Platforms, Experts, Tools: Specialised Cyber-Activists Network

MAPPING STUDY

"Countering online hate speech with automated monitoring tools"

By LICRA (Ligue Internationale Contre le Racisme et l'Antisémitisme)

Project funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020)

Abstract:

This report aims at presenting, defining and analysing existing automated tools to monitor online hateful content as well as its limits beyond the spread of online hate speech across the Web 1.0 and 2.0. The report provides an in-deep analysis of ways in monitoring online hate speech that are available and straightforward for individual experts, non-profit organisations and Human rights activists. Automated intelligent technologies that provide to CSO's, decision makers and online activists a better environment in conducting monitoring project are at the heart of a new way to tackle any form of hate speech across the World Wide Web, including social media. In this mapping study, we will analyse which tools are available to individuals to counter hate speech and how to improve the prospects of removal of reported hate speech.

Keywords:

Artificial intelligence, civil society, counter speech, crawlers, data collection, hate ontology, hate speech, IT companies, monitoring, tools, social media, software.

About the Project

The EU-funded project **sCAN** – *Platforms, Experts, Tools: Specialised Cyber-Activists Network* (2018-2020), coordinated by Licra (International League Against Racism and Antisemitism), aims at gathering expertise, tools, methodology and knowledge on cyber hate and developing transnational comprehensive practices for identifying, analysing, reporting and counteracting online hate speech. This project draws on the results of successful European projects already realised, for example the "*Research, Report, Remove project: Countering Cyber-Hate phenomena*" and *"Facing Facts*", and strives to continue, emphasize and strengthen the initiatives developed by civil society for counteracting hate speech.

Through cross-European cooperation, the project partners will enhance and (further) intensify their fruitful collaboration. The **sCAN** project partners will contribute to selecting and providing relevant automated monitoring tools to improve the detection of hateful content. Another key aspect of **sCAN** will be the strengthening of the monitoring actions (e.g. the monitoring exercises) set up by the European Commission. The project partners will also jointly gather knowledge and findings to better identify, explain and understand trends of cyber hate at a transnational level. Furthermore, this project aims to develop cross-European capacity by providing e-learning courses for cyber-activists, moderators and tutors through the Facing Facts Online platform.

sCAN will be implemented by ten different European partners, namely ZARA – Zivilcourage und Anti-Rassismus-Arbeit from Austria, CEJI-A Jewish contribution to an inclusive Europe from Belgium, Human Rights House Zagreb from Croatia, Romea from Czech Republic, Respect Zone from France, jugendschutz.net from Germany, CESIE from Italy, Latvian Centre For Human Rights from Latvia and the University of Ljubljana, Faculty of Social Sciences from Slovenia.

The sCAN project is funded by the European Commission Directorate – General for Justice and Consumers, within the framework of the Rights, Equality and Citizenship (REC) Programme of the European Union.

Legal disclaimer

This mapping study was funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020).

Project funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020): Project reference number: JUST/REC-AG-2017/REC-RRAC-ONLINE-AG-2017/785774

Table of content

1. Introduction

1.1. Context

In the early 21st century, the democratization of Internet and e-communication support (mobile devices, tablets and computers) and technologies (Artificial intelligence, software and applications) has resulted in a rise of online content and virtual interactions between users. Internet has become a borderless space in which people can interact easier by not divulging their real identity. As it is hard to figure out an identification mapping of who are using the internet, users can freely express their feelings and opinions on any issue without risking to put their identity in danger and being prosecuted. Since the Internet is seen as a borderless space of freedom of expression, civil society, activists, political groups and organisations are able to produce and share diverse content to approve, criticize or promote value and ideas supported by other users and groups. The rising of social media platforms, also called the Web 2.0, like Facebook, Twitter, YouTube and Instagram has strengthened the movement of content. Millions of posts and tweets are published on a daily basis. As a result, the manifestation of hate online content spread out significantly amongst European countries in the last decade. If the Internet, including social media, has opened up new arenas for exchanging opinions, developing access to Freedom of speech, democratic value and foundation, both for the public and for the media, at the same time, hate speech is spread widely and frequently on new platforms. Hate speech may cause fear, violence and social conflict and as a consequence can be the reason why people withdraw from the public debate. Hate content aims at targeting vulnerable groups by producing hate speech against them. It has become crucial to understand how online hate in the public sphere especially could affect our democracies.

Ronald Eissens, general director of the INACH network, gives us a global overview on the evolution of the Internet mapping as he wrote in 2017 in the "Executive Foreword" of the project "Research – Report - Remove: countering cyber hate phenomena":

"*During the age of explorers, the maps that were produced all had white areas on them; terra incognita, unknown land. Invariably, there would be a drawing of a giant octopus or another fabled animal, with the caption 'Here be monsters'. Early internet users knew where those monsters were and their names […]. The contemporary net is no longer mappable for concentration-points of hate. Sure, there are new*

main boosters although they are relatively young […]. But online hate and incitement are now all over, embedded in blogs, postings, snaps, tweets, profiles, groups and grams"[1].

1.2. What is hate speech?

Since hate speech phenomena become more and more a threat for democratic systems and human rights values, it is crucial to develop new systems for monitoring, collecting and analysing online hateful content in order to counter it.

There is no universal definition for hate speech. Hate speech is a complex issue depending on a lot of factors: context, language, legal framework, etc. As a consequence, it is not easy to recognize and to define hate speech.

In the sCAN project, hate speech is defined as intentional or unintentional public discriminatory and/or defamatory statements; intentional incitement to hatred and/or violence and/or segregation based on a person's or a group's real or perceived race, ethnicity, language, nationality, skin colour, religious beliefs or lack thereof, gender, gender identity, sex, sexual orientation, political beliefs, social status, property, birth, age, mental health, disability, disease[2].

1.3. Challenges in countering online hate speech

We observe that the online content is moving faster through the internet and social media and could be extensively uncontrolled by human moderators because of the continuous information flow moving across platforms. Analysing hate speech has become a challenge as we see its diversity and impact. The main question we can raise so far is how to define a way to find out efficiently online hate speech through a massive place of data's movement. Building relevant counter-narratives for tackling online hate speech requires first to research, identify, collect and analyse online hate content. For NGOs, this work used to be manual or based on public online reporting forms. It involves considerable efforts and time for identifying hate speech and then taking measure such as reporting, removing, developing counter-speech or debunking. At the scale of the Internet, civil society's manual efforts for countering online hate speech look like a drop in the ocean. It has become necessary to develop new automated tools for improving this core work.

[1] INACH, Manifestations of Online Hate Speech (2017)
[2] INACH, Project Research-Report-Remove: Countering cyber-hate phenomena (2016-2018)

2. Legal framework and actions taken for countering hate speech

2.1. International and European legal framework on hate speech

Since people across nations and borders can exchange information and materials, decision-makers, public authorities and civil society organisations have raised the importance in establishing a legal framework to regulate online hate speech. European law-makers have to find legal solutions for solving obstacles caused by differences in the legal frameworks of countries where major Internet industry companies are hosting.

International human rights law recognizes explicitly the right of freedom of expression for each individual and groups and call the states to protect and promote these rights. It is however required that the states must prohibit severe forms of hate speech through criminal law and other legal frameworks in order to guarantee to its citizens a safe space of expression and equality.

The initiative of having a supranational legal framework regulating online hate speech was taken by the Council Europe. The Convention on Cybercrime is the first multilateral text aiming at countering computer-based crime by strengthening the cooperation between authorities amongst nations that have ratified the convention. However, the internet hate speech section of the convention has been removed to allow the U.S. to ratify the treaty. On 2010, the Council of Europe responded by adding an additional protocol of the Convention on Cybercrime calling the parties to "*criminalize acts of racist and xenophobic nature committed through computer systems*[3]". This has been signed by 32 member states and ratified by 15 without the U.S. support, arguing that the additional protocol is irrelevant with the U.S. First Amendment.

The consolidation of a legal framework within the European Union against cyber hate started with the first resolutions and decisions on racism and intolerance taken by the European Parliament and the Council of the European Union in the nineties. This framework introduces different levels of responsibility. They focused on discrimination in the field of employment and social affairs, approaches of educational systems[4]. The Tampere Council in 1999 and the European Parliament in 2000

[3] J. Banks, "Regulating hate speech online, International Review of Law, n°24, 2010.
[4] Joint Action 96/443/JHA: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31996F0443

considered that more steps needed to be adopted for fighting racism and xenophobia. The Hague programme was implemented in order to maintain action against racism, xenophobia and antisemitism.

The Framework Decision adopted in 2008 by the Council of the European Union was decisive for providing harmonization of laws and regulations of Member States with regard to offences involving racism and antisemitism. This decision constituted a crucial step forward for improving the legal European framework against hate speech but still it was only limited to racism, antisemitism and xenophobia.

The legal European framework to tackle cyber hate was the initiative of the European Parliament and of the Council of 8 June 2000 with the adoption of the Directive 2000/31/EC[5]. The e-Commerce Directive "*foresees that Internet intermediary service providers should not be liable for the content that they hold and transmit passively. At the same time when illegal content is identified, intermediaries should take effective action to remove it*" (European Commission, 2016). This Directive has led to the development of general take-down procedures. This type of procedure starts when a report or a notification is sent by someone to the hosting service provider concerned about illegal content and is concluded when the hosting service provider takes action against the illegal content, most of the time removal of the content.

Because of the explosion of online hate content and the powerful leading of a few IT companies in the last decade, in 2015, the European Commission without proposing any new law, insisted on them in "*need[ing] for clarification of the notice-and-action procedures*" and "*therefore undertak[ing] further analysis in the coming year to explore whether EU action is warranted in these areas*"[6].

As a results, in 2016 the central outcome of the EU Internet Forum between the Commission and Facebook, Microsoft, Twitter and Google/YouTube, was a new "Code of Conduct on Countering Illegal Hate Speech Online". Some of the main commitments are listed below:

- "*The IT Companies to have in place clear and effective processes to review notifications regarding illegal hate speech on their services so they can remove or disable access to such content. The IT*

[5] Directive 2000/31/EC "on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, "Directive on electronic commerce": https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32000L0031&from=EN

[6] European Commission, Digital single market, "Commission updates EU audiovisual rules and presents targeted approaches to online platform", 25 May 2016 : http://europa.eu/rapid/press-release_MEMO-16-1895_en.htm

companies to have in place Rules or Community Guidelines clarifying that they prohibit the promotion of incitement to violence and hateful conduct.

- *Upon receipt of a valid removal notification, the IT Companies to review such requests against their rules and community guidelines and where necessary national laws transposing the Framework Decision 2008/913/JHA, with dedicated teams reviewing requests.*
- *The IT Companies to review the majority of valid notifications for removal of illegal hate speech* ***in less than 24 hours and remove or disable access to such content,*** *if necessary.*
- *The IT Companies* ***to encourage the provision of notices and flagging of content*** *that promotes incitement to violence and hateful conduct at scale by experts, particularly via partnerships with CSOs [civil society organizations, S.E.], by providing clear information on individual company Rules and Community Guidelines and rules on the reporting and notification processes. The IT Companies to endeavour to strengthen partnerships with CSOs by widening the geographical spread of such partnerships and, where appropriate, to provide support and training to enable CSO partners to fulfil the role of a 'trusted reporter' or equivalent, with due respect to the need of maintaining their independence and credibility*".

In order to test the implementation of the Code of conduct by the IT Companies, the Commission with the help of civil society organisations specialised in hate speech organised monitoring exercises. Since May 2016, four exercises have been carried out to ensure and improve compliance with the Code of Conduct's commitments.

2.2. From manual moderation system to automated content moderation

Websites, forums and social media have strengthened interactions of users located in different places around the world. At the same time, the necessity to propose actions against illegal content increased with concerns raised by NGOs and states arguing that these contents would have an impact on individuals and public space. It has been a long – legal – struggle for European public authorities and

CSOs to persuade social media companies to develop moderation tools, include hate speech in their terms and conditions and take actions to respond to hate speech issues[7].

Ten years ago, Facebook moderators' team was composed of 12 people while 120 millions of users were registered on the platform. IT companies, as for example Twitter, refused to adapt to any European framework invoking the First amendment to the United States Constitution[8].

Nonetheless, in a decade, IT companies' attitudes have changed: moderation system as well as general take-down procedures have become the standard in Europe. Moderation systems or content moderation are methods used to monitor and regulate user-generated posts through implementing a set of pre-arranged rules and guidelines. Social media companies employ content moderators to manually inspect or remove content flagged by users as hate speech. There are different types of moderation systems: pre-moderation, post-moderation, distributed moderation, reactive moderation and automated moderation.

The web 1.0 and forums are more likely using the post-moderation method – and sometimes the pre-moderation method. Due to the amount of data traffic, IT Companies and social media are more able to develop reactive moderation relying on end-user judgment and action in reporting or flagging. Due to the large amount of reported content, the human moderation system shows limits as it cannot filter a hundred percent of the content produced by users on a daily basis. The system requires to have permanent moderators, 24 hours per day, 7 days a week, to moderate and attribute sanctions or not to any hate content reported. Moreover, moderator bias could also affect the moderation quality: "*While human moderators can account for context while moderating comments, their intrinsic biases may affect their decisions, leading to audience frustration. As a result, human moderation is not sustainable or scalable as your publication and community grows*[9]".

Social media companies are also notoriously secret about their content moderation practices – including the creation of manuals and the training of workers – so it's difficult to know much about who makes the content decisions. Even if Facebook created a new website dedicated to transparency

[7] For example, case Yahoo!, inc. v. Licra in 2000: French court ruled that Yahoo was liable and should seek to eliminate French citizen's access to the sale of Nazi merchandise: https://scholarship.law.berkeley.edu/cgi/viewcontent.cgi?article=1435&context=btlj

[8] First amendment to US Constitution: "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances".

[9] D. Seaman, "Human vs. Machine: The Moderation wars", Viafoura,: https://viafoura.com/blog/human-vs-machine-moderation-wars/

regarding the platform, it is quite difficult to have some data on hate speech. For example, in the "Community standards enforcement preliminary report" covering the period from October 2017 through March 2018, and including metrics on enforcement of hate speech, data of hate speech violations on Facebook are "*not available*". Facebook is "*unable to provide reliable data for this time period*"[10]. Nevertheless, Facebook is able to share data regarding moderated content: "*In Q1 2018, we took action on around 2.5 million pieces of content, up from around 1.6 million in Q4 2017. This increase is primarily due to improvements to our detection technology*". This "detection technology" which can be translated in Artificial Intelligence detection is crucial because it has an impact on Facebook's moderation system. Indeed, Facebook is also searching and flagging content before end users report it. On the same report, they note: "*The amount of content we flagged increased from around 24% in Q4 2017 because we improved our detection technology and processes to find and flag more content before users reported it".* In 2018, Facebook has again increased its content moderation team to 7, 500 employees. Even if AI seems to be the wave of the future, humans play more than a supporting role in moderation field[11].

IT companies and social companies tend progressively to use automated systems to sort and filter contents which match specific criteria written into the source code of the tool. Automated moderation run by specifically designed technical tools, using specific content moderation applications to filter hateful content. Detecting illegal content becomes automated and less time-consuming. IP addresses of users classified as abusive can also be blocked. However, the lack of human expertise may limit the monitoring process of such digital tools. Specific platforms are targeted by these tools, mainly for saving time and resources. Platform's administrators are called to set up the tool by putting in criteria to which the software will find out in an efficient way online hate content. The system would work 24 hours per day, 7 days a week without a human behind the computer. Relevant data picked up by the automated tool are sent to the moderator. Lastly, a team of moderator can review the data reported by the automated tool and provide adequate sanctions towards users who do not respect platform's term and conditions.

[10] Facebook, "Community Standards Enforcement Preliminary Report", Transparency Facebook, April 2018: https://transparency.facebook.com/community-standards-enforcement#hate-speech
[11] A. NG, "Inside Facebook, Twitter and Google's AI battle over your social lives", CNET, 17 July 2018: https://www.cnet.com/news/inside-facebook-twitter-and-googles-ai-battle-over-your-social-lives/

3. From the rise of automated tools monitoring internet content…

3.1. Concepts and categories of automated tools

A range of different tools are currently available in the internet market with various plans in regards to the technical support, functionality and capacity to scrape online data. Such tools which are not set up in any extraction process of online hate speech can be used if users configure the software and add specific requests to detect online hate speech through keywords. Nowadays, individuals and companies have developed various automated tools to monitor the Internet.

Regarding the web 1.0 and web 2.0, web scraping or web spidering are "*computer software technique[s] of extracting information from websites. This technique mostly focuses on the transformation of unstructured data (HTML format) on the web into structured data (database or spreadsheet)*"[12]. It is possible to use web crawler or web spiders which are "a *program or automated script which browses the World Wide Web in a methodical, automated manner*"[13]. These types of programs or software are used by search engines to automatically visit pages, to index them and to collect information or data: "*When a crawler visits a website, it picks over the entire website's content" (text, images, source code etc.)*[14]. Crawlers usually extract XML data into an excel sheet.

Beyond crawlers and automated programmes, Artificial Intelligence is now used for monitoring hate speech. Machine-learning and deep-learning based techniques are becoming more efficient in tracking written and video content. Artificial Intelligence is "*the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages*"[15]. There are several fields where Artificial Intelligence can be applied. First, expert systems are computer systems which can be programmed to make expert-decisions in real-life situations. The integration of machines, software, and specific information allows the system to impart reasoning, explanation, and advice user for solving complex issues. Furthermore, chatbots using Natural Language Processing (NLP), also known

[12] S. Rail, "Beginner's guide to Web scraping in Python (using BeautifulSoup)", Analytics Vidhya, 22 October 2015: https://www.analyticsvidhya.com/blog/2015/10/beginner-guide-web-scraping-beautiful-soup-python/
[13] Sciencedaily, Reference term, web crawler: https://www.sciencedaily.com/terms/web_crawler.htm
[14]Sciencedaily, Reference term, web crawler: https://www.sciencedaily.com/terms/web_crawler.htm
[15] Oxford dictionary, definition of artificial intelligence.

as talkbots, IM bots, interactive agents, or Artificial Conversational Entities, can recognize natural human language if communicating directly with a user and conduct a conversation via auditory or textual methods. Artificial Neural Networks (ANN), mostly used in machine learning, intend to simulate intelligence by brain-inspired systems which are intended to replicate the way that humans learn. ANN has become a major part of artificial intelligence. Machine learning is an application of artificial intelligence (AI) based on algorithms that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Deep learning describes algorithms attempting to model high level abstractions in data to determine a high level meaning.

3.2. Artificial Intelligence, IT Companies and hate speech

Mainstream social media such Facebook, Twitter, Google or Microsoft tend to complement and maybe further replace moderators by Artificial Intelligence, monitoring violations of their policy rules. AI has potential in countering the spread of hate speech. However, its development requires sufficient resources that individuals, NGOs or activists do not have as it supposes to have a team of data analyst, social scientists and developers. Moreover, results showed that AI can extract millions of data that require time and effort to sort, analyse and identify online hate speech across social media. Is artificial intelligence really relevant for countering hate speech? How do IT Companies improve their AI technologies programmes?

In 2018, Facebook published its report welcoming the positive feedback in using AI to counter online crimes and violence including hate speech and frauds. During the first three months of 2018, about 583 million of accounts have been deleted according to Guy Rosen, Facebook's VP of Product Management[16]. On the Facebook research website, the 2, 2 billion users platform is very clear on their use of "machine intelligence": "*Facebook Artificial Intelligence researchers seek to understand and develop systems with human-level intelligence by advancing the longer-term academic problems surrounding AI. Our research covers the full spectrum of topics related to AI, and to deriving knowledge from data: theory, algorithms, applications, software infrastructure and hardware infrastructure. Long-term objectives of understanding intelligence and building intelligent machines are bold and ambitious, and we know that making significant progress towards AI can't be done in isolation. That's why we actively engage with the research community through publications, open source software,*

[16] Setra, "En trois mois, Facebook a supprimé 583 millions de comptes fake », presse citron, 17 May 2018: https://www.presse-citron.net/en-trois-mois-facebook-supprime-583-millions-de-comptes-fake/

participation in technical conferences and workshops, and collaborations with colleagues in academia"[17].

Nonetheless, if child abuse or adult nudity images are easily detected by Facebook algorithms, it is much more difficult with hate speech detection: Mark Zuckerberg said it was "*one of the hardest*" problems. Facebook took down 21 million pieces of adult nudity and sexual activity during the period, 96 percent of which was flagged by its artificial intelligence. Regarding hate speech, the company admitted its artificial intelligence has a hard time finding hate speech. The company removed 2.5 million contents of hate speech content during the first quarter of the year 2018. However, only 38% of the hateful content were identified by its technology. Why is hate speech so hard to detect? CSOs are aware of the constant evolution and trends of hate speech, of the importance of context, of social events and of the evolving hateful language. The analysis of hate speech is crucial before starting monitoring hateful content. For now, it appears that AI technologies need to be taught about what hate speech is and need to be fed with thousands and thousands examples of hateful data: this means that human expertise is initially necessary in the process. Humans are also required after data collection by AI technologies in order to identify content of the "grey zone" as hate speech or not.

Since 2016, Google has started using AI to fight online hate speech[18]. Alphabet, Google's parent company, with its subsidiary company Jigsaw[19] has developed a project that focuses on key words in order to measure the "*perceived impact a comment might have on a conversation*" by assigning it a "toxicity" score. Regarding hate speech, the programme named "Conversation AI" has launched the project "Perspective" which is "*an API that makes it easier to host better conversations. The API uses machine learning models to score the perceived impact a comment might have on a conversation. Developers and publishers can use this score to give real-time feedback to commenters or help moderators do their job, or allow readers to more easily find relevant information, as illustrated in two experiments below. We'll be releasing more machine learning models later in the year, but our first model identifies whether a comment could be perceived as "toxic" to a discussion*"[20].

[17] Facebook, "Advancing the field of machine intelligence", Facebook AI Research: https://research.fb.com/category/facebook-ai-research/

[18] WNYC, "Google is Fighting Online Hate Speech With Artificial Intelligence", 13 October 2016: https://www.wnyc.org/story/how-google-fighting-online-hate-speech-through-ai/

[19] https://jigsaw.google.com/

[20] Alphabet, Perspective API, Jigsaw, 2016 :https://www.perspectiveapi.com/#/

Twitter said it would promote "healthy conversation" (in an attempt to avoid the term hate speech) by using a combination of human moderation and machine learning to detect trolls and minimize the appearance of their posts on the platform. Twitter has decided to support academics with a programme called "Measuring healthy conversation". Studies will be launched for understanding "the prevalence of some of the social media's most problematic content" in the form of "*peer-reviewed, publicly available, open-access research articles and open source software whenever possible*"[21]. One of the two projects selected, led at Leiden University in the Netherlands, will aim to create algorithms that can distinguish between "incivility" and "intolerance" in Twitter conversations in order to determine the prominence of echo chambers and uncivil discourse on Twitter[22].

3.3. Artificial Intelligence, Private companies and hate speech

Even though mainstream social media are on their way to developing tools countering online hate content, private companies have also developed a range of technology to get easier searches of specific online content which could be used at least for marketing and business plans and further proposed to companies in need of cleaning their own content[23].

Why are private companies interested in working in the area of hate speech? Why are they interested in cooperating with or even sometimes "helping" CSOs by providing otherwise expensive monitoring tools free of charge?

While IT companies continue to improve their AI tools and at the same time try to restore confidence in their platforms after many scandals, many start-ups and private companies are developing new automated tools using Artificial Intelligence to help fix new issues across the media and the Internet industry.

As an example, Factmata has developed a technology mixing artificial intelligence, algorithm and expert knowledge to deal with hate speech, propaganda, fake news and clickbait[24]. Based in London, Factmata proposes an anti-fake-news AI platform and services by providing a scoring system for the

[21] Twitter, "Measuring Healthy conversation", 2018: https://blog.twitter.com/official/en_us/topics/company/2018/measuring_healthy_conversation.html
[22] Twitter, "Measuring Healthy conversation", 2018: https://blog.twitter.com/official/en_us/topics/company/2018/measuring_healthy_conversation.html
[23] A. Oboler, "How technology can be used to combat online hate speech", World Economic Forum, The Conversation, 13 March 2018: https://www.weforum.org/agenda/2018/03/technology-and-regulation-must-work-in-concert-to-combat-hate-speech-online
[24] Factmata website: https://factmata.com/technology.html

content across the web, enhancing quality and credibility of textual content (e.g. articles, comments and user interactions) published through mainstream traditional mass media.

Factmata works mainly with newspaper companies such as The New York Times, Bloomberg or The Guardian and "*analyse[s] millions of new URLs daily, digging into individual articles and their actual content on a sentence by sentence level, to assess risk. This maximises the number of places people can safely advertise, while providing the best, granular protection available against deceptive content*"[25]. Factmata is supported by a community of users fact-checking or marking news articles for quality with the help of AI. As for social media, the young start-up needs also human background back-stopping in order to improve AI detection and results.

This start-up has perfectly understood how hate speech could be a threat not only for "public relations" but also constitute "a business and legal risk if not checked". In an interview with TechCrunch, Factmata's CEO and founder Dhruv Ghulati said: "*We are taking a long-term perspective on this. We think in five to ten years, will we have a new news platform that puts the user at its core? From the tech perspective, it is well known that this space has been dominated by social media platforms. That market is there, but there is a huge chunk that is not, and we think there is a huge opportunity to revamp safety in that market*.[26]"

Other companies such as Crimson Hexagon provide Artificial Intelligence services based on machine learning, image analytics, neural networks and natural language processing. This Company initially proposes commercial service to "*global brands [for] better understand[ing] their consumers*". It "*allows clients to analyse audiences, track brand perception and campaign performance, and even detect competitive and market trends*". But this service can also be quite relevant regarding hate speech monitoring. Some organisations from the Civil Society are already using this kind of AI technologies – for example, the OCCI (Online Civil Courage Initiative). Nonetheless, most of the time those solutions are not financially accessible for individuals, Human rights activists, local and national NGOs.

[25] Factmata website: https://factmata.com/technology.html

[26] I. Lunden, "Factmata closes $1M seed round as it seeks to build an 'anti fake news' media platform", TechCrunch, January 2018: https://techcrunch.com/2018/02/01/factmata-closes-1m-seed-round-as-it-seeks-to-build-an-anti-fake-news-media-platform/

4. …To the existing limits and challenges of automated monitoring tools

4.1. Financial and material obstacles

Most of automated software available on the market for free are not usually granting full access to users. A free software offers very limited functions which do not efficiently detect online hate content. In order to gain full and unlimited access to the functions offered by downloadable software (i.e. Octoparse, HTTrack), users have to subscribe and pay monthly.

Moreover, most of the advanced internet social media automated monitoring tools developed by IT companies are not available for individuals, NGOs and human rights activists because of the costliness of the software. The private market also proposes a large range of pre-coded internet monitoring tools which cost around 1,000€ to 5,000€.

Building Artificial Intelligence requires experienced team of developers with a high level of qualification, resources, long-term supervision and updating process to keep effectively the monitoring capacity for the detection while expectations from users are changing. AI development is expensive and not accessible to individuals, NGO or Human Rights activists.

However, across all current categories of automated monitoring tools, the detection of online hate speech has limited effectiveness. Each tool is effective only for a specific platform. The rising of social media platforms in addition to Facebook and Twitter (e.g. Instagram, Snapchat and YouTube) makes it necessary to keep developing AIs, crawlers or software exclusively targeting contents of the selected platform and requires significant human and financial resources.

4.2. Evading detection from automated monitoring system

A hate ontology – like that developed within this project – is required to guarantee the effectiveness of the tool in the detection and collection of terms used in online hate speech. A definition of keywords within the national context in which they are used through the Web 1.0 or 2.0 is an unavoidable step that would help to pick up accurate online hate speech. However, a word-based technology does not guarantee that every content will be scraped as users tend to develop strategies of diversion and bypassing of legal boundaries that regulate online content.

Many obstacles are still challenging the detection of hateful terms. For example we observe that users are using "coded languages" also called "offensive language" including abbreviation and emoji. They are more and more interested in a way to escape sanctions from automated scanning tools and moderators. On June 2018, Facebook sent an internal document to its moderators showing which emoji are being used as hate speech. In addition to the textual content, emoji's attachment is potentially an indicator of hate speech, sexual harassment and violence against minors as defined below[27].

Facebook's internal document sent to its moderators

Indicators of..	Emojis	Related IS
Condemnation		Hate Speech, Bullying, Graphic violence, Dangerous Individuals & Orgs
Attack, Harm, Call to Action		Dangerous Individuals and Orgs, Coordinating Harm, Bullying, Harassment, Hate Speech, Protection of Minors, Credible Violence
Sexual orientation		Bullying, Hate Speech
Exclusion		Hate Speech
Dehumanising comparison		Hate Speech
Hate speech (designated dehumanizing comparisons)		Hate Speech
Protected Characteristic	National flags	Hate Speech

Moreover, users are being aware that typing out a full word could expose them to sanctions from the moderation system. Therefore, strategies have been developed to replace any letter(s) by an asterisk in order to avoid detection from the automated scanning programme of the social online platform[28]. For example, users use harmful terms towards black, LGBTQI+ and others vulnerable people like h*e, f*g or b*tch.

The study "All you need is 'love': Evading hate speech detection", developed by the Aalto University (Finland) and the Padua University (Italy) and published in August 2018, proves that it is possible to

[27] S. Wolfe, "A leaked document shows which emoji Facebook associates with site violations", Business Insider France, 7 June 2018: http://www.businessinsider.fr/us/facebook-emojis-hate-speech-2018-6

[28] T. Davidson, D. Warmsley, M. Macy, I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language", ICWSM, March 2017: https://arxiv.org/pdf/1703.04009.pdf

evade different hate speech classifying algorithms simply by altering text, including Google's perspective AI made for detecting toxic comments and hate speech. During their research, contrary to human readers, AI detection was not working when the machine learners included the following in the text: new typos, use of leetspeak (changing some letters in numbers or symbols), adding extra or non-hateful words, insert and remove whitespaces between words. All seven hate-speech classifiers were significantly derailed by at least some of the researchers' methods. At the end of their research, they decided to combine the two most powerful approaches already tested: whitespaces removal and word appending – "love". The message remains totally understandable for humans, but the algorithms do not know what to do with it. The only thing they can really process is the word "love." According to New Scientists, on Google's perspective API "you're pretty smart for a girl" was deemed 18% similar to comments people had deemed toxic, whereas "i love Fuhrer" was only 2% similar[29].

Evading hate speech detection is possible not only by substituting different characters for letters or using symbols, but also by changing machine learning system itself. Indeed, most of the time algorithms based on machine learning are fed and trained on data gathered and generated by humans. As a results, the system can learn a different way than initially intended. In 2016, Microsoft developed a Twitter bot named "Tay", described as an experiment in "conversational understanding". People (mainly trolls) started tweeting the bot with hateful messages. The bot started repeating same kind of messages back to Twitter's users. Socialhax.com collected screenshots of tweets published by Tay before their removal. Many of the tweets saw Tay referencing Hitler, denying the Holocaust or supporting Trump's immigration plans (to "build a wall"). This phenomena of "data poisoning attacks" is possible when the machine is corrupted by hateful data.

Out of this issue of "data poisoning attacks", there also exists another risk caused by an "algorithmic bias". For one of the researcher of the study "All you need is 'love': Evading hate speech detection", Tommi Gröndahl: "*[…] hate-speech data sets are fine as long as we are clear what they are: they reflect the majority view of the people who collected or labelled the data*"[30]. The "algorithmic bias" is caused by several human biases. It "*occurs when AI and computing systems act not in objective fairness, but according to the prejudices that exist with the people who formulated, cleaned and structured their*

[29] D. Heaven, "This AI can tell true hate speech from harmless banter", NewScientists, 11 October 2017: https://www.newscientist.com/article/mg23631471-800-this-ai-can-tell-true-hate-speech-from-harmless-banter/

[30] L. Matsakis, "To break a hate-speech detection algorithm, try 'love'", Wired, 26 September 2018: https://www.wired.com/story/break-hate-speech-algorithm-try-love/

data"[31]. These human biases have effects on the development of Artificial Intelligence system from the creation of the algorithm to the interpretation of the collection of data. In a review published in June 2018 titled "*A review of possible effects of cognitive biases on interpretation of rule-based machine learning models*", a team of European experts identified 20 cognitive biases "*that can give rise to systematic errors when inductively learned rules are interpreted"*[32].

4.3. Data protection and Privacy

Collecting data from users can be controversial in term of personal information and private life recorded (e.g. place, age, friends and political opinions). The recent impact of Facebook-Cambridge Analytica shows clearly the danger of sharing data which would have a potential impact on real political activities and elections[33].

Since 2014, Facebook-Cambridge Analytica has collected around 80 million Facebook accounts. This breach had been used by politicians during the U.S Presidential Elections (2015), Brexit vote (2016) and Mexican Presidential Elections (2018).

In 2016, the European Union adopted the General Data Protection Regulation (GDPR) which is a regulation "on the protection of natural persons with regard to the processing of personal data and on the free movement of such data". In May 2018, this new European legal framework has been in place in all EU countries for ensuring strong data protection for European citizens. With this legislation, firms including IT Companies have to be explicit in their data collection for marketing purposes, have to ask specific permission for collecting and to offer consumers a specific reason for having the information. GDPR is crucial as it improves the protection of European citizens' rights regarding their personal data and clarifies what companies that process personal data must do to safeguard these rights. The first application of this legislation might be with Facebook because of the security breach which affected 30 millions of accounts. In October 2018, the DPC of Ireland declared: "*Facebook data breach. The DPC is concerned that this breach was discovered on Tuesday & affects millions of users. At present*

[31] M. Sears, "AI Bias and the people factor" in AI development", Forbes, 13 November 2018: https://www.forbes.com/sites/marksears1/2018/11/13/ai-bias-and-the-people-factor-in-ai-development/#36514a239134

[32] T. Kliegr, S. Bahnik, J. Fürnkranz, "A review of possible effects of cognitive biases on interpretation of rule-based machine learning models", 27 June 2018: https://arxiv.org/pdf/1804.02969.pdf

[33] K. Granville, "Facebook and Cambridge Analytica: What you need to know as fallout widens", The New York Times, March 2018: https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html

Facebook is unable to clarify the nature of the breach & risk to users. We are pressing Facebook to urgently clarify these matters. #dataprotection »[34]. The Company could face 1, 63 billion $ fine under GDPR legislation[35].

Recent restrictions on data privacy do not allow developers to create new tools which could retrieve data from the whole of Facebook. Most of Facebook's retrieving tools have been closed following the Cambridge-Analytica's scandal. However, it is still possible to capture posts, comments and accounts of specific public-pages and groups if users know the URL. Online hate speech could also be detected if a directory of Facebook pages and groups is made beforehand as a ground for automated monitoring tools analysing online hate speech. Social networks could be classified according to the publicity or the privacy of their content. It indicates how difficult or not a monitoring of public hateful content could be done. For example, Twitter which is rather a public social network would be easier to monitor compared to Snapchat which is private and ensures encrypted data.

Categorization of some social media

Rather public	Half-private	Private/ Encrypted
• Twitter • Youtube • Dailymotion • Pinterest • Reddit • RuTube	• Facebook • Instagram • Linkedin • VK.com	• Telegram • WhatsApp • Snapchat • Curiouscat • Discord • Facebook Messenger

Criteria:

* Rather public: Content is freely available to logged and unlogged users

* Half-private: Users can control the access of their personal information

* Private, Encrypted: Accounts, texts, photos and videos are encrypted

[34] A. Chérif, "Piratage,Que risque Facebook avec la RGPD?", La Tribune, 1 October 2018 : https://www.latribune.fr/technos-medias/piratage-que-risque-facebook-avec-le-rgpd-792325.html

[35] O. Solon, "Facebook faces $1.6bn fine and formal investigation over massive data breach", The Guardian, 3 October 2018: https://www.theguardian.com/technology/2018/oct/03/facebook-data-breach-latest-fine-investigation

5. Case-study: sCAN project

In the sCAN project, one of the main goals is to select and provide relevant automated monitoring tools to improve the detection of hateful content. Some criteria have been chosen regarding the choice of automated tools before their testing, in two phases. The research aims at focusing on available and inexpensive tools with a comprehensive interface in order to improve the monitoring process and collection of data. In this part, we will explain some researches and testing initiated by the project. Most NGOs do not have an IT desk or coding skills. In order to help them with monitoring, automated tools have to be simple, easy to use and less time-consuming than manual monitoring. Another key-point is that automated tools have to be used in many different languages, including different alphabets.

5.1. Examples of automated tools selected

The following automated monitoring tools have been chosen for the first testing campaign of the project. They have been selected in order to match the expectations of NGOs and activists wanting a first-hand tool for detecting online content without technical expertise. The following list is of course non-exhaustive.

The following tools were selected for the first testing campaign:

SociScraper is a free software developed by individual users. It scrapes data from Twitter, Instagram and YouTube with various search settings and sorting options which allow users to extract data by keywords, number of followers, likes and views.

TAGS v 6.1 is a crawler dedicated to Twitter. The crawler allows users to set up and run automated collection of search results on Twitter. It scrapes data by keywords which are recorded in a Google Drive Excel sheet. The excel sheet offers information related to the account, tweet, date and sometimes other type of information if the user made it public.

Inteltechniques is a crawler for Facebook developed by Michael Bazzell. The crawler scrapes public posts released by Facebook users or pages, public images and public videos. Users can set up the crawler through the insertion of keywords amongst Facebook public posts.

Getting data from the Web 1.0 is also possible with scraping tools downloading web material and pages into a computer directory. The first step before monitoring the Worldwide Web is to carry out an

identification and listing of potential websites in which hidden hateful content can be found out by the software. Then, using keywords allows the tools to scrape specific data that users will further sort and analyse. Amongst the Web 1.0, there are some ways in scrapping and recording data:

Search Engine: Most platforms containing social interactions, such as online newspapers and forums, include search engine options which users can use to search for specific content by date and keywords.

Scrapy is an application framework operating with Python for crawling web sites and extracting structured data which can be used for a wide range of useful applications, like data mining, information processing or historical storage[36].

HTTrack is a free software allowing users to download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer.

5.2. Methodology based on hate ontologies

A part of our project is to list relevant keywords in each language related to hate categories based on our analysis. Examples will be available on the project's blog http://www.scan-project.eu. These keywords are chosen because of their significant use in hate speech on a specific context of each country. In order to select these keywords, each organisation part of the project had to identify users' trends – in terms of words, theories, rhetoric, etc. - in expressing hate towards individuals and groups. Hate speech analysis is crucial for identifying these keywords. Because of the evolving nature of hate speech, the analysis is ongoing and keywords have to be updated in accordance with social events, national and European context.

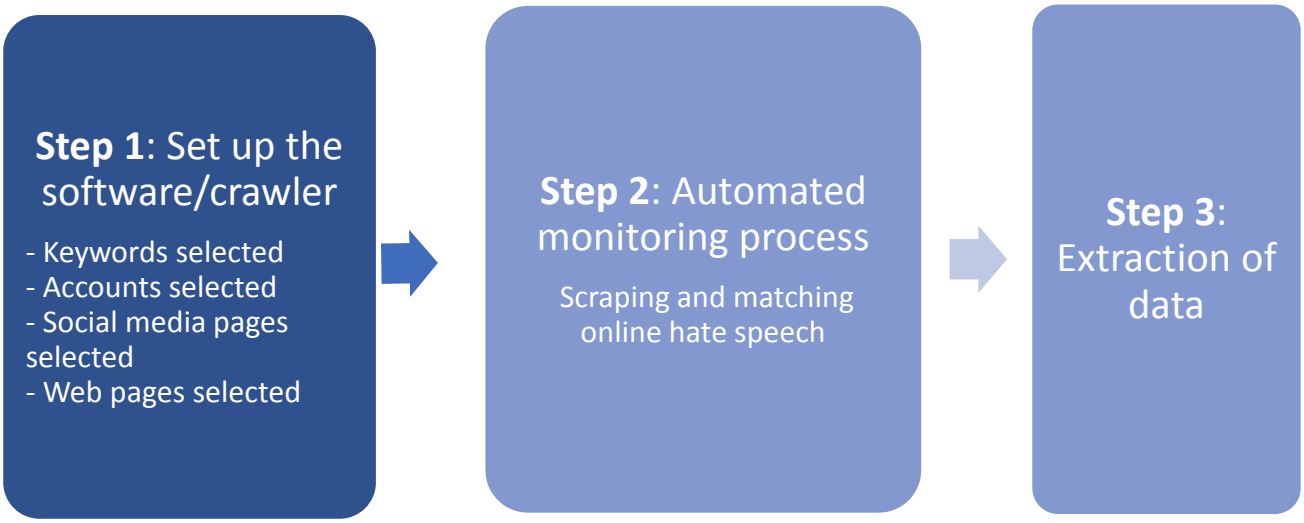
Consensus has been reached within the framework of the sCAN project in the definition of categories that facilitate the collection of relevant content by hate categories of keywords such as racism, antisemitism, anti-Muslim hatred, anti-Refugee hatred, antigypsyism, anti-LGBTQI+ hatred, etc.

Why using keywords in monitoring with automated tools?

[36] Python is "an interpreted, object-oriented, high-level programming language with dynamic semantics". Created in the late 1980s it allows programmers to use different programming styles to create simple or complex programs, get quicker results and write code almost as if speaking in a human language (Technopedia, definition, https://www.techopedia.com/definition/3533/python).

The choices of keywords constitute a significant step after the selection of tools to explore the Internet and its content. For many crawlers, software and algorithms selected it is necessary to include keywords to start the monitoring process. In TAGS.6.1, for example, the crawler can select tweets by searching a word or a collection of words.

The web 1.0 crawlers need to be fed with a list of websites or blogs due to the hugeness of the Internet. For scraping data on the entire Internet, using a basic search engine including the advanced search tools and plug-in options seems to be the best option.



5.3. Extraction, collection and sorting process: challenges and limits

Due to the large amount of data generated on a daily basis, monitoring online hate speech is a hard work for individuals and cyber-activists wanting to contain the spread of hateful content through the Internet. It requires to cover several platforms at the same time.

Automated tools aim at making the collection of data easier and faster for those who carry out a long-term monitoring strategy with limited human and financial resources by saving time during the selection and sorting process. Afterwards, the definition of which content might be qualified as hate speech is an exhausting task needing to be improved by automated monitoring tools.

LICRA (International League against Racism and Antisemitism) and jugenschutz.net have carried out a testing campaign to see how automated scraping tools work in the detection of online hate speech through the Web 1.0 as well as social media featuring in the European Commission's Monitoring Exercises such as Facebook, Twitter, Instagram and YouTube. We observe that the activists looking for online hate speech spend most of their time filtering the selected content into an excel sheet. For example, if the tool scrapes over 400 results, the sorting process may take about one hour and half per keyword and platforms[37].

How can we improve this data collection step?

The methodologies of text mining and data mining are relevant for data analysis. Text mining or text analytics is the process of deriving relevant information from text data. Text mining is the process of exploring, analysing and structuring large amount of unstructured text data helped by software that can identify concepts, topics or keywords in the selected data. This methodology is nowadays easier to use due to the development of deep learning algorithms that can analyse large amount of unstructured data. Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Text mining is similar in nature to data mining, but with a focus on text instead of more structured forms of data. However, one of the first steps in the text mining process is to organize and structure the data in some fashion so it can be subjected to both qualitative and quantitative analysis[38]. These methodologies are helpful in order to transform text or data in a structured data: encoding process is crucial for extracting relevant information in order to start data analysis.

Many free and open-source text mining and data mining tools are available, for example Carrot2 which can automatically organize small collections of documents (search results but not only) into thematic categories. Another web mining module is Pattern for the Python programming language. It is relevant for hate speech monitoring due to available tools for data mining on Google, Twitter and Wikipedia API, and is a web crawler[39].

[37] sCAN Project, Testing campaign realized from 2nd October 2018 to 16th October 2018.
[38] Searchmetrics glossary, "The dictionary of search engine optimization and content marketing" Searchmetrics
[39] Top 27 free software for text analysis, text mining, tex analytics, PAT research: https://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/

Social network analysis tools are software which facilitates the quantitative or qualitative analysis of social networks by describing features of a network either through numerical or visual representation. It generally uses network or graph theory to examine social structures. The main components are nodes (people) and the edges that connect them[40].

It is possible to have access to free social network analysis tools as for example "Social Network Visualizer" which is a cross-platform tool that allows to draw social networks on a virtual canvas. Either load field data from a file (in supported format) or crawl the internet to create a social network of connected webpages[41].

[40] Social network analysis software. Retrieved from: https://en.wikipedia.org/wiki/Social_network_analysis_soft-ware

[41] Rankred, 22 free social networks analysis tools: https://www.rankred.com/free-social-network-analysis-tools/

6. Conclusion

The Web 1.0 and social media have strengthened interactions of users located all around the world. Internet free speech space has also become a vector in the spread of hate speech. In the last decade, main IT companies' attitudes have changed regarding hateful content: from a first amendment philosophy to the development of moderation standards and take-down procedures. One of the solution for countering hate speech is to provide NGO's and individual activists with capacity-building in monitoring the Internet, including social media, by proposing easy-to-use and efficient tools for the collection of data by means of keywords that are regularly updated in regards to the national context and events.

Nowadays, there are a range of automated monitoring tools for the detection of online hate speech with various plan in regards to the technical support, functionality and capacity to scrape online data. This paper aims to present tools, methods and programmes developed so far. Beyond web spidering, Artificial Intelligence and its multiple fields are now used for monitoring hate speech. Social media companies tend to develop Artificial Intelligence as a new technology to moderate online content in an efficient way. Start-ups and private companies are also involved in developing new technologies for tackling hate speech. However, these ways are not easily available to CSOs or online activists: human and material challenges are still necessary. Moreover, many obstacles are still challenging the detection of hateful content. Evading hate speech detection is possible not only by corrupting and coding text, but also by attacking a machine learning system itself.

The sCAN project aims at focusing on available and inexpensive tools with a comprehensive interface in order to improve the monitoring process and collection of data. Human contribution in terms of knowledge and analysis is still crucial in monitoring hate speech. Testing campaigns will be developed during the 2 years project for testing and integrating the most efficient automated tools whilst taking into account the existing barriers and limitations of intelligence technology.

7. Bibliography

ALPHABET, Perspective API, JIGSAW, 2016: https://www.perspectiveapi.com/#/

J. BANKS, "Regulating hate speech online", International Review of Law, n°24, 2010.

A. CHERIF, "Piratage, Que risque Facebook avec la RGPD?", La Tribune, 1 October 2018 : https://www.latribune.fr/technos-medias/piratage-que-risque-facebook-avec-le-rgpd-792325.html

T. DAVIDSON, D. WARMSLEY, M. MACY, I. WEBER, "Automated Hate Speech Detection and the Problem of Offensive Language", ICWSM, March 2017: https://arxiv.org/pdf/1703.04009.pdf

EUROPEAN COMMISSION, Digital single market, « Commission updates EU audiovisual rules and presents targeted approaches to online platform, 25 May 2016: http://europa.eu/rapid/press-release_MEMO-16-1895_en.htm

FACEBOOK, "Community Standards Enforcement Preliminary Report", Transparency Facebook, April 2018: https://transparency.facebook.com/community-standards-enforcement#hate-speech

FACEBOOK, "Advancing the field of machine intelligence", Facebook AI Research: https://research.fb.com/category/facebook-ai-research/

K. GRANVILLE, "Facebook and Cambridge Analytica: What you need to know as fallout widens", The New York Times, March 2018: https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html

T. GRÖNDAHL, L. PAJOLA, M. JUUTI, M. CONTI, N. ASOKAN, "All you need is 'Love': evading hate speech detection", Aalto University (Finland) and University of Padua (Italy), August 2018: https://arxiv.org/pdf/1808.09115.pdf

D. HEAVEN, "This AI can tell true hate speech from harmless banter", NewScientists, 11 October 2017: https://www.newscientist.com/article/mg23631471-800-this-ai-can-tell-true-hate-speech-from-harmless-banter/

I. LUNDEN, "Factmata closes $1M seed round as it seeks to build an 'anti fake news' media platform", TechCrunch, January 2018: https://techcrunch.com/2018/02/01/factmata-closes-1m-seed-round-as-it-seeks-to-build-an-anti-fake-news-media-platform/

INACH, International Network Against Cyber Hate, "Manifestations of Online Hate Speech", 2017: http://www.inach.net/project-research-report-remove-countering-cyber-hate-phenomena/

INACH, International Network Against Cyber Hate, Project Research-Report-Remove: Countering cyber-hate phenomena, 2016-2018: http://www.inach.net/manifestations-of-online-hate-speech/

T. KLIEGR, S. BAHNIK, J. FÜRNKRANZ, "A review of possible effects of cognitive biases on interpretation of rule-based machine learning models", 27 June 2018: https://arxiv.org/pdf/1804.02969.pdf

S. KÖFFER, D. M. RIEHLE, S. HÖHENBERGER, J. BECKER, "Discussing the value of automatic hate speech detection in online debates", University of Münster, Research Centre for Information Systems, March 2018:
https://www.researchgate.net/publication/325367326_Discussing_the_Value_of_Automatic_Hate_Speech_Detection_in_Online_Debates

MANDOLA PROJECT, "Mandola Monitoring Dashboard", September 2016: http://mandola-project.eu/publications/

L. MATSAKIS, "To break a hate-speech detection algorithm, try 'love'", Wired, 26 September 2018: https://www.wired.com/story/break-hate-speech-algorithm-try-love/

A. NG, "Inside Facebook, Twitter and Google's AI battle over your social lives", CNET, 17 July 2018: https://www.cnet.com/news/inside-facebook-twitter-and-googles-ai-battle-over-your-social-lives/

A. OBOLER, "How technology can be used to combat online hate speech", World Economic Forum and the Conversation, 13 March 2018: https://www.weforum.org/agenda/2018/03/technology-and-regulation-must-work-in-concert-to-combat-hate-speech-online

POSITIVE MESSENGERS to counter online hate speech project, "Media content analysis on online hate speech, national report, Italy", 2017: https://positivemessengers.net/en/

S. RAIL, Beginner's guide to Web scraping in Python (using BeautifulSoup), Analytics Vidhya, 22 October 2015: https://www.analyticsvidhya.com/blog/2015/10/beginner-guide-web-scraping-beautiful-soup-python/

H. M. SALEEM, K. P DILLON, S. BENESCH AND D. RUTHS, "A web of hate: Tackling Hateful speech in online social spaces", School of computer science, McGill University, Montreal, School of Communication, The Ohio State University, Ohio, Berkman Centre for Internet and Society, Harvard University, Massachusetts:
https://www.researchgate.net/publication/320163517_A_Web_of_Hate_Tackling_Hateful_Speech_in_Online_Social_Spaces

D. SEAMAN, "Human vs. Machine: The Moderation wars", Viafoura:
https://viafoura.com/blog/human-vs-machine-moderation-wars/

SEARCHMETRICS GLOSSARY, “The dictionary of search engine optimization and content marketing”, Searchmetrics: https://www.searchmetrics.com/glossary/

M. SEARS, “AI Bias and the people factor” in AI development”, Forbes, 13 November 2018: https://www.forbes.com/sites/marksears1/2018/11/13/ai-bias-and-the-people-factor-in-ai-development/#36514a239134

O. SOLON, “Facebook faces $1.6bn fine and formal investigation over massive data breach”, The Guardian, 3 October 2018: https://www.theguardian.com/technology/2018/oct/03/facebook-data-breach-latest-fine-investigation

SETRA, “En trois mois, Facebook a supprimé 583 millions de comptes fake », presse citron, 17 May 2018: https://www.presse-citron.net/en-trois-mois-facebook-supprime-583-millions-de-comptes-fake/

TWITTER, “Measuring Healthy conversation”, 2018: https://blog.twitter.com/official/en_us/topics/company/2018/measuring_healthy_conversation.html

WNYC, Google is Fighting Online Hate Speech with Artificial Intelligence, 13 October 2016: https://www.wnyc.org/story/how-google-fighting-online-hate-speech-through-ai/

S. WOLFE, “A leaked document shows which emoji Facebook associates with site violations”, Business Insider France, 7 June 2018: http://www.businessinsider.fr/us/facebook-emojis-hate-speech-2018-6