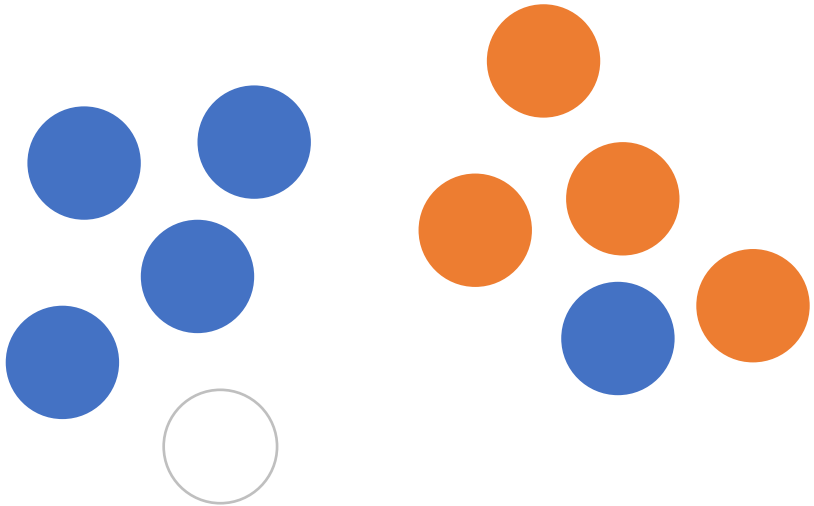
Fig. 1 Sample selection process: This flowchart depicts the decision process by which potential subjects were discovered, vetted and ultimately included or excluded

Supervised methods

- Support Vector Machines (SVMs)
- Naïve Bayes
- Decision trees
- Random forest
- Linear regression & logistic regression
- Neural networks

etc..

Supervised methods



P&I Policy & Internet

Explore this journal >

Article

Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making

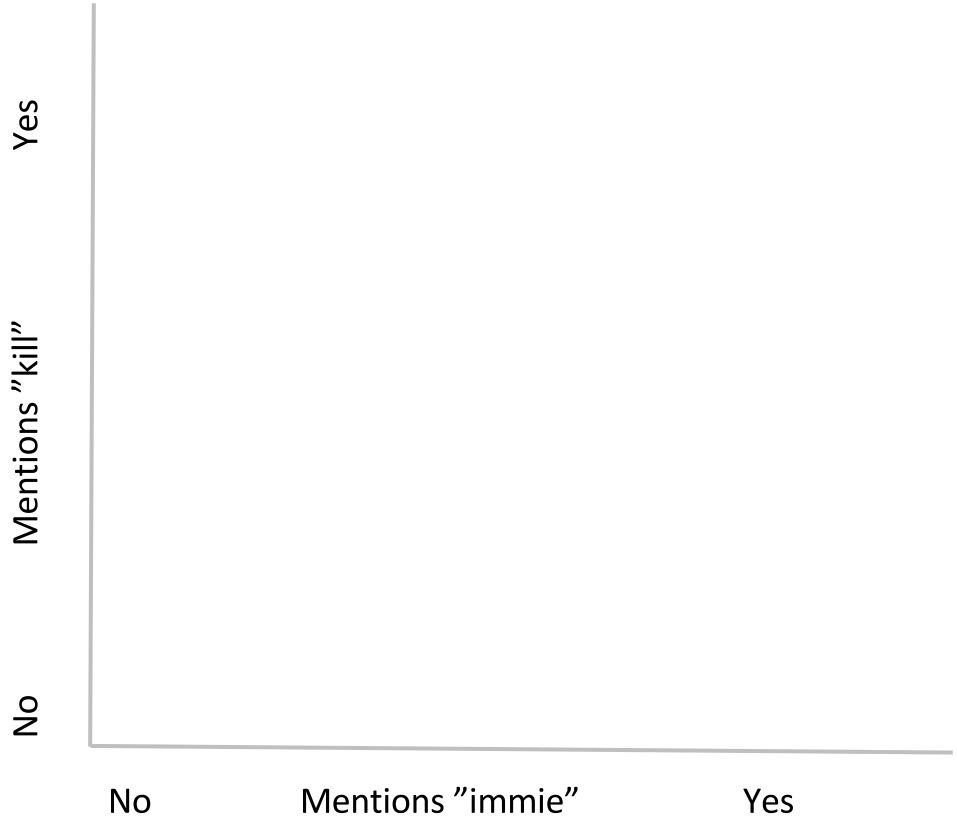
Pete Burnap, Matthew L. Williams

First published: 22 April 2015 Full publication history

DOI: 10.1002/poi3.85 View/save citation

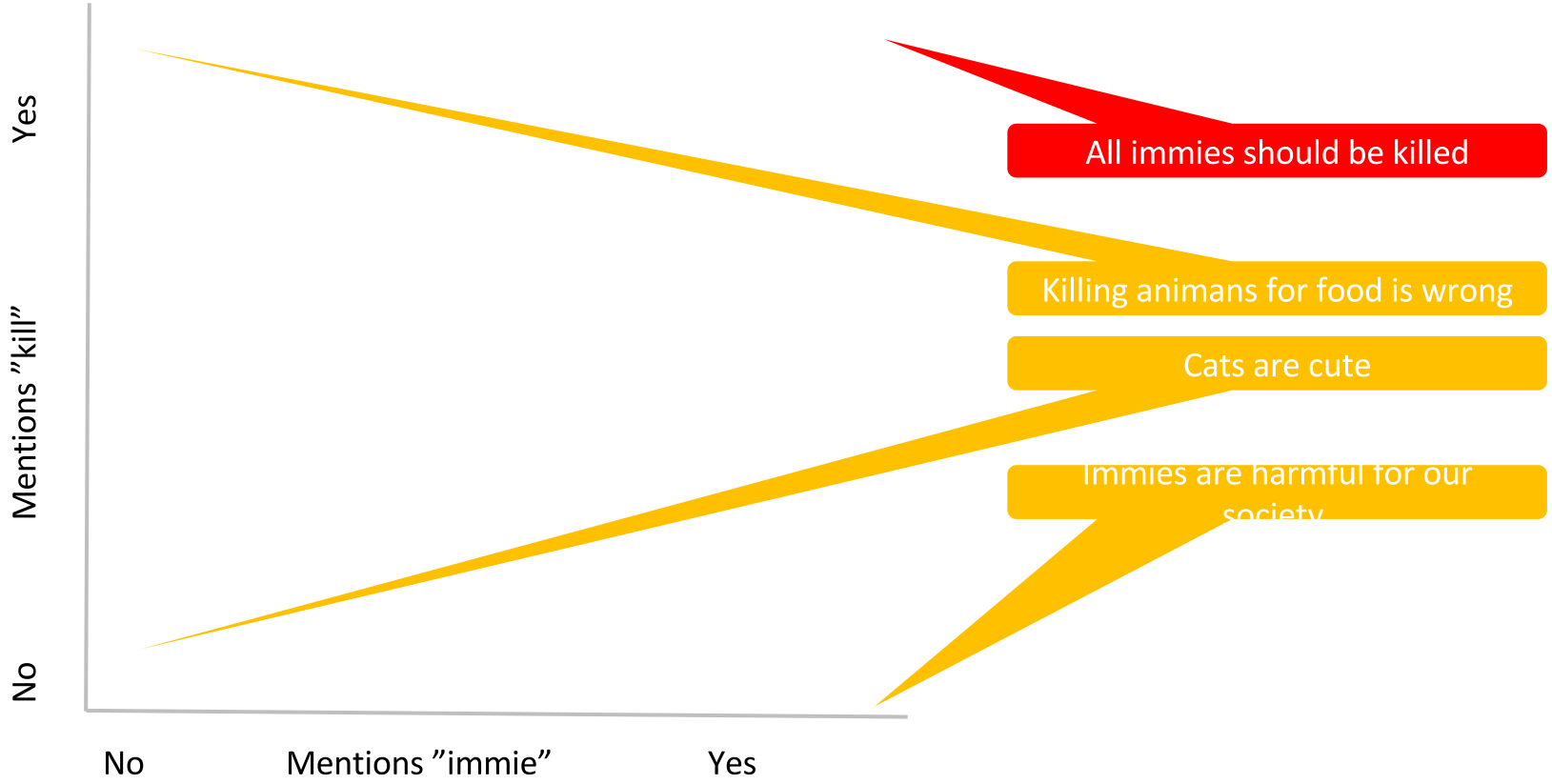
Cited by (CrossRef): 12 articles Check for updates | Citation tools

Supervised machine learning



Supervised machine learning

Supervised machine learning



Black boxes?

Unsure if this will be perceived as toxic (0.61) Learn more

SEEM WRONG?

I think some Finns are not nice.

Likely to be perceived as toxic (0.78) Learn more

I think some women are not nice.

Likely to be perceived as toxic (0.74) Learn more

SEEM WRONG?

I think some Canadians are not nice.

Unsure if this will be perceived as toxic (0.57) Learn more

SEE

I think some men are not nice.

Perspective was created by Jigsaw and Google's Counter Abuse Technology team in a collaborative research project called Conversation-AI. We are also open sourcing experiments, models, and research data to explore the strengths and weaknesses (e.g. potential unintended biases) of using machine learning as a tool for online discussion.

Context matters!

TABLE 3. 2005 Senate to House Classification Accuracies (Percent)

	2005 Senate cross validation	2005 House prediction
majority baseline	55.0	51.5
svm-bool	73.7	51.5
svm-ntf	55.6	51.5
svm-tfidf	69.7	65.8
nb-bool	81.0	51.5
nb-tf	86.0	67.6

-30 %-units

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying Party Affiliation from Political Speech. *Journal of Information Technology & Politics*, *5*(1), 33–48.

Interventions and other initiatives

HATE SPEECH

TURNING THE PAGE OF HATE:
A MEDIA CAMPAIGN FOR
TOLERANCE IN JOURALISM

When it comes to hate speech, journalists and editors must pause and take the time to judge the **potential impact** of offensive, inflammatory content.

The following test, developed by the EJN and based on international standards, highlights questions in the **gathering**, **preparation** and **dissemination** of news and helps place what is said and who is saying it in an **ethical context**.

DONT **SENSATIONALISE!**
AVOID THE ***RUSH*** TO PUBLISH
TAKE A **MOMENT OF REFLECTION**

EthicalJournalismNetwork.org

EJN

Existing measures against hate speech

Monitoring **Building and using tools to detect hate speech**	Automatic detection vs. manual work; technologies for monitoring	International monitoring (e.g. UN, EU, European Council)	Official monitoring (e.g. police; ombudsman)	Un/semi-official monitoring e.g. Kenya (UMATI), Finland (our example)	News media's online moderation (WAN-Ifra report online)	Discussion boards/fora/ Internet platforms (peers vs. company service) (e.g. Perspective/ Google)
Limiting **Efforts to control and limit hate speech**	Legislation: international, regional, national Rabat Action Plan (OHCHR) (Incitement to national, racial and religious hatred)	Legal praxis e.g. EU-level (European Court of HR); national	Implemen-tation of laws; changing laws; societal and institutional practices	Official policies Also: Support for free expression and equality in society (e.g. Camden Priciples) Social policy	Policies of internet companies	Online moderation

Existing measures against hate speech

Countering **Creating and implementing various practices to counter hate speech online**	Media Literacy and education campaigns, e.g nohatespeech; Torjun vihapuhetta (opposing hate speech, Min. Edu in Finland); Toolkits for media education and for dealing with hate speech Education of perpetrators?	Raising awareness: Hate speech materials, e.g. Article19 toolkit "Hate speech explained"; Ethical Journalism Network: "5 point test for hate speech" for journalists; Equality trainings	Building tools for hate speech detection: HateBase	Making cases known and visible, e.g. video campaigns; media articles; comic books; poetry	Support for plural media landscape and inter-group dialogue In media: creating violent-free debates, constructive, or conciliatory journalism	Active role of researchers?
Mobilizing **Activating people to act against hate speech; grassroots movements**	E.g. Czech republique www.hatefree.cz	Myanmar: Panzagar ("flower speech")	Finland: #lääppijät			

Existing measures against hate speech

Lobbying **Trying to change the practices of the largest Internet intermediaries (Google, Twitter, FB, Youtube…)**	International discussions (e.g. EU, European Council, the UN)	NGO's, e.g. Anti-Defamation League; Online Hate Prevention Institute actively lobbying				
Assessing **Evaluating the current situation / the problem of hate speech / the existing measures to counter it**	Research: hate speech, users, producers, victims, online affordances and practices, technology, legislation, societal contexts …	Assessment of existing initiatives	Discussions: local, national, regional, international, intergovern-mental			

Workshop 2

Discussion in three smaller groups:

- What could be the role of media industry, governments, non-governmental organizations, individuals and other actors in prevention of and interventions against hate speech?
- How do you think any content that the **automated systems** believe to contain hate speech should and could be processed further?
- What kind of **research or research-activism** do you think is needed to advance anti-hate speech work?

http://www.tinyurl.com/lesshatetrello

Thank you!

https://rajapinta.co/tag/hate-speech/

https://github.com/futurice/spice-hate_speech_detection

AoIR 2017, Tartu, Estonia