Publications of the Ministry of Justice Finland

Reports and guidelines

2021:19

Utilisation of artificial intelligence in monitoring hate speech

MINISTRY OF JUSTICE FINLAND

Publications of the Ministry of Justice, Reports and guidelines 2021:19

Utilisation of artificial intelligence in monitoring hate speech

Laura Kettunen, M.A.
Mari-Sanna Paukkeri, D.Sc. (Tech.), Utopia Analytics

Ministry of Justice, Finland, Helsinki 2021

Julkaisujen jakelu
Distribution av publikationer

Valtioneuvoston julkaisuarkisto Valto

Publikations-arkivet Valto

julkaisut.valtioneuvosto.fi

Julkaisumyynti
Beställningar av publikationer

Valtioneuvoston verkkokirjakauppa

Statsrådets nätbokhandel

vnjulkaisumyynti.fi

Publication sale

Online bookstore of the Finnish Government

vnjulkaisumyynti.fi

Publication distribution

Institutional Repository for the Government of Finland Valto

julkaisut.valtioneuvosto.fi

This report was produced as part of the Facts against Hate project, which was funded by the European Union's Rights, Equality and Citizenship Programme (2014–2020).

The authors assume full responsibility for the contents of the publication, which do not necessarily represent the views of the European Commission or the Finnish Ministry of Justice.

Ministry of Justice, Finland

ISBN pdf: 978-952-259-811-0
ISSN pdf: 2490-0990

Layout: Government Administration Department, Publications
Helsinki 2021 Finland

Description sheet
30 August 2021

Utilisation of artificial intelligence in monitoring hate speech

Publications of the Ministry of Justice, Reports and guidelines 2021:19		**Subject**	Reports and guidelines
Publisher	Ministry of Justice, Finland		
Authors	Laura Kettunen, Mari-Sanna Paukkeri		
Language	English	**Pages**	44

Abstract

The report is produced by project Facts Against Hate coordinated by Ministry of Justice. One of the objectives of the project is to pilot the use of artificial intelligence (AI) in monitoring hate speech. The aim of the monitoring is to gain an overall picture of hate speech. Goals include gaining an understanding of the channels in which hate speech occurs and discerning differences in hate speech on the different platforms.

The report presents the findings of AI-assisted hate speech monitoring. The approach employed was a combination of human analysis and machine learning.

The dataset for the report consisted of around 12 million comments and online posts from September–October 2020. According to the findings, hate speech as defined in this report was detected in around 150,000 messages a month, or in 1.8% of total messages, on public Finnish-language online platforms. Over the two-month period reviewed, 1 September – 31 October 2020, a total of 298,032 hate speech messages were identified and 97% of these were detected in various discussion forums. The next most common platform for hate speech messages was Twitter (2.5%). The dataset does not include closed groups and private accounts on Facebook.

The authors assume all responsibility for the contents of the publication. The contents do not necessarily represent the views of the Ministry of Justice of Finland or the European Commission, which is funding the Facts Against Hate project.

Keywords	hate speech, artificial intelligence, AI, racism, harassment		
ISBN PDF	978-952-259-811-0	**ISSN PDF**	2490-0990
URN address	http://urn.fi/URN:ISBN:978-952-259-811-0		

Kuvailulehti
30.8.2021

Tekoälyn hyödyntäminen vihapuheen seurannassa

Oikeusministeriön julkaisuja, Selvityksiä ja ohjeita 2021:19		**Teema**	Selvityksiä ja ohjeita
Julkaisija	Oikeusministeriö		
Tekijä/t	Laura Kettunen, Mari-Sanna Paukkeri		
Kieli	englanti	**Sivumäärä**	44

Tiivistelmä

Raportti on tuotettu oikeusministeriön koordinoiman Tiedolla vihaa vastaan -hankkeen toimesta. Hankkeen yhtenä tavoitteena on pilotoida tekoälyä vihapuheen seurannassa. Seurannan tavoitteena on saada kokonaiskuva vihapuheesta. Tavoitteena on muun muassa muodostaa käsitys siitä, millaisissa kanavissa vihapuhetta esiintyy ja millaisia eroja eri alustoilla esiintyvässä vihapuheessa on.

Raportissa esitellään tuloksia tekoälyn avulla tehdystä vihapuheen seurannasta. Lähestymistapana on käytetty ihmistyön ja koneoppimisen yhdistelmää.

Raportin aineistona oli noin 12 miljoonaa suomenkielistä kommenttia ja nettikirjoitusta syys–lokakuulta 2020. Tulosten perusteella tämän raportin määritelmän mukaista vihapuhetta esiintyy julkisilla suomenkielisillä alustoilla verkossa noin 150 000 viestiä kuukaudessa, eli 1.8 prosenttia kaikista viesteistä. Kahden kuukauden tarkasteluajanjaksolla 1.9.–31.10.2020 tunnistettiin 298 032 vihapuheviestiä, joista 97 % esiintyi erilaisilla keskustelupalstoilla. Seuraavaksi yleisin alustatyyppi oli Twitter (2.5 %). Facebookin suljetut ryhmät ja ei-julkiset tilit eivät ole mukana aineistossa.

Julkaisun sisällöt ovat täysin tekijöiden vastuulla, eivätkä ne välttämättä edusta Tiedolla vihaa vastaan -hanketta rahoittavan Euroopan komission tai oikeusministeriön näkemyksiä.

Asiasanat	vihapuhe, tekoäly, rasismi, häirintä		
ISBN PDF	978-952-259-811-0	**ISSN PDF**	2490-0990
Julkaisun osoite	http://urn.fi/URN:ISBN:978-952-259-811-0		

Presentationsblad
30.8.2021

Utnyttjande av artificiell intelligens vid uppföljning av hatretorik

Justitieministeriets publikationer, Utredningar och anvisningar 2021:19		**Tema**	Utredningar och anvisningar
Utgivare	Justitieministeriet		
Författare	Laura Kettunen, Mari-Sanna Paukkeri		
Språk	engelska	**Sidantal**	44

Referat

Rapporten har producerats inom ramen för projektet Fakta mot hat som koordineras av justitieministeriet. Ett av projektets mål är att testa artificiell intelligens vid uppföljningen av hatretorik. Syftet med uppföljningen är att få en helhetsbild av hatretoriken. Målet är bland annat att bilda en uppfattning om i vilka kanaler hatretorik förekommer och vilka skillnader som finns i hatretoriken på olika plattformar.

I rapporten presenteras resultaten av den uppföljning av hatretorik som gjorts med hjälp av artificiell intelligens. Som utgångspunkt användes en kombination av mänskligt arbete och maskininlärning.

Rapportens material består av cirka 12 miljoner finskspråkiga kommentarer och inlägg på nätet från september till oktober 2020. Enligt resultaten är antalet meddelanden som uppfyller definitionen på hatretorik enligt denna rapport på offentliga finskspråkiga onlineplattformar cirka 150 000 per månad, dvs. 1,8 procent av alla meddelanden. Under den två månader långa granskningsperioden 1.9–31.10.2020 identifierades 298 032 hatmeddelanden, varav 97 procent förekom på olika diskussionsforum. Den näst vanligaste plattformen för hatmeddelanden var Twitter (2,5 procent). Facebooks slutna grupper och konton ingår inte i materialet.

Författarna ansvarar helt och hållet för publikationens innehåll, och innehållet representerar nödvändigtvis inte synpunkterna hos Europeiska kommissionen, som finansierar projektet Fakta mot hat, eller justitieministeriet.

Nyckelord	hatretorik, artificiell intelligens, rasism, trakasseri		
ISBN PDF	978-952-259-811-0	**ISSN PDF**	2490-0990
URN-adress	http://urn.fi/URN:ISBN:978-952-259-811-0		

Contents

Summary

This report is produced by project Facts Against Hate coordinated by Ministry of Justice. The objective of the project is to develop the monitoring of hate speech by piloting new tools that specifically monitor online hate speech.

The project has tested the possibilities of artificial intelligence to recognize hate speech in the online environments. The approach was to combine human evaluation with machine learning. The goal, among others, was to understand what are the main channels of hate speech and what kind of differences there are in the hate speech published on different online platforms.

The definition of hate speech was based on academic research in the field of social sciences. In the definition work hate speech categories were produced, which were then used to manually identify examples of hate speech from data. These annotations were used as training data for Utopia AI Moderator, a language-independent tool that utilizes text analytics and machine learning. The data set consisted of circa 12 million comments and posts in Finnish from September to October 2020.

The results show that there are, according to the definition of this report, about 150 000 hate speech messages published on the Finnish publicly available social media platforms every month, about 1.8% of all messages. During the two months' analysis period from 1 September to 31 October 2020, a total of 298 032 hate speech messages were identified, out of which 97% appeared on various discussion forums. The next largest platform types are different kinds of Twitter messages (2.5%) and Instagram messages (0.2%). Blogs, news comments and messages on public Facebook cover less than 0.02% of all identified hate speech. The data set does not include private Facebook groups or accounts.

Ylilauta.org seems to be the most significant platform for hate speech (285 000 messages, that is 96% of all messages identified as hate speech). The second largest volume of hate speech is on twitter.com (7 450 messages), Suomi24.fi (1 931 messages), hommaforum.org (1 600 messages) and vauva.fi (796 messages).

The data set analysed in the report opens a possibility to examine the potentially growing hate speech arenas. Among the public international social media platforms Twitter seems to be the most prominent with 7 450 messages identified as hate speech, 0.14% of all tweets. The results show that retweets play a significant role in the circulation of hate speech messages: 39% of all tweets identified as hate speech are duplicates.

Out of the hate speech analyzed further by artificial intelligence, 62% was labeled as being contemptuous or stigmatizing for a group, 32% as an insult or other expression of hate, 4% as hate expressions related to individual's characteristics, and 2% as hate expression towards a professional group.

The report also reviews the themes of hate speech. A hate speech vocabulary analysis reveals that the most common word in the data is *muslimi* (English *muslim*). It appears in 26% of the hate speech identified by the AI. Other common words were *islam, neekeri, homo* and *suvakki, poliisi, huora, terroristi, somali, ählämi ja rasisti*.

It is worth noticing that a small part of the users seem to produce the majority of the hate speech. The 10 most common usernames in the hate speech messages produce approximately 11% of all identified hate speech. The single most active author of hate speech in this data set is a Twitter account that published 352 tweets identified as hate speech during the analysis period.

1 Introduction

This report was produced as part of the Facts against Hate project coordinated by the Ministry of Justice, which aims for more effective work against hate crimes and hate speech. One of the project's objectives is to develop the monitoring of hate speech by testing new tools for targeted monitoring of online hate speech.

For the purposes of this report, Utopia Analytics tested the potential of an AI-assisted moderation product as a tool for identifying and monitoring hate speech in digital environments. A one-off monitoring exercise was carried out in autumn 2020.

The aim of the project was to examine online hate speech across a broad front and to build a picture of the platforms on which hate speech occurs, the differences between hate speech found on different platforms, and the potential links between hate speech in online environments and events of the real world.

Research ethical guidelines were followed when analysing the data. Research must not violate the protection of privacy or harm the participants in any way. Our data set consists exclusively of messages that were marked as public by the posters on social media platforms. The examples used in the report are extracts of posts included in the data set, in which names and usernames have been removed. In some cases, the language of the message was corrected just enough to ensure that a short example can be understood on its own in the same way as in the context of the entire message.

2 Defining hate speech

The findings discussed in this report are based on a definition of hate speech formulated specifically for this project and for online discussions. The definition relies on prior studies and expert consultations. To begin with, main classes and subcategories of different hate speech types were formed, after which finishing touches were put on the definition by manually annotating social media messages as hate speech of a specific category and non-hate speech. This made it possible to test the effectiveness of the classification with authentic messages.

The goal was to produce a definition that is as comprehensive as it can be when classifications of this type are used. The definition was not influenced by conceptions of what artificial intelligence can or cannot learn. Consequently, formulating a definition of hate speech and applying it to a social media data set to manually identify hate speech was one of the project tasks. Training artificial intelligence by means of manually compiled examples was another, separate task.

Rather than commenting on how any other party should define hate speech, this report welcomes discussion on the subject and considers further research important.

2.1 Premises

Hate speech is not an offence category but describes a group of acts of a certain type. Hate speech refers to communication that spreads or incites hatred against a person or group of people for a reason related to the person. The classification of hate speech presented in this report is based on definitions used in Finnish research, the harassment provision in the Non-Discrimination Act (section 14), the discrimination provision in the Act on Equality between Women and Men (section 7), the provision on ethnic agitation in the Criminal Code (section 10), and a report produced by the Office of the Prosecutor General (2012) on spreading punishable hate speech.

Hate speech may constitute an offence under the Criminal Code of Finland, discrimination prohibited under the Non-Discrimination Act or the Act on Equality between Women and Men, or some other expression that is generally harmful. It may take the form of written or other communication. The definition of hate speech used in this report is broader than the definition of punishable hate speech prohibited under the Criminal Code, as the authors also wished to take other legislation into account. The authors' purpose is not to label certain messages as punishable, as this is a task for the law enforcement authorities. Neither does the report comment on whether an individual expression is a violation of law.

In this report, hate speech means communication whose meaning or tone is degrading, humiliating, threatening, hostile, aggressive or dehumanising (for example, comparing people to animals or parasites). The communication may be related to personal characteristics or stigmatise a group of people. Personal characteristics include a person's age, language, appearance, religion or belief, gender, sexual orientation, ethnic background or physical functional capacity. The expressions may also be directed at individuals because they are presumed to belong to a national, ethnic, religious, sexual or other group. (Knuutila et al. 2019, Report of the Office of the Prosecutor General 2012).

The Committee of Ministers of the Council of Europe defines as hate speech all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred. This includes intolerance expressed by aggressive nationalism and discrimination and hostility against minorities, migrants and people of immigrant origin. According to the International Committee on Antisemitism, online hatred comprises any racism, anti-Semitism, religious extremism, homophobia or other phobia related to sexual orientation, narrow-mindedness against persons with disabilities, political hatred, spreading of rumours, gender-related hatred, violent pornography, promotion of terrorism, cyber bullying, harassment and persecution, expressions aiming to silence anyone speaking back (including shaming, defamation and name-calling) as well as speech that stigmatises groups spread on any electronic device. (Pöyhtäri 2015)

Hate speech can be used in an attempt to influence decision-making. It may target professional groups, persons working for a cause or a group, or persons in the public eye because of their profession or for some other reason. If the criticism against a representative of a professional group is exclusively levelled at their role as a professional, this is not regarded as hate speech in the report. If the criticism is directed at their personal characteristics or a group of people they represent, however, this is considered hate speech. (Knuutila et al. 2019) Prior studies have found that professional groups exposed to hate speech include politicians and decision-makers, journalists, researchers, police officers, prosecutors and judges. (See e.g. Hiltunen 2017, Association of Finnish Lawyers 2019, Committee for Public Information 2015, Pöyhtäri, Haara & Raittila 2013).

Hate speech typically sets and maintains a tone of language in which it is acceptable to judge certain individuals, minorities, nationalities, cultures, ethnic groups or religions as being of lesser value than the speaker or some others because of their (alleged) characteristics, or to attempt to destroy them. In addition, the issue of power relations and exercise of power is essentially linked to hate speech, and it is important to ask who speaks and what their position is. (Pöyhtäri 2015) While a post may be neutral, the context in which it appears may actually mean that the speaker is targeting a person with hatred.

Online shaming can additionally be interpreted as hate speech in some cases (Korpisaari 2019). In its broadest sense, hate speech also includes trolling and doxing (sharing someone's personal data online). The term hate speech has also been used in the context of cyber hate and cyber violence as well as toxic speech. (Laaksonen et al. 2020). The definition used in this report does not distinguish between shaming, trolling or doxing, however.

2.2 Ambiguous expressions

Hate speech may not be angry in its emotional content, or even express strong emotions. It may be neutral and calm in style, despite its intention to degrade or stigmatise. (Knuutila et al. 2019). Consequently, the context is essential in many cases when assessing whether an expression is hate speech or not.

This is why recognising hate speech is not always straightforward. In case of ambiguous expressions, not only the semantic content of the text but also its style and context have been assessed in this report to try and grasp the writer's intention. For the purposes of this assessment, the essential question is whether the author's intention is to argue a point and justify a position, or to intentionally stigmatise or degrade. For example, justifying a political opinion and the argumentation related to it have not necessarily been interpreted as hate speech, even if the position or opinion the argumentation concerns could be seen as discriminatory. Criticism levelled at political opinions or other ideologies has not been interpreted as hate speech, even if it were vulgar in style. In difficult and unclear cases, the definitions in the harassment provision of the Non-Discrimination Act, the discrimination provision in the Act on Equality between Women and Men, the provision on ethnic agitation in the Criminal Code, as well as the Office of the Prosecutor General's report on hate speech have been used as guidelines.

Under the harassment provision in section 14 of the Non-Discrimination Act, the deliberate or de facto infringement of the dignity of a person is harassment if the infringing behaviour relates to a reason referred to in section 8 of this Act (age, origin, nationality, language, religion, belief, opinion, political activity, trade union activity, family relationships, state of health, disability, sexual orientation or other personal characteristics), and as a result of the reason, a degrading or humiliating, intimidating, hostile or offensive environment towards the person is created by the behaviour.

Under the discrimination provision in section 7 of the Act on Equality between Women and Men, sexual harassment means unwanted conduct (of a sexual or non-sexual nature) by which a person's psychological or physical integrity is violated intentionally or factually, in particular by creating an intimidating, hostile, degrading, humiliating or offensive

atmosphere. On the other hand, an expression of opinion where a certain group is threatened, defamed or insulted on the basis of its race, skin colour, birth status, national or ethnic origin, religion or belief, sexual orientation or disability or a comparable basis, comprises ethnic agitation.

The Office of the Prosecutor General's report (2012) takes a stand on the type of speech that can be identified as hate speech or racism prohibited under law. On this basis, a practice often applied on discussion forums is to prohibit defamation and content that violates the protection of privacy, among other things. Legislation-based guidelines for moderating online content include respect for human dignity and a prohibition of inciting violence. (Pöyhtäri 2015, Report of the Office of the Prosecutor General 2012). In this project, posts which are insulting based on one of the reasons specified in the cited provisions were also interpreted as hate speech (for example, the words *autisti* or *vammainen* [English *autistic, disabled*] used as an insult). In the case of online posts, it is not always possible to know whether an insult is related to a real characteristic of the target. However, insults of this type present a specific group of people in a degrading light and normalise degrading speech associated with the group in question.

2.3 Hate speech definition

In this report, hate speech is defined as follows: Hate speech means degrading, humiliating, threatening, hostile, aggressive or dehumanising expressions that

1. are related to personal characteristics

a. age
b. gender, gender identity or expression of gender
c. sexual orientation
d. ethnic background (skin colour, origin, language)
e. religion or belief
f. political opinion
g. physical functional capacity
h. appearance
i. nationality
j. other

2. stigmatise or generalise a group of people based on their

a. age group
b. gender, gender identity or gender expression
c. sexual orientation
d. ethnic background (skin colour, birth, language)

e. religion or belief
f. political opinion
g. physical functional capacity
h. appearance
i. nationality
j. other

3. target a representative of a professional group, focusing on personal characteristics rather than their professional role, or a group of people the professional represents

a. politicians and decision-makers
b. public servants
c. journalist
d. prosecutors and law enforcement authorities, judges, police officers
e. researchers, experts
f. public personalities, social media influencers
g. persons working for a cause or a group
h. other.

4. Other expressions which, based on their intention or context, can be interpreted as

a. being one of the above, or
b. inciting or persuading others to take discriminatory action or use violence against an individual or a group of persons, or considering such actions acceptable

3 Data sets

The report analyses social media comments collected on the public Internet. The larger data set was purchased from an external service provider, Mohawk Analytics, which uses crawlers for the public Finnish-language social media and stores messages appearing on the major sites, sometimes within seconds after publishing. This is referred to as the **'*Social media data set*'** in the report. Another smaller set was collected manually on Ylilauta and Hommaforum sites, which block crawlers. This data set is referred to as the **'*Ylilauta and Hommaforum data set*'**.

By looking at the *Social media data set*, a comprehensive picture of the Finnish-language social media can be obtained. The material covers the largest discussion forums and news sites, many blogs and, naturally, discussions on the large international social media giants' platforms. The data was collected in almost real time, and it contains other information in addition to the post itself, including a link to the original message, information about the site and the site type, the time stamp of the post, the time the post was collected and, in some cases, information about the poster, a previous message or a thread. This information is already available while the discussion is still active.

Despite the many advantages of the *Social media data set*, it also has its limitations in the type of analysis conducted for this report. The key limitation is that the material lacks social media messages not available on the public Internet. For example, it does not include messages posted in closed Facebook groups or comments on private Facebook accounts. In addition, the report does not include WhatsApp messages, dating site content, and forums requiring login maintained by different communities.

Automated collection of messages using crawlers also sets certain limitations: the crawlers that collect the messages are configured manually, and some sites may simply be missing from the crawler's collection lists. The source code of some sites also prohibits crawlers. Additionally, a crawler may accidentally skip a part of the site and fail to collect the messages contained in this section. When messages are collected through interfaces, such as Twitter's and Facebook's, it is necessary to specify separately the users whose messages are regarded as the Finnish-language social media as well as the languages which can exist in the messages that are included in this set. This report does not discuss in detail the proportion of the various social media services that the data comprises or the share of the messages posted on each site that is included in the data set.

In addition, the *Social media data set* has features related to the time at which the posts were published and crawled that are essential in terms of the report's findings: sites subject to premoderation (including many news commenting sites) moderate the most unpleasant messages before they are published, whereas sites using post-moderation (including discussion forums) initially publish all comments and only moderate them later. Consequently, this data set presumably contains a large number of comments which the crawler has collected before they were removed from the site. However, the data contain no information about whether or not the comments were deleted on the site.

The *Social media data set* does not contain certain discussion forums considered particularly significant and prominent in terms of hate speech, including Ylilauta and Hommaforum. The source code of these sites prohibits crawlers. A small sample of messages posted on Ylilauta and Hommaforum was collected manually for this project. Collecting posts manually is slow, and collecting all messages posted within the time period of this study was not possible. Manual collection was the only way to collect these posts, however, without breaching the sites' terms of use. Scaling the analysis of a small, manually collected sample to all messages on the site creates a source of error in the findings. For example, it is possible that the sample is randomly extracted from threads that contain more or less hate speech than the rest of the site on average. It should also be noted that the total numbers of posts on these sites are estimates. However, some uncertainty is also associated with the number of messages in the *Social media data set*.

Comprehensive analysis of social media messages is challenging in general. The numbers of messages and websites discussed in this report are large, however, which is why we may assume that they contain the greatest part of the Finnish-language social media, and that their content includes most types of hate speech.

3.1 Millions of Finnish posts and online messages

Finnish-language social media contents were studied using the *Social media data set*,. The data set contains 11,975,002 messages, most of them in Finnish. The average number of messages is around 196,000 per day and 6 million a month. The data set includes the site platform types. For their distribution, see the following Table.

Table 1. The *Social media data set* collected on public Finnish social media platforms between 1 September and 31 October 2020 by platform type. (*) Does not include Ylilauta and Hommaforum.

Platform type	Posts	Posts/month
Twitter	5,478,192	2,739,096
Discussion forum*	2,596,214	1,298,107
Instagram	2,482,086	1,241,043
Facebook (public)	774,747	387,374
News comment	591,767	295,884
Blog	51,595	25,798
Other	401	201
Total	**11,975,002**	**5 987 501**

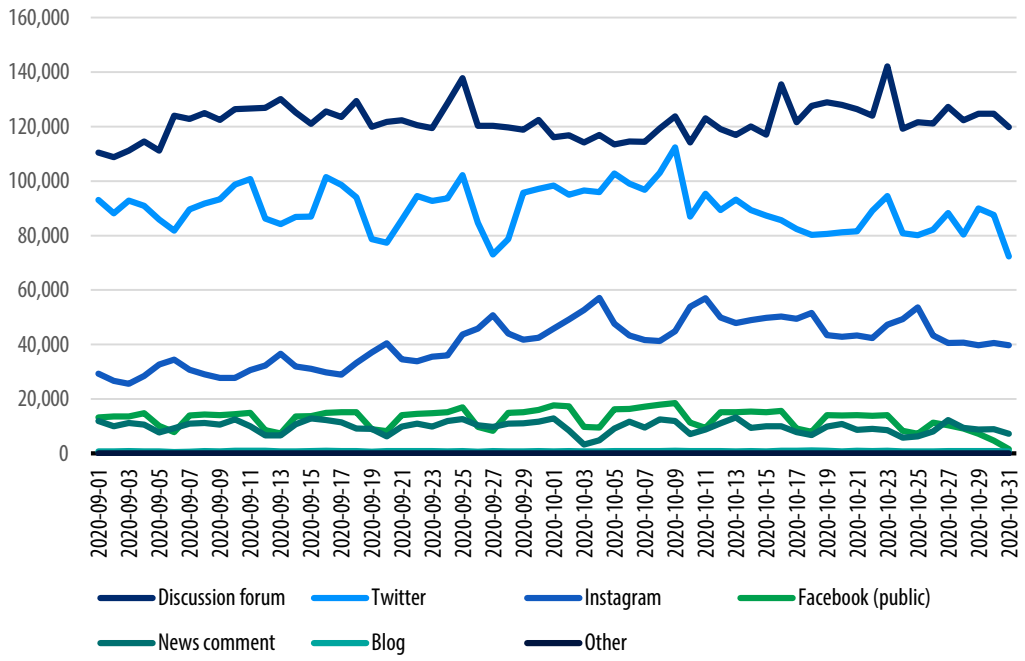
3.2 Ylilauta and Hommaforum

Unlike the other platforms discussed in this report, two sites considered important in terms of hate speech, Ylilauta and Hommaforum, prevent data crawling on their sites, and are consequently not included in the *Social media data set*. Thus messages posted on Ylilauta and Hommaforum were additionally collected manually in January 2021. The manual collection started from the beginning of the thread list of the site to select threads which contained more than ten messages posted during the review period, and from threads in the most frequently read topic areas. The aim was to obtain as comprehensive a sample as possible of the messages on the sites. A total of 3,205 messages were collected manually from Ylilauta from the categories Body and sex, Games, Spectator sports, Miscellaneous, YouTube and streams, and Society and politics. On Hommaforum, 343 messages posted under the categories Kylänraitti, Mylly and Tupa were collected. Ylilauta homepage links to the site's statistics, according to which 79,000 messages were posted a day during the period covered by the study in September and October 2020. On Hommaforum the age of the 100th newest message was monitored, which produced the average of about 300 messages a day in January 2021. While the number of messages published in September and October 2020 may have been different, this is likely to be the correct order of magnitude.

3.3 Data set distributions

The following Figure shows the distribution of messages between different types of social media platforms during the study period.

Figure 1. Distribution of messages between different types of social media platforms between 1 September and 31 October 2020.



Weekly fluctuations can be clearly seen in the data: especially on public Facebook, where the content is largely produced by companies and associations, most messages are posted on working days, whereas the weekends are quieter. On the other hand, the number of Instagram messages goes up on weekends. While the number of messages on Twitter is often smaller at weekends, there also are some exceptions.

The Finnish-language public social media is divided between a few large platforms, including Twitter, Instagram and Facebook. In total, the data set contains 187 websites, including Finland's largest discussion forums ylilauta.org, vauva.fi and suomi24.fi. The following table shows all platforms in the data set on which there was more than 2,000 messages in September–October 2020.

Table 2. Sites with more than 2,000 messages between 1 September and 31 October 2020 and their platform types. (*) The number of posts on Ylilauta is based on the platform's own statistics, while the number on Hommaforum is an estimate.

Platform	Platform type	Messages
twitter.com	Twitter	5,478,192
ylilauta.org*	Discussion forum	4,818,000
www.instagram.com	Instagram	2,482,085
www.vauva.fi	Discussion forum	1,100,589
www.facebook.com	Facebook (public)	774,747
keskustelu.suomi24.fi	Discussion forum	600,271
www.is.fi	News comment	258,010
www.iltalehti.fi	News comment	212,414
forum.hevostalli.net	Discussion forum	166,261
www.reddit.com	Discussion forum	150,441
www.demi.fi	Discussion forum	97,330
www.hs.fi	News comment	70,050
keskustelu.kauppalehti.fi	Discussion forum	42,974
puheenvuoro.uusisuomi.fi	Blog	35,511
yle.fi	News comment	33,815
kaksplus.fi	Discussion forum	32,855
www.tiede.fi	Discussion forum	30,693
futisforum2.org	Discussion forum	24,991
matkapuhelinfoorumi.fi	Discussion forum	24,841
murha.info	Discussion forum	24,775
www.punkinfinland.net	Discussion forum	20,342
tappara.co	Discussion forum	19,322
maanpuolustus.net	Discussion forum	18,434
hommaforum.org*	Discussion forum	18,300
keskustelu.pakkotoisto.com	Discussion forum	16,629
vesabbs.com	Discussion forum	15,718
www.masinistit.com	Discussion forum	12,510
www.autostadium.fi	Discussion forum	11,677
www.fillarifoorumi.fi	Discussion forum	11,332
www.supervuoro.com	Discussion forum	11,075

Platform	Platform type	Messages
foorumi.hifiharrastajat.org	Discussion forum	9,943
www.ilvesfoorumi.com	Discussion forum	9,237
ask.fm	Blog	9,003
www.btcf.fi	Discussion forum	8,335
lampopumput.info	Discussion forum	6,990
murobbs.muropaketti.com	Discussion forum	6,874
konekansa.net	Discussion forum	6,857
www.metsalehti.fi	Discussion forum	6,207
hopeinenomena.fi	Discussion forum	6,163
www.vrcf.fi	Discussion forum	5,449
www.kotiverstas.com	Discussion forum	5,405
www.ts.fi	News comment	5,129
forum.ylikerroin.com	Discussion forum	5,047
www.aamulehti.fi	News comment	4,998
nyymichan.fi	Discussion forum	4,804
keskustelu.anna.fi	Discussion forum	4,655
www.uusisuomi.fi	News comment	4,380
yhteiso.elisa.fi	Discussion forum	4,186
paakallo.fi	Discussion forum	4,033
hymy.fi	Discussion forum	4,013
muusikoiden.net	Blog	3,880
uusi.keskustelukanava.agronet.fi	Discussion forum	3,803
www.digicamera.net	Discussion forum	3,659
www.mersuforum.net	Discussion forum	3,627
kitina.net	Discussion forum	3,084
opelclubfinland.fi	Discussion forum	3,037
forums.offipalsta.com	Discussion forum	2,835
tuki.dna.fi	Discussion forum	2,707
www.nesretro.com	Discussion forum	2,442
paihdelinkki.fi	Discussion forum	2,421
www.koripallo.com	Discussion forum	2,376
www.lily.fi	Blog	2,338
www.perhokalastajat.net	Discussion forum	2,216

4 Methodology

The application of a theoretical definition of hate speech to practice is one of the key outcomes of this report. An approach based on collaboration between a human researcher and a machine learning system was selected, in which the researcher manually annotates individual social media messages as hate speech and non-hate speech. These examples produced by the researcher were used as training data for a machine learning system, a context-sensitive AI tool developed for identifying inappropriate posts. The AI learns to identify hate speech and is able to process a volume of posts infinitely larger than what could be analysed using human resources. The advantage of this approach compared to identifying hate speech based on automated word lists, for example, is that the researcher can meticulously keep to the predefined hate speech categories, regardless of the individual words contained in the posts.

The definition of hate speech categories was not influenced by ideas of what artificial intelligence can or cannot learn, and the categories were formed independently of the AI tool. As the AI tool used in the project is based on statistical modelling and therefore requires a sufficient volume of training data, in this project the AI was trained to identify whether or not a comment contains hate speech in general, as well as to recognise the four main categories in which a reasonable amount of training data was obtained. A more detailed analysis and classification into subcategories was carried out manually.

In this project, the AI tool was trained to only assess the content of the message and its semantic meaning. The selected AI tool learns to recognise the semantic context of the text; consequently, its operating logic is different from modelling based on the occurrence of individual words. On the other hand, the training data does not contain the surface level context, including who wrote the message or on which platform the message was posted. The training data consists exclusively of data in text format, and images, videos or links are thus disregarded in manual annotation and the AI modelling. This results in modelling that is as fair as possible for different platforms and posters: what counts is the semantic content of the message.

It would also have been possible to include the message preceding the analysed message (to which it responds) in the modelling. In some cases, the previous message may play an important role in interpreting a new one. The information on the previous message was not available in all cases in this data set, however, for which reason it was excluded from the analysis. In the future, it would be interesting to also examine how taking the textual context into account would change the results.

4.1 Annotation

The first messages analysed by a researcher were filtered from a large volume of messages based on the word list which the Police University College uses when drawing up its annual hate crime report. As additional filters were used words related to certain events of the real world which occurred during the period covered by the study and which could be expected to elicit hate speech (such as the Pride Week). The word list was only used in the first phase to facilitate manual annotation by a researcher, rather than as part of the AI model. Using the word lists, a much higher number of messages classified as hate speech came up for annotation by the researcher than what annotation in random order would have produced. This represented an effort to not only maximise the volume of AI training data but also draw the researcher's attention to the topics around which most hate speech revolves right at the beginning of the project. Of the messages filtered using the word lists, 8.7% were labelled as hate speech in the annotation phase.

In the final set identified as hate speech by the AI tool, 66% of the messages had originally come up when filtered by the word list. On the other hand, 6.9% of messages identified as non-hate speech had also come up using the same word lists. This indicates that while the word lists worked well enough as a preliminary device and improved the efficiency of annotation, they did not have a significant impact on the final AI model for recognising hate speech. The same word lists have also been used in earlier studies, as a result of which a significant number of relevant words has been accumulated on them. It should be noted, however, that the use of word lists may influence the project outcome, and some areas of hate speech, perhaps new ones, may have been overlooked when compiling the word list.

4.2 Practical choices and interpretations

During the annotation process, certain posts containing particularly degrading words were categorically interpreted as hate speech. These words included such degrading words referring to ethnic background as *neekeri, jutku,* ähläm̈i [English *nigger, kike, Dune coon*] and gender or sexual orientation, including *huora, hintti, hinttari* [*whore, fag, faggot*]. Some words in certain subject areas could appear not only in a degrading context but also in a less offensive or positive one. For example, the word *homo* [*gay*] is used in both a positive tone and a degrading and stigmatising sense. When this word came up, an attempt was made to interpret the intention of the poster and the tone of the message.

Due to their historical meaning, the context also influenced the interpretation of some words referring to ethnic background and nationality, including *ryssä* and *mustalainen* [*Russki, Gypsy*] in this report. *Mustalainen* is a term that has earlier also been used by the authorities in Finland, and while it no longer is acceptable in official contexts, it also

appeared in the material in contexts that were interpreted as non-degrading. An example of a comment of this type, in which the poster was not regarded as having a degrading intention and the post was not categorised as hate speech, could be:

> *"Being blonde [hair, skin or eyes] is a recessive property. […] I remember in primary school in Eastern Finland in the 1970s, there were no pupils with dark hair or brown eyes. There was one in the higher comprehensive, though, and this pupil's father was a Gypsy."*

A post including the word *ryssä* was not interpreted as hate speech if it appeared in the context of war history or referred to Russia as a global political actor. As an example, the following would not have been interpreted as hate speech on these grounds:

> *"They did their job. The Russki did not take our country."*

The material contains a large number of posts which were in the grey area. In these cases, it is impossible to say unequivocally if the author intends to degrade and stigmatise, or if they primarily seek to argue for or against a particular political position or opinion. In some cases, a different interpretation could also have been justified from the one made in this project.

4.3 Notes on the definition

The following observations, which affected the categorisation, were also made while annotating the material:

1. The definition did not provide a clear category for expressions that, rather than referring to the actual characteristics of any individual or group, use ethnicity, gender, sexual orientation, etc. as an insult. In the annotation process, such comments were put in category 4 (Intention or insult). See also section 5.4. Examples: *"huora", "tuommoista jutkupaskaa"* [*"whore", "that kike shit"*].

2. One of the original categories in the definition was dropped in the annotation stage. The excluded category would have included degrading and stigmatising posts *targeting an individual because they are assumed to belong to a group of people*. This definition of hate speech is relevant in court cases related to discrimination, for example. In the categorisation of posts, the difference between these posts and those placed in category 1 (Personal characteristics) was not clear enough, which is why this category was excluded from the definition.

3. Hate speech relating to a profession was divided between categories 2 (Group of people or generalisation) and 3 (Professional group). Generalised degrading comments about representatives of a profession were put in category 2. On the other hand, hate speech which targets a representative of a professional group but which aims the criticism at personal characteristics not related to the professional role were put in category 3.

4. In the subcategories of messages identified as hate posts, Muslims were in many cases referred to as an ethnic rather than a religious group. In the annotation of these cases, the criticism was interpreted as being associated with ethnic background (d). Comments referring to Islam or some other religion, or explicitly to religious questions, were interpreted as belonging to category e (religion or belief).

4.4 Utopia AI Moderator

Utopia AI Moderator is a language-independent SaaS solution based on text analytics and machine learning which can learn a specific site's moderation policy. In this project, it was used to identify hate speech as defined by the researcher. Hate speech is much less common and more serious than the inappropriate content that Internet service providers often remove from their websites. As it occurs less often, hate speech is more difficult to identify for an artificial intelligence tool based on statistical modelling than more commonly occurring inappropriate content. Because of the seriousness of hate speech, however, it is easier for AI tools like Utopia AI Moderator to learn about it, as hate speech differs semantically to a significant degree from other social media comments.

Utopia AI Moderator does not itself determine what is hate speech or inappropriate content and what is not. It always relies on human decisions made by the social platform in question about acceptable or inappropriate content. Utopia AI Moderator assumes that each social platform has determined its own terms of use and clearly informed its users of what kind of content is acceptable on the site and what is not. Utopia AI Moderator also presumes that human moderators have acted consistently in their task, ensuring that the AI does not receive conflicting information about what types of content should be accepted or rejected. In practice, there is always some variation in human decisions, which makes the consistency of the AI valuable.

Utopia AI Moderator consists of machine learning models developed by Utopia, a SaaS interface, and regular updates and maintenance of the machine learning models. Companies that use Utopia AI Moderator monitor particularly those messages which the system has labelled as borderline cases, or content of a new type. Through these

human decisions, the system constantly receives more training data, and the machine learning models maintained and developed further by Utopia's AI experts are kept up to date following the customer companies' moderation policies as they take shape and are gradually updated.

Utopia AI Moderator was originally developed for the Finnish language, but as a language-independent tool, it can be used to moderate user generated content in all languages. The product understands the informal language of social media without being confused by misspellings, everyday language expressions, foreign-language words or phrases, amusing emojis or any other features not found in grammar books or dictionaries. It is also context sensitive. With sufficient and consistent training data, Utopia AI Moderator learns to understand any comments that human moderators have understood.

4.5 Modelling

Utopia AI Moderator was used to perform two different tasks in the project: firstly, to produce data sets for the researcher to speed up the identification of messages that contain hate speech. Once the researcher had manually annotated a sufficient number of messages for statistical analysis, they were used as training data for the final Utopia AI Moderator model.

The final training data for Utopia AI Moderator comprised 18,925 messages, of which 2,471 were labelled as hate speech by the researcher, while the rest did not contain hate speech as defined in this report. In addition, 1,902 messages had been annotated, placed in hate speech categories and used to train another machine learning model to identify these categories.

5 Results

The AI learned to recognise hate speech based on the training data annotated by the researcher, despite the fact that the volume of the data was rather small and categorising the posts was a challenging task even for a human. Based on a manually annotated test set which was not included in the training data, the detection accuracy of the AI was 98.6%. The AI also arranged the messages based on the level of certainty at which it identified them as hate posts. In some cases, the AI labelled a post differently from the human annotator. Posts 'incorrectly' identified as hate speech with a high level of certainty were almost without exception borderline cases. It was consequently found that even a moderately sized set of training data was sufficient to quickly implement a highly accurate AI model that can automatically process immense volumes of social media posts in real time and label the most obvious cases of hate speech. The artificial intelligence model trained in the project can be used to identify hate speech in social media posts. The accuracy of the AI model is also sufficient for comparing different websites. If the Utopia AI Moderator model trained is this project were put in production use, its findings would be assessed by professionals of hate speech recognition. They would regularly compare a small number of decisions made by the AI tool to the selected definition of hate speech and use their assessments as new training data for the AI system. Consequently, the training data would build up and the accuracy of the AI model would improve as the product is in use.

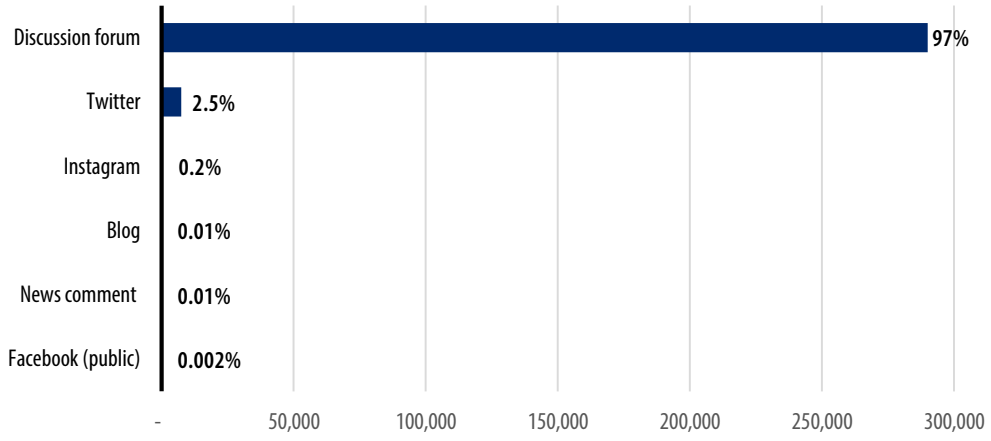
The smaller data set collected on Ylilauta and Hommaforum sites was not used to train the AI model in this project. Had the manually collected set been larger in size, posts on these websites could also have been used for training purposes. In other words, artificial intelligence was not applied to analysing the *Ylilauta and Hommaforum data set*. These posts were ultimately all annotated manually, similarly to the part of the *Social media data set* used to train the AI. This also ensured that the volume of hate speech calculated on the basis of a small sample was as accurate as possible when scaled to the larger data set.

In manual annotation, 190 out of the 3,205 posts collected from Ylilauta, or 5.9%, were identified as hate speech. On Hommaforum, this proportion was 30 out of 343 posts (8.7%). These percentages of the estimated total volumes of the sites would mean approx. 285,000 hate posts on Ylilauta and approx. 1,600 on Hommaforum during the time period covered by the study between 1 September and 31 October 2020. In particular, the large volume of hate speech on Ylilauta raises questions about the comparability of these results with the sites on which the *Social media data set* was collected. As both Ylilauta and Hommaforum have prohibited automated collection of content, manual sampling of this type is practically the only way to assess the volume of hate speech on the sites. In the *Social media data set*, the proportion of posts annotated as hate speech by a human researcher was less than 1%, whereas it was 5.9% on Ylilauta and 8.7% on Hommaforum. Even if the estimated amounts were not exactly correct, we can confidently say that the proportion of hate speech on these websites is significantly higher than on other Finnish social media platforms.

5.1 Hate speech platforms

The data sets used for the study indicate that the volume of hate speech as defined in this report on public Finnish-language online platforms is approx. 150,000 posts a month. Over a review period of two months between 1 September and 31 October 2020, 298,032 hate posts were identified, 97% of which appeared on different discussion forums. The next most common platforms were Twitter (2.5%) and Instagram (0.2%). Blogs, news comments and the public Facebook contain less than 0.02% of all identified hate speech. When looking at the results, it should be remembered that the data do not include private groups or accounts on Facebook (see Chapter 3). This report completely overlooks hate speech in Facebook's private groups in practice, and we may presume that Facebook's role as a platform of hate speech is considerably larger than what this data set indicates. The percentage of hate speech (298,032 posts) out of all 16.8 million messages is 1.8%.

Figure 2. Distribution of messages identified as hate speech between different public social media platform types. The figure for Facebook only includes the share of the public Facebook.



The most significant discussion forum containing hate speech is Ylilauta (estimated 285,000 posts, 98% of posts identified as hate speech on discussion forums, 96% of posts identified as hate speech on all platforms). The platforms that contain the next largest number of hate speech, however significantly less than Ylilauta, are suomi24.fi (1,931 posts), Hommaforum (estimated 1,600 posts) and vauva.fi (796 posts). The proportion of hate speech among all posts on the platform was 8.7% on Hommaforum, 5.9% on Ylilauta, 1.3% on murha.info and 0.62% on nyymichan.fi. The average proportion of hate posts published on these discussion forums is 3.9%. This is mainly explained by hate speech on Ylilauta and Hommaforum. Without these two platforms, the proportion of hate speech on the discussion forums would be 0.13%.

All sites where at least five hate posts were identified in this report are listed in the following table.

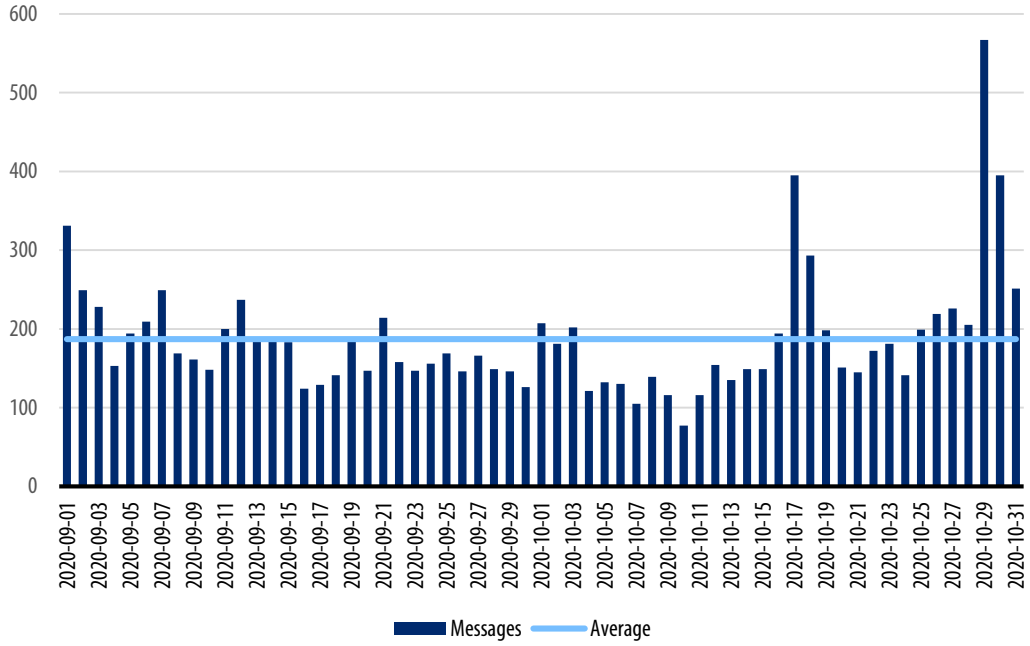
Table 3. Messages identified as hate posts on different platforms and their proportions of posts on each platform. (*) The total proportions of hate posts on Ylilauta and Hommaforum are estimates.

Platform type and site	Hate posts	Total posts	Posts per month	Proportion of hate posts
Twitter	**7,450**	**5,478,192**	**2,739,096**	**0.14%**
twitter.com	7,450	5,478,192	2,739,096	0.14%
Discussion forum	**290,076**	**7,432,514**	**3,716,257**	**3.9%**
ylilauta.org*	285,000	4,818,000	2,409,000	5.9%
keskustelu.suomi24.fi	1,931	600,271	300,136	0.32%
hommaforum.org*	1,600	18,300	9,150	8.7%
www.vauva.fi	796	1,100,589	550,295	0.07%
murha.info	313	24,775	12,388	1.3%
www.demi.fi	109	97,330	48,665	0.11%
vesabbs.com	75	15,718	7,859	0.48%
kaksplus.fi	45	32,855	16,428	0.14%
www.tiede.fi	33	30,693	15,347	0.11%
forum.hevostalli.net	32	166,261	83,131	0.02%
nyymichan.fi	30	4,804	2,402	0.62%
maanpuolustus.net	26	18,434	9,217	0.14%
www.punkinfinland.net	17	20,342	10,171	0.08%
www.reddit.com	16	150,441	75,221	0.01%
keskustelu.anna.fi	11	4,655	2,328	0.24%
keskustelu.kauppalehti.fi	11	42,974	21,487	0.03%
keskustelu.pakkotoisto.com	9	16,629	8,315	0.05%
futisforum2.org	6	24,991	12,496	0.02%
Other	16	244,452	122,226	0.01%
Instagram	**453**	**2,482,086**	**1,241,043**	**0.02%**
www.instagram.com	453	2,482,086	1,241,043	0.02%
Blog	**27**	**51,595**	**25,798**	**0.05%**
puheenvuoro.uusisuomi.fi	20	35,511	17,756	0.06%
Other	7	16,084	8,042	0.04%
News comment	**21**	**591,767**	**295,884**	**0.004%**
www.is.fi	10	258,010	129,005	0.004%
www.hs.fi	5	70,050	35,025	0.01%
Other	6	263,707	131,854	0.002%
Facebook (public)	**5**	**774,747**	**387,374**	**0.001%**
www.facebook.com	5	774,747	387,374	0.001%
Other	**0**	**401**	**200**	**0.0%**
Total	**298,032**	**16,811,302**	**8,405,651**	**1.8%**

5.2 Distribution of hate speech over time

The *Social media data set* processed by the AI, which contains platforms other than Ylilauta and Hommaforum, can also be examined in more detail on a day-by-day basis. In September and October 2020, 11,432 posts (0.10%) in the *Social media data set* contained hate speech meeting the definition used in this report. This equals 187 posts identified as hate speech per day on average. The hate posts are distributed over the period covered by the study as follows:

Figure 3. Distribution of messages identified as hate posts by the AI in the *Social media data set* between 1 September and 31 October 2020. In total, the data set contains 11,432 posts identified as hate speech by the AI. The average daily number of posts identified as hate speech is 187 messages.



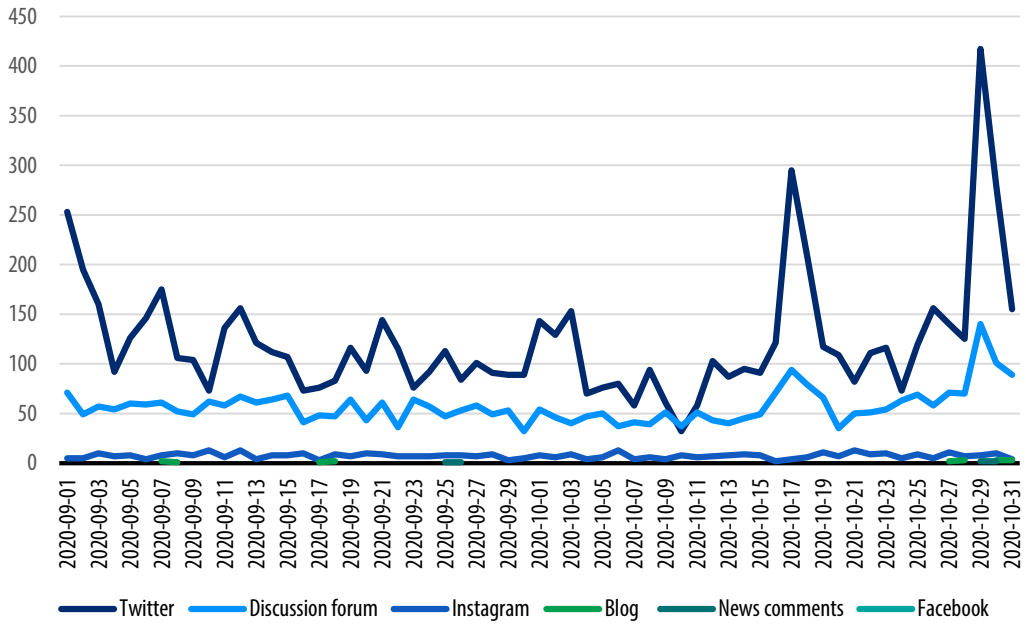
During the time period covered by the study, at least three events occurred which attracted a great deal of attention on the social media and which may have elicited hate speech:

a. Helsinki Pride week 7 September–13 September 2020 (Pride parade and hybrid event on 12 September).
b. On 3 October 2020, a demonstration by Elokapina movement leads to an incident where the police use OC spray as a coercive measure.
c. On 8 October 2020, Trendi magazine publishes an interview with Prime Minister Sanna Marin with an image of PM Marin posing in a business suit without a chemise.

During the Pride week, especially on its first day (7 September) and on the actual day of the event (12 September), higher numbers of identified hate posts can be seen: 249 posts (7 September) and 237 posts (12 September), exceeding the daily average of identified hate posts by approx. 30%. The number of hate posts on the day of Elokapina demonstration on 3 October (202 posts) also appears higher than average. On the other hand, while there was a large volume of discussion about the incident on the social media during the days following the demonstration, the volume of hate speech appears to be even lower than usual on those days. Likewise, the volume of hate speech appeared to be lower than average on the day Prime Minister Marin's interview was published in Trendi, and on the following days.

The following graph describes the daily occurrence of posts identified as hate speech in the *Social media data set*.

Figure 4. Occurrence of posts identified as hate speech during the period covered by the study in the *Social media data set*. The discussion forums do not include Ylilauta and Hommaforum sites.

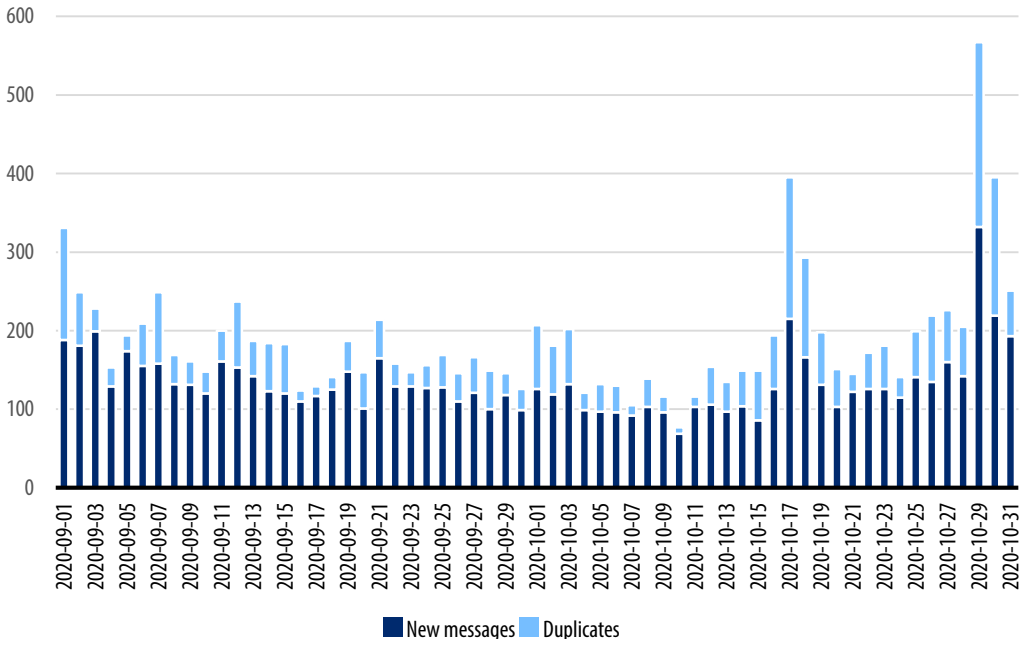


Spikes were registered in hate speech volumes on 17 and 29 October. For a more detailed discussion, see section 6.1.

5.3 Twitter and the power of retweets

Figure 3 shows all posts identified as hate speech by the AI, regardless of whether they were original posts or duplicates of a message already posted. As there was a considerable number of reposted messages or duplicates in the *Social media data set*, it makes sense to examine the proportion of hate speech they account for. In practice, almost all reposted messages were retweets. The Figure below shows how the proportion of retweets of all hate speech is particularly high on days when a large number of hate messages is posted. The proportion of retweets among Twitter messages is 39%.

Figure 5. Distribution of posts identified as hate speech between 1 September and 31 October 2020 in the *Social media data set,* including information on posts with content identical to a message already posted in this data set (duplicate messages).



The results indicate that by retweets, the volume of a Twitter message can be strongly amplified. While we cannot reliably determine that retweets represent intentional decisions made by separate individuals to repeat hate messages, this possibility cannot be ruled out.

Among the twenty user accounts that had the highest number of tweets in the data set, there are four on which more than 98% of the Twitter messages are retweets. The posters behind two of these four accounts use a screen name, while two use a first and last name that seem authentic. This group also contains nine accounts that only retweet rarely

(less than 3% of the messages are retweets). Four of these posters use a screen name. On average, one half of the messages on the remaining seven accounts are new, while the other half are retweets. The posters behind five of these seven accounts use a screen name. The messages on these twenty user accounts which are the most active in tweeting hate speech make up 22% of all hate speech identified on Twitter.

5.4 Categorisation of hate speech

As described in section 2.3 of this report, hate speech was divided into four categories:

1. Hate speech that discriminates against an individual because of their personal characteristics (age, language, appearance, religion or belief, gender or gender expression, sexual orientation, ethnic background or physical functional capacity). Examples: *"ruma transuhan se on", "thaimaan apina"* [*"an ugly trans it is", "an ape from Thailand"*].

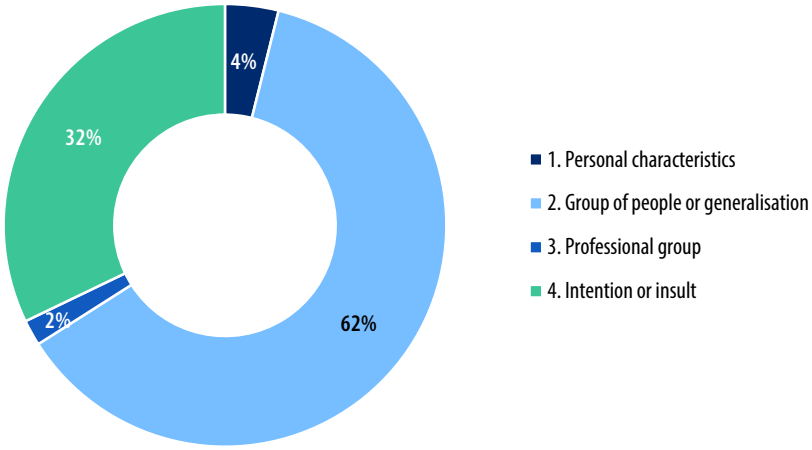
2. Hate speech that stigmatises or generalises a group of people. *"Raiskaukset ryöstöt väkivalta islam."* [*"Rapes muggings violence Islam.."*]

3. Hate speech that targets a professional group. *"[poliitikon nimi] […] Toivottavasti huora olet saanut koronan ja kuolet"* [*"[name of politician] […] I hope you whore have got the coronavirus and will die"*], *"Onkohan [poliitikon nimi] ja näil suvakeilla oikeesti joku poliittinen paritusrinki?"* [*"I wonder if [name of politician] and these sjws really have some sort of political pimping ring?"*]

4. Other expressions interpreted as hate speech, including insults and incitement to violence. The expressions were placed in this category if it was not clear whether the degrading speech is directed at the real characteristics of a person, or whether the poster intends to insult them in general. Insults were interpreted as hate speech if the reason cited in the definition of hate speech is used as an insult while targeting disparaging and stigmatising speech at this group/characteristic. Incitement to violence has also been included in this category. Examples: *"vitu vammanen"*, *"homo"* [*"fuckin' handicapped", "gay"*].

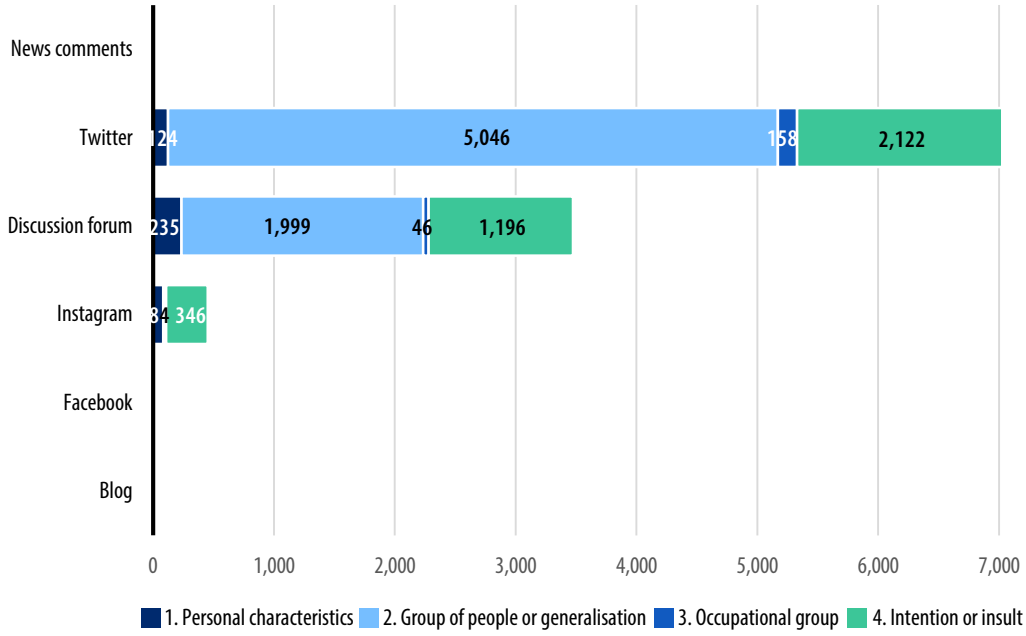
The following Figure shows the distribution of the categories in the *Social media data set*. Category 2 (Group of people or generalisation), which accounted for 62% of hate speech, emerged as the largest one in the data set processed by the AI(excluding Ylilauta and Hommaforum). The second largest group were category 4 messages (Intention or insult) (32%). Personal characteristics were the subject in 4% of the messages, and professional group in 2%.

Figure 6. Hate posts identified in the *Social media data set* divided into four categories.



In quantitative terms, the largest proportion of messages placed in category 2 (degrading message about a group of people or generalisation) was published on Twitter. The second most common category of hate speech on Twitter appears to be 4 (degrading intention or insult). The same categories rank the highest on discussion forums where, however, they are more or less equal in size. On Instagram, on the other hand, category 4 is by far the largest.

Figure 7. Distribution of hate speech categories between different platform types in the *Social media data set*.



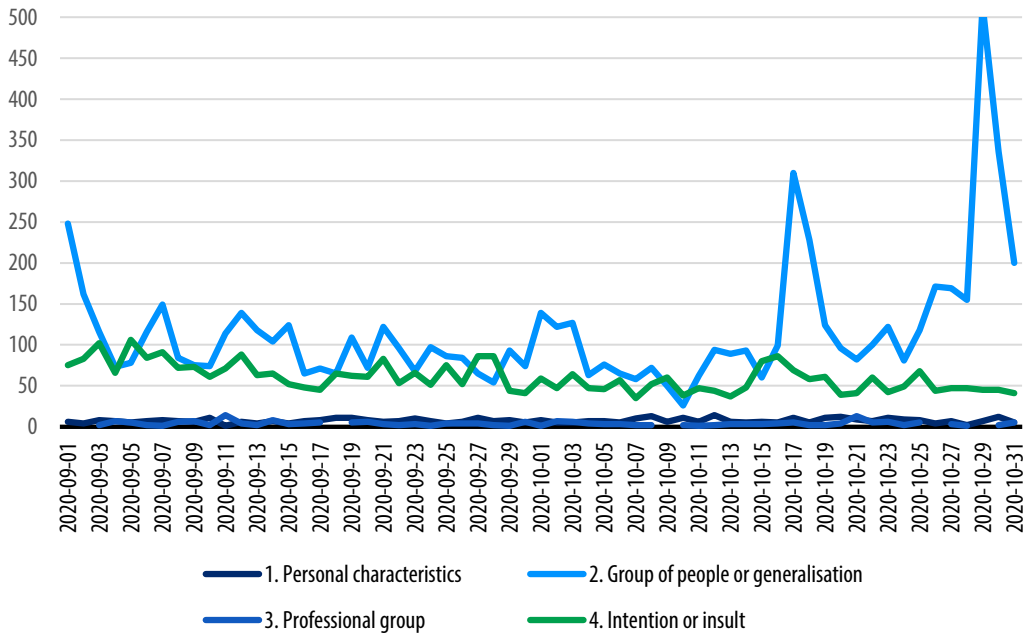
6 Hate speech themes

In this section, we examine the content of hate speech from different perspectives: by examining hate posts identified by the AI on a day-by-day basis, by calculating which words often associated with hate speech are repeated the most frequently in the data set, and by placing comments in the subcategories of the report's definition.

6.1 Examination of spikes in hate speech

Significant spikes can be observed in the volume of identified hate speech on 17–18 October and 29–30 October (See Figure 3). Below we take a closer look at the messages posted on 17 and 29 October. The hate posts from these days consist exclusively of category 2 posts (group of people or generalisation). The following Figure shows the distribution of hate speech categories over the period during which the *Social media data set* was collected.

Figure 8. Hate speech types day by day in the *Social media data set.*



Twitter tweets accounted for a clearly larger than average proportion of all identified hate speech on these two days. Of both spikes in the volume of hate speech on 17 and 29 October, Twitter tweets accounted for 75%, whereas on other days included in the Figure, their share was 65%.

6.1.1 17 Oct 2020, the day following a French teacher's murder

On 16 October 2020, a murder was committed in Paris as a teacher was stabbed to death and beheaded. YLE reported on the incident as follows: "A teacher who showed a cartoon of Prophet Muhammad to his class was killed in front of the school in France – an Islamist terror attack" (Yle 16 October 2020).

On 17 October, the AI identified 395 hate messages, of which 180 (46%) had been reposted (retweets, excluding two messages). When we examined the posts manually, we found that the dominant theme of identified hate speech is ethnic background or religion. The three most frequently retweeted posts refer to Muslims and violence. Two of them refer explicitly to the beheading.

The word *muslimi* [*Muslim*] (in all word combinations and inflections) occurs in 53% of the hate posts identified by the AI on 17 October. The frequency of this word's occurrence is clearly higher than average in posts identified as hate speech in the entire data set, which is 26%. The word *Islam* occurs in 33% of the posts, and the word *terroristi* [*terrorist*] in 17%, whereas the frequency of these words is 15% and 4% in posts identified as hate speech in the entire data set. Some messages contain more than one of these words. In addition, the word *mielenosoitus* [*demonstration*] occurs frequently in retweets of the same Twitter tweet (11%).

When we examine the themes of hate posts identified by the AI, it is important to note that hate speech is an inexact term. Interpretation and discretion are needed to label some of the messages, and even after this, the categorisation of a message as hate speech may not be straightforward. In the training stage, artificial intelligence forms an understanding of where hate speech is placed in the semantic space. Borderline cases and interpretations of the poster's intention made in the annotation stage make up an important part of the AI training data. By default, some of the hate speech identified by artificial intelligence is also borderline cases. What may influence the results is the fact that in this project, external factors of the message were not accounted for in the AI training data, including links referred to in the message.

6.1.2 Widespread tweet on 29 October 2020

Another spike was observed in the data set on 29 October, on which day a total of 567 messages were identified as hate speech by artificial intelligence. Of these, 228 were duplicates (41%). The most common word in this day's posts again was *muslimi* (all inflections and word combinations). On 29 October, it occurs in as many as 72% of all messages identified as hate speech (on average, this word occurs in 26% of all hate posts in the data set). The word *Islam* appears in 34% of the messages (average for all hate posts in the data set is 15%). Of all messages identified as hate speech on 29 October, only one sixth (17%) do not contain the words *islam* or *muslimi*.

A single message, which was retweeted 138 times, accounts for one quarter (25%) of the posts identified as hate speech on that day. At the same time, the tweet in question is the single most frequently occurring message identified as hate speech in the *Social media data set*. It appears in the data set a total of 263 times, accounting for 2.3% of all hate speech identified in the *Social media data set*.

6.2 Words occurring in hate posts

Of the 11,432 messages identified as hate speech by the AI in the *Social media data set*, 2,942 messages (26%) contain the word *muslimi* with different inflections and in all word categories and combinations. *Islam* and its different forms appeared in 1,665 posts (15%), *neekeri* [*nigger*] in 1,465 (13%) and *homo* [*gay*] in 1,057 (9.2%). In addition, the words *suvakki, poliisi, huora, terroristi, somali, ählämi and rasisti* [*sjw (social justice warrior), police, whore, terrorist, Somali, Dune coon and racist*] occurred more than a hundred times (over 0.9%). Other frequently occurring words included *Allah, apina, feministi, hintti, juutalainen, koraani, maahanmuuttaja, mielenosoitus, monikulttuurisuus, mustalainen, mutakuono, natsi, nekru, pakolainen, rättipää, ryssä, somppu, sosiaalipummi, suvaitsevainen, vammainen,* [*Allah, ape, feminist, fag, Jewish, Quran, immigrant, demonstration, multiculturalism, Gypsy, darkey, Nazi, nog, refugee, raghead, Russki, Somali (offensive), welfare parasite, social justice warrior, disabled*], all of which occurred at least 30 times (0.3%) in identified hate posts. However, the occurrence of a single word does not categorically make the message a hate post, and the entire content of the post was always assessed. The words listed above mainly come from the list used by the Police University College in its annual report.

The following Table links the most common words to categories of hate speech.

Table 4. Frequently occurring words in messages classified in different categories of hate speech.

	Frequently occurring words (with different inflections)
(1) Personal characteristics	*huora, homo, vammainen, vähemmistö [whore, gay, disabled, minority]*
(2) Stigmatisation of a group	*muslimi, islam, suvakki, poliisi, somali, ählämi, terroristi, rasisti [Muslim, Islam, sjw, police, Somali, Dune coon, terrorist, racist]*
(3) Professional group	*terroristi, pakolainen, homo [terrorist, refugee, gay]*
(4) Insult or other hate speech	*neekeri, homo, huora, suvakki, poliisi, ryssä, vammainen [nigger, gay, whore, sjw, police, Russki, disabled]*

When we look at the most common platform types, different topics are highlighted. In more than a half (51%) of hate posts on Instagram, the term *homo* [*gay*] is used, and based on this data set, Instagram is not a platform used for discussions on religions or immigration classified as hate speech. The second most common word in Instagram hate posts is *huora* [*whore*] (10%). *Homo* is also one of the standard words of hate posts (15%) on discussion forums, even though *muslimi* (17%) is even more common. *Islam* is also frequently referred to on discussion forums (12%). On Twitter, *muslimi* is a key word indicating hate speech: 31% of hate posts identified on Twitter use this word. *Islam* is the second most common word on Twitter (17%). Compared to other platforms, the word *neekeri* [*nigger*] is used frequently on Twitter: it occurs in up to 16% of Twitter messages identified as hate speech. On other platforms, individual words occurred too infrequently to have any statistical significance.

In messages identified as hate speech in the separate data set from Ylilauta and Hommaforum, the words *vammainen* [*disabled*] and *huora* as well as *homo* and *neekeri* were very common. In addition, discussions about identified persons' state of health were prominent in this material. In Hommaforum messages, the word *neekeri* was particularly frequent; in this small data set, it occurred in 15 hate posts (43%).

6.3 A small minority of users produces the majority of hate speech

As is typical of natural distributions, a small proportion of users would appear to produce the greatest part of hate speech. One out of four posters of messages identified as hate speech in the *Social media data set* (24%) did not use a screen name. All anonymous writers in the data set posted their messages on discussion forums. Three quarters (76%) of the messages identified as hate speech on the discussion forums, on the other hand, were posted anonymously, and only 24% with user IDs revealing the poster's name.

In the *Social media data set*, identified hate speech is linked to 2,303 different users. The ten users who came up the most frequently in the hate posts of the data set accounted for 11% of all identified hate speech. The most active producer of hate speech in the *Social media data set* is a Twitter account which published 352 messages identified as hate speech in two months.

6.4 Personal and group characteristics in hate speech

In the *Social media data set* identified as hate speech by the AI, 500 cases which the AI regarded as hate speech with the highest level of certainty were analysed. They were manually classified in the subcategories of the hate speech definition used in this report. If a post was interpreted as hate speech on more than one criteria, it was put in all relevant subcategories. For example, the classification of the expression "*vitun homo*" [*"fucking gay"*] is c (sexual orientation). Such comment as *"Olipa ruma takkutukka!! Kuinka monta ählämiä ja neekeriä sitä on käyttänyt*?" [*"What an ugly shaggy hair!! How many Dune coons and niggers have used it [him or her]?"*] was placed in subcategories b, d, h, as the contempt is motivated by b. gender (presumed), d. ethnic background (use of the word *neekeri*), and h. appearance. In 65% of the messages, hate speech was motivated by ethnic background: in other words, skin colour, origin or language. In 27% of the messages, hate speech was motivated by sexual orientation, and in 20%, gender, gender identity or gender expression.

The distributions of the subcategories were as follows:

Table 5. Subcategories of hate speech and their frequency in the data set.

Subcategory	Occurrences	% of posts
a. age	0	0%
b. gender, gender identity or expression of gender	100	20%
c. sexual orientation	136	27%
d. ethnic background (skin colour, origin, language)	324	65%
e. religion or belief	17	3.4%
f. political opinion	20	4.0%
g. physical functional capacity	22	4.4%
h. appearance	18	3.6%
i. nationality	14	2.8%
j. other	4	0.8%

The most common type of hate posts in these messages was motivated by ethnic background, such as skin colour, origin or language (subcategory d), which accounted for 65% of the messages. Example: *"ei neekereitä Suomeen"* [*"no niggers to Finland"*].

The second most common type of hate speech, or that motivated by sexual orientation (subcategory c), was found in 27% of the messages ("*vitun homo*") ["*fucking gay"*]. One out

of five posts (20%) included hate speech related to gender (subcategory b) (**"***olet huora*") ["*you are a whore"*].

Some of the hate posts were identified on multiple criteria and belonged to more than one subcategory: Almost a quarter (23%) of the analysed messages belong to at least two subcategories, while 6% belong to at least three.

Table 6. Hate speech identified on multiple criteria in figures: number of different categories of hate speech per analysed message.

Subcategories	Posts	Proportion (%)
1	387	77%
2	83	17%
3	22	4%
4	4	1%
5	4	0.8%
TOTAL	**500**	**100%**

The most common type of hate speech identified on multiple criteria is messages referring to both gender (b) and ethnic background (d), which occurs in 4% of the posts ("ähläimit *huumanneet matupatjan", "neekerimiesten nussima huora*") ["*Dune coons drugged an immigrant mattress", "a whore fucked by nigger men"*]. The second most common type identified on multiple criteria refers to gender (b), sexual orientation (c) and ethnic background (d) in the same message (3% of the messages). At most, a single message could contain up to five criteria for identifying hate speech.

The small data set divided into subcategories would appear to indicate that the most common motivations for hate speech vary slightly from platform to platform. On Twitter, 91% of the posts identified as hate speech by the AI tool with the highest level of certainty referred to ethnic background (d). Hate speech is motivated by ethnic background in more than one half of the cases (60%) on discussion forums and only about one out of ten cases on Instagram (12%). It would appear that hate speech on Instagram is more frequently motivated by sexual orientation (c), which is referred to in 58% of the hate posts on this platform. On Twitter, sexual orientation is only referred to in 4% of hate posts. Considerably more hate speech identified on multiple criteria occurs on discussion forums than on Twitter and Instagram (40% of hate posts). The longer length of messages is likely to be a partial explanation for this.

7 Conclusions

A key outcome of this project was an AI model capable of identifying online hate speech. The AI learned to identify hate speech even with a relatively small set of training data. Artificial intelligence compares texts to other texts with a similar semantic meaning and gradually forms an understanding of the area occupied by hate speech in the so-called semantic space. The artificial intelligence model created in the project can be used to identify hate posts among other messages. If more training data can be provided for the AI in the future, it will learn to distinguish borderline cases from hate speech as defined in this report with increasing levels of certainty.

A large *Social media data set* consisting of approximately 12 million messages, of which 11,432 were identified as hate posts by the AI, can also be regarded as one of the project's outcomes. This report contains a numeric breakdown of the findings as well as an overview of hate speech content based on smaller data sets. However, the data set also lends itself to in-depth qualitative analysis in which a selected area can be examined, for example the content of hate posts published on a particular platform or the content of a specific topic area of hate speech, in particular.

Hate speech is an inexact term for which narrower or broader definitions can be drawn up. In this project, a more extensive definition of hate speech was used than in the Finnish Criminal Code, for example. Labelling posts as hate speech or non-hate speech often required interpretation and discretion. From the AI perspective, these borderline cases are placed on the margins of the area occupied by hate speech in the semantic space, whereas obvious cases are found in the middle. As the *Social media data set* acquired for this project was examined by the AI, the situations where the AI's decisions differed from those made by a human annotator were always borderline cases. Consequently, it is conceivable that artificial intelligence can reliably filter online material to find messages that can be categorised as hate posts at the highest level of certainty. When we examine the themes of AI-identified hate speech, however, we should be aware of the fact that by default, some of the messages in the data set are always borderline cases.

Our assessment indicates that Ylilauta appears to serve as the platform for a large part of online hate speech in Finland. In the quantitative analysis, it was assessed as a platform with a particularly high volume of hate speech: it accounted for 96% of all online hate posts identified in this project. Ylilauta is a subcultural imageboard predominated by the ideal of free discussion and anonymity. It has become known for its humour in which political correctness has no place, jokes and trolling, and there is little monitoring of the threads, apart from clear breaches of law. (Vainikka 2019). The definition of hate speech used in this project is broader than that contained in the Finnish Criminal Code. The

prevailing discussion culture on Ylilauta may be part of the reason for so many of the messages posted on this platform being

identified as hate speech in our project. In the light of our results, however, it could be asked if the tone used on Ylilauta is something that the larger body of social media users see as hostile and offensive, and if the platform has become a community that provides support for hate speech. Consequently, targeting measures and dialogue directly at platforms of this type could benefit the work against hate speech, in addition to intervening in the activities of international digital giants.

Of the large social media platforms, Twitter was the most significant one in terms of hate speech: 2.5% of messages identified as hate speech were posted on Twitter. The large number of retweets among messages identified as hate posts was an interesting finding. This indicates that while the number of hate speech producers may be small, effective retweeting may make the hate speech phenomenon look larger than its size. It should be remembered, however, that these proportions would change by default if private accounts and groups on Facebook were also included. The number of retweets in the data set was the highest on days when the volume of hate speech was also great in other respects. When the content of hate speech spikes was examined, retweets were found to play a significant role in them. A few frequently retweeted individual messages could to a large extent explain a spike. The role of ordinary social media users as spreaders of hate speech is thus emphasised, and their behaviour related to retweeting, or refraining from doing so, is of great importance.

PROJECT TEAM

The members of the project team that produced the report were:

Definition of hate speech, annotation, the report: Laura Kettunen, M.A.
Commenting on the report: Reeta Pöyhtäri, D.Soc.Sc.
Utopia Analytics:
- Report: Mari-Sanna Paukkeri, D.Sc. (Tech.)
- Data modelling: Jaakko Väyrynen, D.Sc. (Tech.)
- Project management: Kari Kemppi, M.Sc. (Eng.)

Ministry of Justice

REFERENCES

Council of Europe Committee of Ministers recommendation to Member States on "Hate Speech" R (97) 20 (1997). https:// search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680505d5b

Hiltunen, I. (2017). Painostuksen monet muodot. *Journalisti* 6 (19). https://www.journalisti.fi/artikkelit/2017/7/painostuksen-monet-muodot/

International Press Institute (2018). Online Attacks on Journalists in Finland: Overview and Best Newsroom Practices. https://ipi.media/countering-online-harassment-in-newsrooms-finland/. Accessed on 10 September 2019.

Knuutila, Aleksi; Kosonen, Heidi; Saresma, Tuija; Haara, Paula & Pöyhtäri, Reeta (2019). The impact of hate speech on public decision-making (in Finnish with English abstract). Publications of the Government´s analysis, assessment and research activities.

Korhonen, N.; Jauhola, L.; Oosi, O. & Huttunen, H. P. (2016). "I often find myself thinking how I should be or where I shouldn't go" – Survey on hate speech and harassment and their influence on different minority groups (in Finnish with English abstract). Publications of the Ministry of Justice, no. 7. http://urn.fi/ URN:ISBN:978-952-259-496-9

Korpisaari, Päivi. (2019). Sananvapaus verkossa – yksilöön kohdistuva vihapuhe ja verkkoalustan ylläpitäjän vastuu. *Lakimies* 7–8/2019 pp. 928–952.

Laaksonen, S.-M.; Haapoja, J.; Kinnunen, T.; Nelimarkka, M. & Pöyhtäri, R. (2020). The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring. *Front. Big Data* 3:3. doi: 10.3389/fdata.2020.00003

Association of Finnish Lawyers (2019). Lakimiesliitto vaatii järeitä toimia oikeudenhoidon ammattilaisten suojaamiseksi vainoamiselta. Association of Finnish Lawyers, 21 May 2019. https://www.lakimies-liitto.fi/uutiset/lakimiesliitto-vaatii-jareita-toimia-oikeudenhoidon-ammattilaistensuojaa-miseksi-vainoamiselta/

Mäkinen et al. (2019). Words Are Actions: More Efficient Measures against Hate Speech and Cyberbullying (in Finnish with English abstract). Ministry of the Interior publication.

Pöyhtäri R.; Haara, P. & Raittila, P. (2013). *Vihapuhe sananvapautta kaventamassa*. Tampere: Tampere University Press.

Pöyhtäri, Reeta (2015). Vihapuhe haasteena uutismedialle ja journalismille. *Vihapuhe Suomessa*, ed. Neuvonen Riku. Edita Publishing Oy.

Rauta, J. (2018). Poliisin tietoon tullut viharikollisuus Suomessa 2017. Reports of the Police University College 131. https://www.theseus.fi/bitstream/handle/10024/154780/POLAMK_Rap131_web.pdf

Ruotsalainen, M. (2017). Vihapuheen nousu julkisessa keskustelussa. *Jätkät ja jytkyt: Perussuomalaiset ja populismin retoriikka*, eds. E. Palonen & T. Saresma, 181–198. Tampere: Vastapaino.

Committee for Public Information (2015). Kysely tutkijoiden asiantuntijaroolissa saamasta palautteesta: Tulosyhteenveto. Committee for Public Information , 22 December 2015.

Vainikka, Eliisa (2019). Naisvihan tunneyhteisö. Anonyymisti esitettyä verkkovihaa Ylilaudan ihmissuhdekeskusteuissa. *Media & Viestintä* 42 (2019): 1–25.

Report of the Office of the Prosecutor General's working group (2012). Rangaistavan vihapuheen levittäminen Internetissä. https://www.valtakunnansyyttajanvirasto.fi/material/attachments/ valtakunnansyyttajanvirasto/vksvliitetiedostot/tyoryhmat/6Jqa-1QEsJ/17-34-11_tyoryhma- raportti.pdf

Ministry of Justice Finland
PL 25
00023 Valtioneuvosto, Finland
www.ministryofjustice.fi

ISSN 2490-0990 (PDF)
ISBN 978-952-259-811-0 (PDF)

MINISTRY OF JUSTICE
FINLAND