Oikeusministeriön julkaisuja
Justitieministeriets publikationer

Selvityksiä ja ohjeita
Utredningar och anvisningar

2021:15

Tekoälyn hyödyntäminen vihapuheen seurannassa

OIKEUSMINISTERIÖ
JUSTITIEMINISTERIET

Oikeusministeriön julkaisuja, Selvityksiä ja ohjeita 2021:15

Tekoälyn hyödyntäminen vihapuheen seurannassa

Laura Kettunen, FM
Mari-Sanna Paukkeri, TKT, Utopia Analytics

Oikeusministeriö 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

Tämä raportti on tuotettu osana Tiedolla vihaa vastaan -hanketta (Facts against Hate), joka on saanut rahoitusta Euroopan Unionin perusoikeus , tasa-arvo- ja kansalaisuusohjelmasta (2014–2020).

Oikeusministeriö

ISBN pdf: 978-952-259-893-6
ISSN pdf: 2490-0990

Taitto: Valtioneuvoston hallintoyksikkö, Julkaisutuotanto
Helsinki 2021

Kuvailulehti
22.4.2021

Tekoälyn hyödyntäminen vihapuheen seurannassa

Oikeusministeriön julkaisuja, Selvityksiä ja ohjeita 2021:15		**Teema**	Selvityksiä ja ohjeita
Julkaisija	Oikeusministeriö		
Tekijä/t	Laura Kettunen, Mari-Sanna Paukkeri		
Kieli	suomi	**Sivumäärä**	48

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	978-952-259-893-6	**ISSN PDF**	2490-0990
Julkaisun osoite	http://urn.fi/URN:ISBN:978-952-259-893-6		

Presentationsblad
22.4.2021

Utnyttjande av artificiell intelligens vid uppföljning av hatretorik

Justitieministeriets publikationer, Utredningar och anvisningar 2021:15		**Tema**	Utredningar och anvisningar
Utgivare	Justitieministeriet		
Författare	Laura Kettunen, Mari-Sanna Paukkeri		
Språk	finska	**Sidantal**	48

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	978-952-259-893-6	**ISSN PDF**	2490-0990
URN-adress	http://urn.fi/URN:ISBN:978-952-259-893-6		

Description sheet
22 April 2021

Utilisation of artificial intelligence in monitoring hate speech

Publications of the Ministry of Justice, Reports and guidelines 2021:15		**Subject**	Reports and guidelines
Publisher	Ministry of Justice, Finland		
Authors	Laura Kettunen, Mari-Sanna Paukkeri		
Language	Finnish	**Pages**	48

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	978-952-259-893-6	**ISSN PDF**	2490-0990
URN address	http://urn.fi/URN:ISBN:978-952-259-893-6		

Sisältö

Tiivistelmä

Tämä raportti on tuotettu oikeusministeriön koordinoiman Tiedolla vihaa vastaan -hankkeen toimesta. Yksi hankkeen tavoitteista on kehittää vihapuheen seurantaa testaamalla verkossa tapahtuvan vihapuheen kohdennetun seurannan uusia välineitä.

Raporttia varten on pilotoitu tekoälyn mahdollisuuksia oppia tunnistamaan vihapuhetta digitaalisissa ympäristöissä. Lähestymistapana on ihmistyön ja koneoppimisen yhdistelmä. Tavoitteena on muun muassa muodostaa käsitys siitä, millaisissa kanavissa vihapuhetta esiintyy ja millaisia eroja eri alustoilla esiintyvässä vihapuheessa on.

Vihapuheen määrittely tehtiin aikaisempaan sosiaalitieteelliseen tutkimukseen pohjaten. Määrittelyssä muodostettiin luokkia, joihin nojautuen datasta tunnistettiin manuaalisesti vihapuheen esimerkkejä. Näitä annotointeja käytettiin opetusaineistona Utopia AI Moderatorille, joka on kieliriippumaton tekstianalytiikkaa ja koneoppimista hyödyntävä järjestelmä. 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 on Twitter-viestit (2.5 %) ja Instagram-viestit (0.2 %). Blogit, uutiskommentit ja Facebookin julkiset viestit kattavat alle 0.02 % kaikesta tunnistetusta vihapuheesta. Facebookin suljetut ryhmät ja ei-julkiset tilit eivät ole mukana aineistossa.

Merkittävin vihapuheen julkaisualusta näyttäisi olevan ylilauta.org (285 000 viestiä, 96 % vihapuheeksi tunnistetuista viesteistä). Seuraavaksi eniten vihapuhetta näyttäisivät sisältävän twitter.com (7 450 viestiä), suomi24.fi (1 931 viestiä), hommaforum.org (1 600 viestiä) ja vauva.fi (796 viestiä).

Raportissa käsitelty aineisto avaa tarkasteluun vihapuheen mahdollisesti nousevat areenat. Kansainvälisten sosiaalisen median alustojen osalta merkittävin vihapuheen julkinen julkaisualusta näyttäisi olevan Twitter, josta löytyi tarkasteluvälillä 7 450 kpl vihapuheeksi tunnistettua viestiä, 0.14 % kaikista tviiteistä. Tuloksista ilmenee, että Twitterin uudelleentviittauksilla on merkittävä rooli vihapuheviestien levittämisessä. Vihapuheeksi tunnistetuista Twitter-kirjoituksista kaikkiaan 39 % on duplikaatteja.

Tekoälyn avulla tarkemmin analysoidusta vihapuheesta 62 % luokiteltiin ihmisryhmää leimaavaksi tai yleistäväksi, 32 % solvaukseksi tai muuksi vihapuheeksi tulkittavaksi ilmaisuksi, 4 % henkilökohtaisiin ominaisuuksiin liittyväksi vihapuheeksi ja 2 % ammattiryhmään kohdistuvaksi vihapuheeksi.

Raportissa luodaan katsaus myös vihapuheen teemoihin. Vihapuhesanastoanalyysin mukaan koko aineiston yleisin yksittäinen sana on *muslimi*. Se esiintyy 26 %:ssa kaikista tekoälyn vihapuheeksi tunnistamista teksteistä. Muita yleisiä sanoja olivat *islam, neekeri* ja *homo* sekä *suvakki, poliisi, huora, terroristi, somali, ählämi* ja *rasisti*.

On huomionarvoista, että pieni osa kirjoittajista näyttäisi tuottavan suurimman osan vihapuheesta. Vihapuheviesteissä useimmin esiintyvät 10 käyttäjää vastaavat 11 %:sta kaikesta tunnistetusta vihapuheesta. *Sosiaalisen median aineiston* innokkain vihapuhekirjoittaja on Twitter-tili, joka julkaisi tarkastelujakson aikana 352 vihapuheeksi tunnistettua viestiä.

Referat

Den här rapporten har producerats inom ramen för projektet Fakta mot hat som koordineras av justitieministeriet. Ett av målen med projektet är att utveckla övervakningen av hatretorik genom att testa nya verktyg för riktad övervakning av hatretorik online.

Projektet har testat möjligheterna för artificiell intelligens att lära sig känna igen finskspråkiga hatretorik i digitala miljöer. Metoden är en kombination av mänskligt arbete och maskininlärning. Syftet är bland annat att bilda en uppfattning om i vilka kanaler hatretorik förekommer och skillnaderna mellan hatretorik på olika plattformar.

Definitionen av hatretorik gjordes baserat på tidigare socialvetenskaplig forskning. I definitionen bildades kategorier, som användes för att manuellt identifierade exempel på hatretorik från datan. Dessa anteckningar användes som läromedel för Utopia AI Moderator, ett språkoberoende system som använder textanalys och maskininlärning. Materialet bestod av cirka 12 miljoner kommentarer och online artiklar på finska från september till oktober 2020.

Baserat på resultaten och definitionen i denna rapport förekommer hatretorik i offentliga finska plattformerna i cirka 150 000 meddelanden per månad, eller 1.8 procent av alla meddelanden. Under två månaders granskningsperioden från 1 september till 31 oktober 2020 identifierades 298 032 hatretorikmeddelanden, varav 97 % uppträdde på olika diskussionsforum. Näst vanligaste plattform typerna är olika Twitter meddelanden (2.5 %) och Instagram meddelanden (0.2 %). Bloggar, nyhetskommentarer och offentliga inlägg på Facebook täcker mindre än 0.02% av all hatretorik som identifierades. Facebooks stängda grupper och icke-offentliga konton ingår inte i materialet.

Den signifikantaste plattformen för publicering av hatretorik verkar vara ylilauta.org (285 000 inlägg, 96 % av meddelanden som identifierats som hatretorik). Näst mest hatretorik innehöll twitter.com (7 450 inlägg), suomi24.fi (1 931 inlägg), hommaforum.org (1 600 inlägg) och vauva.fi (796 inlägg).

Materialet som behandlas i rapporten öppnar möjligheten att granska växande arenor för hatretorik. När det gäller internationella sociala medier förefaller den viktigaste plattformen för publicering av offentlig hatretorik vara Twitter, som innehöll 7 450 meddelanden som identifierats som hatretorik under granskningsperioden, 0.14 % av alla tweetar. Resultaten visar att retweets på Twitter spelar en viktig roll i spridningen av hatretorikmeddelanden. Totalt 39 % av Twitter-inlägg som identifierats som hatretorik är dubbletter.

Av det hatretorik som analyserades vidare med hjälp av artificiell intelligens klassificerades 62 % som stigmatiserande eller generaliserande av en grupp människor, 32 % som en förolämpning eller annat uttryck tolkat som hatretorik, 4 % som hatretorik relaterat till personliga egenskaper och 2 % som hatretorik riktat mot en yrkesgrupp.

Rapporten ger också en översikt över teman för hatretorik. Enligt ordförrådsanalys av hatretorik är det vanligaste enstaka ordet i hela materialet *muslim*. Det förekommer i 26 % av alla texter som genom artificiell intelligens identifieras som hatretorik. Andra vanliga ord var *islam*, *neekeri*, *homo*, *suvakki, poliisi, huora, terroristi, somali, ählämi* och *rasisti*.

Det är anmärkningsvärt att en liten andel av skribenterna verkar producera det mesta av hatretoriket. De 10 vanligaste användarna av hatretorikmeddelanden står för 11 % av alla identifierade hatretorik. Den ivrigaste hatretorikskribenten på sociala medier är ett Twitter-konto, som publicerade 352 meddelanden som identifierats som hatretorik under översynsperioden.

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 Johdanto

Tämä raportti on tuotettu oikeusministeriön koordinoiman Tiedolla vihaa vastaan -hankkeen toimesta. Hankkeen tavoitteena on viharikosten ja vihapuheen vastaisen työn tehostaminen. Yksi hankkeen tavoite on kehittää vihapuheen seurantaa testaamalla verkossa tapahtuvan vihapuheen kohdennetun seurannan uusia välineitä.

Tätä raporttia varten Utopia Analytics on tutkinut tekoälyyn perustuvan moderointituotteen potentiaalia vihapuheen löytämisen ja seurannan välineenä digitaalisissa ympäristöissä. Seurantaa tehtiin kertaluonteisesti syksyllä 2020.

Projektin tavoitteena on ollut katsoa verkossa esiintyvää vihapuhetta laajasti ja muodostaa käsitys siitä, millaisissa kanavissa vihapuhetta esiintyy, millaisia eroja eri alustoilla esiintyvässä vihapuheessa on ja onko verkossa esiintyvän vihapuheen ja reaalimaailman tapahtumien välillä havaittavissa yhteyttä.

Aineistoja tarkastellessamme noudatamme tutkimuseettisiä ohjeita. Tutkimus ei saa olla vaaraksi kenenkään yksityisyydelle eikä aiheuttaa osallistujille minkäänlaisia haittoja. Käytämme aineistona pelkästään sellaisia viestejä, jotka sosiaalisen median alustoilla on viestien kirjoittajan toimesta merkitty julkisiksi. Raportissa käytetyt esimerkit ovat otteita aineistossa käsitellyistä viesteistä, joista on poistettu nimet ja käyttäjänimet. Joissakin tapauksissa oikeakielisyyttä on sen verran korjattu, että lyhyt esimerkki olisi yksinään ymmärrettävissä samoin kuin koko viestin kontekstissa.

2 Vihapuheen määrittelystä

Tässä raportissa esiteltävät tulokset pohjautuvat vihapuheen määritelmään, joka on tehty erikseen tätä projektia varten ja verkkokeskusteluja ajatellen. Määrittelyssä on käytetty apuna aikaisempia tutkimuksia ja kuunneltu asiantuntijoita. Määrittely aloitettiin muodostamalla pää- ja alaluokat eri vihapuhetyypeistä, jonka jälkeen määrittely viimeisteltiin merkitsemällä sosiaalisen median viestejä manuaalisesti vihapuheeksi luokituksineen ja ei-vihapuheeksi, tarkastellen näin luokituksen toimivuutta oikeilla viesteillä.

Määritelmästä on pyritty tekemään niin kattava kuin se tällä tavalla luokittelemalla on mahdollista. Määritelmään ei ole vaikuttanut ajatus siitä, mitä tekoäly voi tai ei voi oppia. Näin tässä raportissa tehty vihapuheen määritelmä ja sen hyödyntäminen ihmistyönä tehdyssä vihapuheen tunnistamisessa sosiaalisen median aineistosta on yksi erillinen kokonaisuus. Tekoälyn opettaminen ihmistyönä koottujen esimerkkien avulla on toinen erillinen kokonaisuus.

Tämä raportti ei ota kantaa siihen, miten jonkun muun tahon tulisi vihapuhemääritelmä tehdä, vaan toivottaa jatkokeskustelun aiheesta tervetulleeksi ja pitää jatkotutkimusta tärkeänä.

2.1 Premissit

Vihapuhe ei ole rikosnimike vaan kuvaa tietynlaisten tekojen joukkoa. Vihapuheella tarkoitetaan viestintää, joka levittää tai lietsoo vihaa yhtä ihmistä tai ihmisryhmää vastaan henkilöön liittyvän syyn perusteella. Tässä raportissa esitettävä vihapuheen luokittelu on tehty nojautuen kotimaisissa tieteellisissä tutkimuksissa käytettyihin määritelmiin, yhdenvertaisuuslain häirintäpykälään (§14), tasa-arvolain syrjintäpykälään (§7), rikoslain pykälään kiihottamisesta kansanryhmää vastaan (§10), sekä Valtakunnan syyttäjänviraston raporttiin (2012) rangaistavan vihapuheen levittämisestä.

Vihapuhe voi olla rikoslain mukainen rikos, yhdenvertaisuuslaissa tai tasa-arvolaissa kiellettyä syrjintää tai muuten yleisesti haitallista ilmaisua. Se voi olla kirjoitusta tai muuta viestintää. Tässä raportissa vihapuhe määritellään rangaistavaa eli rikoslaissa kiellettyä vihapuhetta laajemmin, sillä määrittelyssä on haluttu ottaa huomioon myös muu lainsäädäntö. Raportin tarkoitus ei ole todeta tiettyjä viestejä rangaistaviksi, sillä sen arvioiminen on lainvalvontaviranomaisten tehtävä. Tässä raportissa ei myöskään oteta kantaa siihen, onko jokin yksittäinen ilmaisu laissa kielletty.

Raportissa tarkoitamme vihapuheella viestintää, jonka merkitys tai sävy on halventava, nöyryyttävä, uhkaava, vihamielinen, hyökkäävä tai epäinhimillistävä (esimerkiksi verrataan ihmistä eläimeen tai loiseen). Viestintä voi liittyä henkilökohtaisiin ominaisuuksiin tai leimata jotakin ihmisryhmää. Henkilökohtaisilla ominaisuuksilla tarkoitetaan esimerkiksi henkilön ikää, kieltä, ulkonäköä, uskontoa tai vakaumusta, sukupuolta, seksuaalista suuntautumista, etnistä taustaa tai ruumiin toimintakykyä. Ilmaisut voivat myös kohdistua yksilöön sen perusteella, että hänen oletetaan kuuluvan johonkin kansalliseen, etniseen, uskonnolliseen, seksuaaliseen tai muuhun ryhmään. (Knuutila et al. 2019, Valtakunnansyyttäjänviraston raportti 2012).

Euroopan neuvoston ministerikomitea on määritellyt vihapuheeksi kaikki ilmaisumuodot, jotka levittävät, lietsovat, edistävät tai oikeuttavat etnistä vihaa, ulkomaalaisvastaisuutta, antisemitismiä tai muuta vihaa. Tämä koskee niin aggressiivista kansallismielisyyttä kuin vähemmistöjen, maahanmuuttajien ja maahanmuuttajataustaisten ihmisten syrjintää ja vihamielisyyttä heitä kohtaan. Kansainvälisen antisemitismin vastainen toimikunnan mukaan verkkovihaa on minkä tahansa elektronisen välineen avulla levitettävä rasismi, antisemitismi, uskontoon liittyvä kiihkoilu, homofobia tai muu seksuaaliseen suuntautumiseen liittyvä fobia, vammaisiin kohdistuva ahdasmielisyys, poliittinen viha, huhujen levittäminen, sukupuoleen liittyvä viha, väkivaltainen pornografia, terrorismin edistäminen, nettikiusaus, ahdistelu ja vaino, vastapuheen vaientava puhe (kuten muiden häpäiseminen, solvaus ja nimittely) ja ryhmiä leimaava puhe. (Pöyhtäri 2015)

Vihapuheella voidaan pyrkiä vaikuttamaan päätöksentekoon. Vihapuheen kohteena voivat olla ammattiryhmät, jonkin asian tai ryhmän puolesta toimivat henkilöt tai julkisuuden henkilöt, jotka ovat esillä ammattinsa tai muun syyn vuoksi. Jos ammattiryhmän edustajaan kohdistuva arvostelu koskettaa toimintaa yksinomaan hänen roolissaan ammattilaisena, tätä ei tässä raportissa katsota vihapuheeksi. Kuitenkin, jos kritiikki kohdistuu henkilökohtaisiin ominaisuuksiin tai johonkin hänen edustamaansa ihmisryhmään, tämä lasketaan vihapuheeksi. (Knuutila et al. 2019) Vihapuheen kohteeksi ovat aiempien tutkimusten mukaan joutuneet ainakin poliitikot ja päättäjät, journalistit, tutkijat, poliisit, syyttäjät ja tuomarit. (Ks. mm. Hiltunen 2017, Lakimiesliitto 2019, Tiedonjulkistamisen neuvottelukunta 2015, Pöyhtäri, Haara & Raittila 2013).

Vihapuheelle on tyypillistä, että se yleensä tuottaa ja ylläpitää puhetapaa, jonka mukaan on hyväksyttävää arvioida jotkut ihmisyksilöt, vähemmistöt, kansallisuudet, kulttuurit, etniset ryhmät tai uskonnot jonkin (väitetyn) ominaisuutensa vuoksi itseä tai muita alempiarvoisiksi, tai pyrkiä tuhoamaan ne. Lisäksi kysymys valtasuhteista ja vallankäytöstä liittyy keskeisesti vihapuheeseen, ja on tärkeää kysyä, kuka puhuu ja missä asemassa hän on. (Pöyhtäri 2015) Kirjoitus voi olla ilmauksena neutraali, mutta konteksti voi olla sellainen, että sanoja itse asiassa maalittaa jonkin henkilön vihan kohteeksi.

Myös maalittaminen voidaan joissakin tapauksissa tulkita vihapuheeksi (Korpisaari 2019). Laajimmassa merkityksessään vihapuheen voidaan katsoa käsittävän lisäksi myös trollauksen ja ns. doksaamisen eli toisen ihmisen henkilötietojen jakamisen julkisuuteen. Vihapuhe-termiä on käytetty myös ns. kybervihan ja kyberväkivallan sekä toksisen puheen yhteydessä. (Laaksonen et al. 2020). Tämän raportin määritelmässä ei kuitenkaan huomioida erikseen maalittamista, trollausta tai doksausta.

2.2 Tulkinnanvaraiset ilmaisut

Vihapuhe ei välttämättä ole tunnesisällöltään vihaista tai edes sisällä voimakkaita tunteita. Vihapuhe voi olla tyyliltään neutraalia ja rauhallista siitä huolimatta, että se pyrkii halventamaan tai leimaamaan. (Knuutila et al. 2019). Konteksti onkin useissa tapauksissa olennainen tekijä, kun arvioidaan, onko jokin ilmaus vihapuhetta vai ei.

Vihapuheen tunnistaminen ei näin ollen ole aina yksiselitteistä. Tässä raportissa tulkinnanvaraisten ilmausten kohdalla on tekstin semanttisen sisällön lisäksi arvioitu tyyliä ja kontekstia ja koetettu tavoittaa kirjoittajan intentiota. Olennaista arvioinnin kannalta on, tulkitaanko kirjoittajan intentioksi argumentoida ja perustella jotakin kantaa vai tarkoituksellisesti leimata tai halventaa. Esimerkiksi poliittisen kannan perustelua ja argumentointia ei ole välttämättä tulkittu vihapuheeksi, vaikka argumentoinnin kohteena oleva kanta tai mielipide voitaisiin nähdä syrjivänä. Poliittisiin mielipiteisiin tai muuhun ajatteluun kohdistuvaa arvostelua ja kritiikkiä ei ole tulkittu vihapuheeksi, vaikka arvostelu olisi alatyylistäkin. Hankalissa ja epäselvissä tapauksissa on palattu määritelmiin yhdenvertaisuuslain häirintäpykälässä, tasa-arvolain syrjintäpykälässä, rikoslain pykälässä kiihottamisesta kansanryhmää vastaan sekä Valtakunnansyyttäjänviraston vihapuhetta koskevassa raportissa.

Yhdenvertaisuuslain häirintäpykälän (§14) mukaan käyttäytyminen on häirintää, jos se tarkoituksellisesti tai tosiasiallisesti loukkaa ihmisarvoa, jos loukkaava käyttäytyminen liittyy (yhdenvertaisuuslain) syrjintäpykälässä (§8) tarkoitettuun syyhyn (ikä, alkuperä, kansalaisuus, kieli, uskonto, vakaumus, mielipide, poliittinen toiminta, ammattiyhdistystoiminta, perhesuhteet, terveydentila, vammaisuus, seksuaalinen suuntautuminen tai muu henkilöön liittyvä syy) ja käyttäytymisellä luodaan halventava, nöyryyttävä, uhkaava, vihamielinen tai hyökkäävä ilmapiiri.

Tasa-arvolain syrjintäpykälän (§7) mukaan seksuaalista- tai sukupuoleen perustuvaa häirintää on henkilön sukupuoleen, sukupuoli-identiteettiin tai sukupuolen ilmaisuun liittyvä ei-toivottu käytös (seksuaalinen tai ei-seksuaalinen), jolla tarkoituksellisesti tai tosiasiallisesti loukataan tämän henkistä tai fyysistä koskemattomuutta ja jolla luodaan uhkaava, vihamielinen, halventava, nöyryyttävä tai ahdistava ilmapiiri. Kiihottamisrikoksesta puolestaan puhutaan silloin, kun on kyse uhkaamisesta, panettelusta, solvaamisesta rodun,

ihonvärin, syntyperän, kansallisen tai etnisen alkuperän, uskonnon tai vakaumuksen, seksuaalisen suuntautumisen tai vammaisuuden perusteella taikka niihin rinnastettavalla muulla perusteella.

Valtakunnansyyttäjänviraston raportti (2012) ottaa kantaa siihen, millainen puhe on tunnistettavissa laissa kielletyksi vihapuheeksi tai rasismiksi. Tähän perustuen keskustelupalstoilla on usein käytäntönä kieltää muun muassa kunnianloukkaukset ja yksityiselämää loukkaavat sisällöt. Verkkosisältöjen moderoinnissa lakiin perustuvia ohjeita ovat muun muassa ihmisarvon kunnioittaminen ja väkivaltaan kehottamisen kielto. (Pöyhtäri 2015, Valtakunnansyyttäjänviraston raportti 2012). Tässä projektissa on tulkittu vihapuheeksi lisäksi sellaiset kirjoitukset, jotka käyttävät solvauksen välikappaleena jotakin yllä mainituissa lakipykälissä määritellyistä syistä (esimerkiksi solvauksena esitetty *autisti* tai *vammainen*). Verkkokirjoituksen tapauksessa ei ole aina mahdollisuutta tietää, liittyykö jokin solvaus kohteensa reaaliseen ominaisuuteen. Tämän kaltainen solvaaminen kuitenkin esittää tietyn ryhmän halventavassa valossa sekä normalisoi ryhmään liittyvää halventavaa puhetta.

2.3 Vihapuheen määritelmä

Tässä raportissa vihapuhe määritellään seuraavasti: Vihapuhe on halventavia, nöyryyttäviä, uhkaavia, vihamielisiä, hyökkääviä tai epäinhimillistäviä ilmaisuja, jotka

1. liittyvät henkilökohtaisiin ominaisuuksiin

a. ikä
b. sukupuoli, sukupuoli-identiteetti tai sukupuolen ilmaisu
c. seksuaalinen suuntautuminen
d. etninen tausta (ihonväri, syntyperä, kieli)
e. uskonto tai vakaumus
f. poliittinen kanta
g. ruumiin toimintakyky
h. ulkonäkö
i. kansallisuus
j. muu

2. leimaavat ihmisryhmää tai yleistävät

a. ikäryhmää
b. sukupuoleen, sukupuoli-identiteettiin tai sukupuolen ilmaisuun liittyvää
a. seksuaalista suuntautumista
b. etnistä taustaa (ihonväri, syntyperä, kieli)
c. uskonnollista tai vakaumuksellista
d. poliittista kantaa

e. ruumiin toimintakykyyn liittyvää
f. ulkonäköön liittyvää
g. kansallisuutta
h. muuta

3. kohdistuvat ammattiryhmän edustajaan siten, että kritiikki kohdistuu ammattiroolin ulkopuolelle henkilökohtaisiin ominaisuuksiin tai johonkin hänen edustamaansa ihmisryhmään

a. poliitikot ja päättäjät
b. virkamiehet
c. toimittajat
d. syyttäjät ja lainvalvonta, tuomarit, poliisi
e. tutkijat, asiantuntijat
f. julkiset henkilöt, somevaikuttajat
g. jonkin asian tai ryhmän puolesta toimivat henkilöt
h. muut

4. Muut ilmaisut, joiden intention voi kontekstin perusteella tulkita

a. olevan jotakin edellä mainituista tai
b. kehottavan tai houkuttelevan syrjivään toimintaan tai väkivaltaan yksityishenkilöä tai ihmisryhmää kohtaan tai pitävän tällaista toiminta hyväksyttävänä

3 Aineistot

Raportissa analysoidaan sosiaalisen median kommentteja julkisesta internetistä. Suurempi käytetyistä aineistoista on ostettu ulkopuoliselta palveluntarjoajalta, Mohawk Analyticsiltä, joka hakurobottien avulla seuraa suomenkielistä julkista sosiaalista mediaa, ja tallentaa viestejä suurimpien sivustojen osalta jopa muutamassa sekunnissa viestin ilmestyttyä sivustolle. Tätä aineistoa kutsutaan tässä raportissa nimellä ***"sosiaalisen median aineisto"***. Toinen, pienempi aineisto on kerätty käsin Ylilauta- ja Hommaforum-sivustoilta, joilta hakurobottien vierailut on estetty. Tätä aineistoa kutsutaan nimeltä ***"Ylilauta ja Hommaforum -aineisto"***.

Sosiaalisen median aineisto antaa kattavan kuvan suomenkielisestä sosiaalisesta mediasta. Se käsittää mm. suurimmat keskustelupalstat ja uutissivustot, blogeja, sekä tietysti suurten kansainvälisten sosiaalisen median jättien alustoilla julkaistuja keskusteluja. Aineisto on kerätty lähes reaaliajassa ja siitä löytyy itse viestin lisäksi muita tietoja, kuten linkki alkuperäiseen viestiin, tieto sivustosta ja sivustotyypistä, viestin aikaleima, keräysaika, ja joissakin tapauksissa myös tieto viestin kirjoittajasta, edellisestä viestistä tai viestiketjusta. Tällainen tieto on saatavilla jo silloin, kun keskustelu on käynnissä.

Sosiaalisen median aineiston monista eduista huolimatta aineistossa on myös rajoitteensa tämän raportin mukaisessa analyysissa. Tärkein rajoitteista on, että aineistosta puuttuvat sellaiset sosiaalisen median viestit, jotka eivät ole saatavilla julkisessa internetissä. Mukana ei ole esimerkiksi suljetuissa Facebook-ryhmissä julkaistuja viestejä tai ei-julkisten Facebook-tilien kommentteja. Lisäksi esimerkiksi WhatsApp-viestit, deittisivustojen sisällöt ja erilaisten yhteisöjen sisäänkirjautumista vaativat palstat jäävät tämän raportin ulkopuolelle.

Myös viestien automaattinen kerääminen hakuroboteilla asettaa rajoitteensa: viestejä keräämiä hakurobotteja konfiguroidaan käsin ja joitakin sivustoja voi vain puuttua hakurobotin keräyslistoilta. Jotkut sivustot myös kieltävät lähdekoodissaan hakurobottien vierailut. Lisäksi hakurobotti voi vahingossa ohittaa sivuston jonkin osan, jonka seurauksena osion viestit eivät päädy hakurobotin poimimiksi. Rajapintojen kautta kerätyistä viesteistä, esim. Twitteristä ja Facebookista, taas täytyy erikseen määritellä, minkä käyttäjien ja minkä kielen sanoja sisältävät viestit lasketaan suomenkieliseksi sosiaaliseksi mediaksi. Tässä raportissa ei käsitellä tarkemmin sitä, kuinka suuri osa erilaisista sosiaalisen median palveluista on mukana aineistossa ja kuinka kattavasti kullakin sivustolla julkaistut viestit ovat päätyneet aineistoon.

Lisäksi *sosiaalisen median aineistossa* on julkaisu- ja keräysajankohtiin liittyviä raportin tulosten kannalta oleellisia ominaisuuksia: ennakkomoderoinnissa olevat sivustot (esimerkiksi useat uutiskommentointisivustot) moderoivat ikävimmät viestit ennen niiden julkaisua, mutta jälkimoderoinnissa olevat sivustot (esimerkiksi keskustelupalstat) julkaisevat aluksi kaikki kommentit ja moderoivat niitä vasta jälkeenpäin. Tässä tutkimusaineistossa on siis oletettavasti paljon kommentteja, jotka hakurobotti on ehtinyt kerätä ennen kuin ne on poistettu sivustolta. Tietoa kommenttien poistamisesta ei aineistossa kuitenkaan ole.

Sosiaalisen median aineistosta puuttuu joitakin vihapuheen kannalta erityisen merkittäviä ja prominenttina pidettyjä keskustelufoorumeita kuten Ylilauta ja Hommaforum. Nämä sivustot ovat lähdekoodissaan kieltäneet hakurobottien käyttämisen. Ylilaudan ja Hommaforumin sisältämiä viestejä on tässä projektissa kerätty erikseen pieni otos käsin. Viestien kerääminen käsityönä on hidasta, eikä kaikkia tutkimusaikavälin viestejä pystytty tämän projektin puitteissa keräämään. Käsin poimiminen on kuitenkin ainoa keino, jos halutaan noudattaa sivustojen käyttöehtoja. Pienestä käsin kerätystä otoksesta tehdyn analyysin skaalaaminen sivuston koko viestimääriin tuo omat virhelähteensä tuloksiin. On esimerkiksi mahdollista, että otos on sattumalta poimittu viestiketjuista, joissa on enemmän tai vähemmän vihapuhetta kuin sivustolla keskimäärin. Lisäksi on huomioitava, että kyseisten sivustojen viestien kokonaismäärät myös ovat arvioita. Kuitenkin myös *sosiaalisen median aineistossa* viestien määrässä on epävarmuutta.

Sosiaalisen median viestien kokonaisvaltaisen analysointi on ylipäänsä haasteellista. Tässä raportissa käsiteltyjen viestien ja sivustojen määrät ovat kuitenkin suuria, ja siksi voi olettaa, että valtaosa suomenkielisestä sosiaalisesta mediasta, sekä sisällöllisesti valtaosa julkaistuista vihapuhetyypeistä on raportissa mukana.

3.1 Miljoonia suomenkielisiä verkkokirjoituksia ja -viestejä

Suomalaisen sosiaalisen median sisältöjä on tutkittu *sosiaalisen median aineiston*, eli julkisesta internetistä kolmannen osapuolen, Mohawk Analyticsin, hakurobottien avulla poimimien viestien perusteella 1.9.–31.10.2020 väliseltä ajalta. Aineistossa on 11 975 002 pääosin suomenkielistä viestiä. Keskimäärin viestejä on noin 196 000 päivässä ja 6 miljoonaa kuukaudessa. Aineistossa on mukana viestin julkaisualustan tai sivuston mukaan luokitellut sivustotyypit, joiden jakauma näkyy seuraavassa taulukossa.

Taulukko 1. *Sosiaalisen median aineisto* suomalaisesta julkisesta sosiaalisesta mediasta aikavälillä 1.9.–31.10.2020 sivustotyypeittäin. (*) Ylilauta ja Hommaforum eivät ole mukana.

Alustatyyppi	Viestejä	Viestejä/kk
Twitter	5 478 192	2 739 096
Keskustelupalsta*	2 596 214	1 298 107
Instagram	2 482 086	1 241 043
Facebook (julkinen)	774 747	387 374
Uutiskommentti	591 767	295 884
Blogi	51 595	25 798
Muu	401	201
Yhteensä	**11 975 002**	**5 987 501**

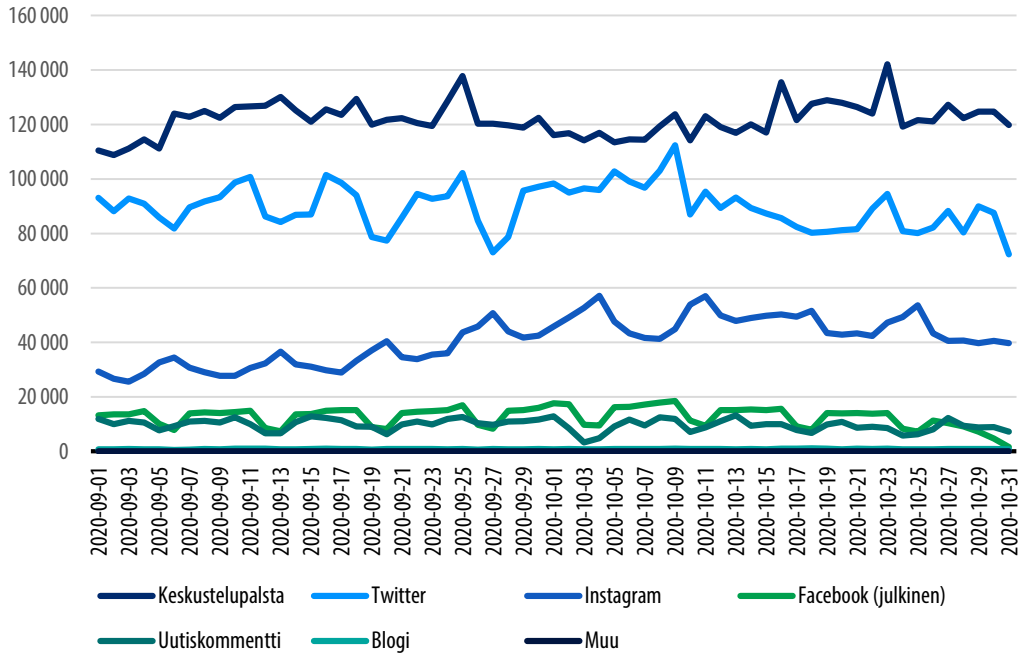
3.2 Ylilauta ja Hommaforum

Koska vihapuheen kannalta kaksi merkittävänä pidettävää sivustoa, Ylilauta ja Hommaforum ovat raportissa esitellyistä muista sivustoista poiketen estäneet hakurobottien tekemän aineiston keräämisen ja ne siis puuttuvat *sosiaalisen median aineistosta*, tehtiin ylimääräinen manuaalinen viestien poiminta tammikuussa 2021 kyseisiltä sivustoilta Näiltä kahdelta sivustolta viestejä poimittiin käsin valitsemalla sivuston antaman keskusteluketjulistan alusta alkaen sellaisia ketjuja, joilla on yli kymmenen viestiä tutkimusaikavälillä, sekä ketjuja eniten luetuilta aihealueilta. Tavoitteena oli saada mahdollisimman kattava otos aineiston viesteistä. Ylilaudalta kerättiin manuaalisesti 3 205 viestiä kategorioista Keho ja seksi, Pelit, Penkkiurheilu, Sekalainen, Tubetus ja striimit, sekä Yhteiskunta ja politiikka. Hommaforumilta kerättiin 343 viestiä kategorioista Kylänraitti, Mylly ja Tupa. Ylilauta linkkaa pääsivuillaan tilastoihinsa, joista tarkasteluaikavälillä syys–lokakuussa 2020 on keskimäärin noin 79 000 viestiä päivässä. Hommaforumilla seurattiin sivuston sadan uusimman viestin julkaisunopeutta ja päädyttiin keskiarvoon noin 300 viestiä päivässä tammikuussa 2021. Syys-lokakuussa 2020 sivuston julkaisumäärät ovat voineet olla erilaisia, mutta suuruusluokka lienee oikeansuuntainen.

3.3 Aineistojen jakaumat

Seuraavassa kuvaajassa esitetään viestien jakautuminen erityyppisille sosiaalisen median alustoille tutkittavalla aikavälillä.

Kuvaaja 1. Viestien jakautuminen aikavälillä 1.9.–31.10.2020 eri julkisen sosiaalisen median alustatyypeille.



Aineistosta on selkeästi havaittavissa viikoittainen vaihtelu: erityisesti julkisen Facebookin, jossa on suurelta osin yritysten ja yhdistysten kirjoittamaa sisältöä, viestit painottuvat viikolle ja viikonloput ovat hiljaisempia. Sen sijaan Instagramin viestimäärät kohoavat viikonloppuisin. Twitterissä viestimäärät ovat usein pienempiä viikonloppuisin, mutta poikkeuksiakin löytyy.

Suomenkielinen julkinen sosiaalinen media jakautuu muutamalle suurelle alustalle, kuten Twitter, Instagram ja Facebook. Kokonaisaineistossa on mukana yhteensä 187 sivustoa, joiden joukossa Suomen suurimmat keskustelupalstat ylilauta.org, vauva.fi ja suomi24.fi. Seuraavassa taulukossa esitellään kaikki aineiston alustat, joilla aineistossa keskustelua esiintyy yli 2 000 viestin verran syys–lokakuussa 2020.

Taulukko 2. Sivustot, joissa esiintyy yli 2 000 viestiä aikavälillä 1.9.–31.10.2020 ja niiden sivustotyypit. (*) Ylilaudan viestimäärä perustuu alustan omaan ilmoitukseen ja Hommaforumin viestimäärä on arvio.

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

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

4 Menetelmät

Yksi tämän raportin tärkeimmistä tuloksista on viedä teoriatasolla tehty vihapuhemääritelmä käytäntöön. Lähestymistavaksi on valittu ihmistyön ja koneoppimisjärjestelmän yhteistyö, jossa tutkija merkitsee käsityönä yksittäisiä sosiaalisen median viestejä vihapuheeksi ja ei-vihapuheeksi. Koneoppimisjärjestelmä eli epäasiallisuuksien tunnistamiseen kehitetty kontekstin tunnistava tekoäly saa nämä tutkijan esimerkit opetusaineistokseen. Se oppii tunnistamaan vihapuhetta ja pystyy käsittelemään valtavan paljon suurempaa kommenttimäärää kuin mitä ihmisvoimin voitaisiin analysoida. Lähestymistavan etu esimerkiksi automaattisiin sanalistoihin pohjautuvaan vihapuheen tunnistamiseen verrattuna on se, että tutkija voi huolellisesti noudattaa ennalta määriteltyjä vihapuheluokkia riippumatta siitä, mitä yksittäisiä sanoja kommenteissa on.

Vihapuheluokkien määrittelyyn ei ole vaikuttanut ajatus siitä, mitä tekoäly voi tai ei voi oppia, vaan luokat on muodostettu tekoälystä riippumatta. Koska käytetty tekoälysovellus pohjautuu tilastolliseen mallinnukseen ja se siis tarvitsee riittävästi opetusaineistoa, tekoäly on tässä hankkeessa opetettu tunnistamaan päätaso, onko kommentti vihapuhetta vai ei, sekä neljä pääluokkaa, joihin opetusaineistoa syntyi kohtuullisesti. Tätä tarkempi analyysi alaluokkiin on tehty henkilöarviointina.

Tekoäly opetetaan tässä hankkeessa niin, että arvioitavana on ainoastaan viestin sisältö ja sen semanttinen merkitys. Hankkeessa käytetty tekoäly oppii tunnistamaan tekstin semanttisen kontekstin, joten sen toimintalogiikka on erilainen kuin yksittäisten sanojen esiintymiseen perustuvassa mallinnuksessa. Sen sijaan pintatason konteksti, eli esimerkiksi se, kuka viestin kirjoitti tai tieto alustasta, jolla viesti on julkaistu, ei ole mukana opetusaineistossa. Opetusaineisto sisältää pelkästään tekstimuodossa olevaa dataa, joten myöskään kuvia, videoita tai linkkejä ei ihmisen tekemässä annotoinnissa eikä tekoälymallinnuksessa oteta huomioon. Näin mallinnus on mahdollisimman tasapuolista eri alustoiden ja kirjoittajien välillä: vain viestin semanttinen sisältö ratkaisee.

Mallinnus olisi ollut mahdollista tehdä myös niin, että viestiä edeltävä viesti (johon analysoitava viesti vastaa) on mukana mallinnuksessa. Joissakin tapauksissa edeltävällä viestillä voi olla suuri merkitys uuden viestin tulkintaan. Tässä aineistossa ei kuitenkaan ollut saatavilla edellisen viestin tietoja kaikissa tapauksissa, ja nämä jätettiin siksi pois analyysista. Tulevaisuudessa olisikin mielenkiintoista tutkia myös sitä, miten tekstikontekstin huomioon ottaminen muuttaisi tuloksia.

4.1 Annotointi

Ensimmäiset tutkijan analysoitavaksi tulevat viestit seulottiin suuresta viestimassasta käyttäen samaa sanalistaa, jota Poliisiammattikorkeakoulu käyttää apuna vuosittaisen viharikosraportin tekemisessä. Lisäksi seulonnassa käytettiin sanoja, jotka liittyivät joihinkin reaalimaailman kyseisen aikavälin tapahtumiin, joiden aikana vihapuhetta saattaisi odotusarvoisesti ilmetä (kuten esimerkiksi Pride-viikko). Sanalistaa käytettiin pelkästään annotointityön ensimmäisessä vaiheessa tutkijan käsin tekemän annotoinnin helpottamiseksi, ei osana tekoälymallia. Sanalistojen avulla valikoitui tutkijan annotoitavaksi paljon enemmän vihapuheeksi luokiteltavia kommentteja kuin satunnaisessa järjestyksessä annotoimalla olisi tullut. Näin haluttiin sekä maksimoida tekoälyn saama oppimisaineisto että kohdistaa tutkijan huomio heti projektin alussa aihepiireihin, joiden ympärillä valtaosa vihapuheesta esiintyy. Sanalistoilla seulotuista viesteistä 8.7 % merkittiin annotoinnissa vihapuheeksi.

Lopullisessa vihapuheeksi tekoälyn avulla tunnistetussa joukossa 66 % viesteistä oli alun perin osunut sanalistoihin. Toisaalta ei-vihapuheeksi tunnistetuista viesteistä 6.9 % osui myös samoihin sanalistoihin. Tämä kertoo siitä, että sanalistat toimivat mainiosti alustavana annotointityön tehostajana, mutta eivät vaikuta merkittävästi lopulliseen vihapuhetta tunnistavaan tekoälymalliin. Samoja sanalistoja on käytetty apuna myös aikaisemmissa tutkimuksissa, joten niille on kertynyt huomattava määrä relevanttia sanastoa. On kuitenkin huomioitavaa, että sanalistojen käytöllä voi olla vaikutusta projektin tuloksiin ja joitakin listan ulkopuolella olevia, ehkä uusia, vihapuheen osa-alueita on voinut jäädä katveeseen.

4.2 Käytännön valintoja ja tulkintoja

Kommentteja annotoidessa tiettyjä erityisen halventavia sanoja sisältävät tekstit tulkittiin kategorisesti vihapuheeksi. Tällaisia sanoja olivat esimerkiksi etniseen taustaan viittaavat halventavat sanat kuten *neekeri, jutku, ählämi* ja sukupuoleen tai seksuaaliseen suuntautumiseen viittaavat *huora, hintti, hinttari.* Jotkut samojen aihealueiden sanat saattoivat esiintyä halventavassa yhteydessä mutta myös vähemmän loukkaavassa tai myönteisessä kontekstissa. Esimerkiksi sanaa *homo* käytetään sekä myönteisessä että halventavassa ja leimaavassa merkityksessä. Sen kohdalla pyrittiin tulkitsemaan kirjoittajan intentiota ja tekstin sävyä.

Etniseen taustaan ja kansallisuuteen viittaavat sanat *ryssä* ja *mustalainen* ovat tässä raportissa myös sanoja, joiden osalta konteksti vaikutti tulkintaan sanojen merkityshistorian vuoksi. *Mustalainen* on ollut suomen kielessä myös viranomaistermi, ja vaikka se ei enää

ole virallisissa yhteyksissä hyväksyttävä, sanaa esiintyi aineistossa myös ei-halventavaksi tulkitussa kontekstissa. Tällainen kommentti, jossa kirjoittajan intentiota ei pidetty halventavana eikä kirjoitusta vihapuheena, voisi olla esimerkiksi seuraava:

> *"Vaaleus on väistyvä ominaisuus. […] Itse muistan itäsuomalaiselta ala-asteelta 1970-l. Että luokassa ei ollut yhtään tummahiuksista tai ruskeasilmäistä oppilasta. Ylä-asteella oli yksi, hänen isänsä oli mustalainen."*

Ryssä-sanan tapauksessa kirjoitusta ei tulkittu vihapuheeksi, jos sana esiintyi sotahistoriallisessa kontekstissa tai jos sanaa käytettiin Venäjästä maailmanpoliittisena toimijana. Esimerkiksi seuraavaa ei näillä perustein olisi tulkittu vihapuheeksi:

> *"Tekivät työnsä. Ryssä ei vallottanu maata."*

Aineiston joukossa on runsaasti kirjoituksia, jotka ovat ns. harmaalla alueella. Tällaisissa tapauksissa ei voi yksiselitteisesti sanoa, onko kirjoittajan intentio halventaa ja leimata, vai pyrkiikö hän ensisijaisesti argumentoimaan tietyn poliittisen kannan tai mielipiteen puolesta. On mahdollista, että joissakin tapauksissa olisi voitu perustellusti tulla myös toisenlaiseen ratkaisuun kuin mihin tulkinnassa päädyttiin.

4.3 Määrittelyyn liittyviä huomioita

Annotoinnissa tuli vastaan lisäksi seuraavia luokitteluun vaikuttavia huomioita:

1. Määrittelyssä ei ollut selkeää luokkaa ilmauksille, joissa ei viitata kenenkään yksilön tai ryhmän todellisiin ominaisuuksiin, mutta käytetään etnisyyttä, sukupuolta, seksuaalista suuntautumista jne. solvauksen välineenä. Annotoinnissa tällaiset kommentit merkittiin luokkaan 4 (Intentio tai solvaus). Katso myös luku 5.4. Esimerkkejä: *"huora", "tuommoista jutkupaskaa".*

2. Annotointivaiheessa jätettiin pois yksi kokonainen luokka, joka määritelmään oli alun perin kirjoitettu. Pois jätettyyn luokkaan olisivat kuuluneet sellaiset halventavat ja leimaavat viestit, jotka *kohdistuvat yksilöön sen perusteella, että hänen oletetaan kuuluvan johonkin ihmisryhmään.* Tällainen vihapuheen määrittely on relevantti esimerkiksi syrjintätapausten oikeuskäsittelyissä. Verkkoviestien luokittelussa ero luokkaan 1 (Henkilökohtaiset ominaisuudet) ei ollut tarpeeksi selkeä, ja tämä luokka jätettiin siksi kokonaan pois määritelmästä.

3. Ammattiin liittyvä vihapuhe jakaantui luokkiin 2 (Ihmisryhmä tai yleistys) ja 3 (Ammattiryhmä). Kommentit, jossa yleistetään jonkin ammatin edustajia halventavasti, merkittiin luokkaan 2. Luokkaan 3 sen sijaan merkittiin sellainen vihapuhe, joka kohdistuu ammattiryhmän edustajaan siten, että kritiikki kohdistuu ammattiroolin ulkopuolelle henkilökohtaisiin ominaisuuksiin.

4. Vihapuheeksi tunnistettujen viestien alaluokittelussa tulkittiin monessa tapauksessa, että muslimeihin viitattiin teksteissä paremminkin etnisenä ryhmänä kuin uskontokuntana. Tällaisissa tapauksissa luokitteluun merkittiin, että kritiikki kohdistuu etniseen taustaan (d). Luokkaan e (uskonto tai vakaumus) tulkittiin liittyväksi ne kommentit, jossa mainittiin islam tai muu uskonto tai viitattiin eksplisiittisesti uskonkysymyksiin.

4.4 Utopia AI Moderator

Utopia AI Moderator on tekstianalytiikkaa ja koneoppimista hyödyntävä kieliriippumaton SaaS-palvelu, joka oppii tietyn sivuston oman moderointilinjan. Tässä projektissa sitä on käytetty tunnistamaan tutkijan määrittelemää vihapuhetta. Vihapuhe on huomattavasti harvinaisempaa ja vakavampaa kuin se epäasiallinen sisältö, jota nettipalveluiden tarjoajat usein poistavat sivustoiltaan. Harvinaisuutensa vuoksi vihapuhetta on tilastolliseen mallinnukseen pohjautuvan tekoälyn vaikeampi tunnistaa kuin yleisempää epäasiallista sisältöä. Toisaalta vakavuutensa vuoksi Utopia AI Moderatorin kaltainen tekoäly oppii vihapuhetta helpommin, koska se semanttisesti merkittävästi poikkeaa muusta sosiaalisen median kommentoinnista.

Utopia AI Moderator ei itse määrittele, mikä on vihapuhetta tai epäasiallista sisältöä ja mikä ei. Se käyttää aina kunkin sosiaalisen median palvelun omia ihmispäätöksiä hyväksyttävästä ja epäasiallisesta sisällöstä. Utopia AI Moderator olettaa, että kukin sosiaalisen median palvelu on määritellyt omat käyttöehtonsa ja kertonut selkeästi käyttäjilleen, minkälainen sisältö on hyväksyttävää ja mikä ei ko. palvelussa. Lisäksi Utopia AI Moderator olettaa, että ihmismoderoijat ovat toimineet johdonmukaisesti moderointityötä tehdessään, jotta tekoälyjärjestelmä ei saa ristiriitaista tietoa siitä, minkälaista sisältöä tulee hyväksyä ja minkälaista ei. Käytännössä ihmispäätöksissä on aina jonkin verran vaihtelua ja tähän tekoäly tuo arvokasta lisää johdonmukaisuudellaan.

Utopia AI Moderator on kieliriippumaton tekstianalytiikkajärjestelmä, joka koostuu Utopian kehittämistä koneoppimismalleista, SaaS-rajapinnasta ja koneoppimismallien säännöllisestä päivittämisestä ja ylläpidosta. Yritykset, jotka käyttävät Utopia AI Moderatoria, seuraavat erityisesti tuotteen merkitsemiä rajatapauksia tai uudenlaisia sisältöjä. Näillä

ihmispäätöksillä tuote saa jatkuvasti hiukan lisää opetusaineistoa ja Utopian tekoäly-asiantuntijoiden ylläpitämänä ja edelleenkehittäminä koneoppimismallit säilyvät ajan tasalla ja noudattavat asiakasyritysten hiljalleen päivittyvää ja edelleenmuotoutuvaa moderointilinjaa.

Utopia AI Moderator on alun perin kehitetty suomen kielelle, mutta kieliriippumattomuutensa vuoksi se käy kaikenkielisen käyttäjien luoman sisällön moderointiin. Tuote ymmärtää sosiaalisen median epätäsmällistä kieltä, eikä se häiriinny kirjoitusvirheistä, puhekielisyyksistä, vieraskielisistä sanoista tai lauseista, hauskoista emojeista tai muustakaan kieliopista tai sanakirjoista poikkeavasta. Se tunnistaa myös kontekstin. Mikäli opetusaineistoa on riittävästi ja se on yhdenmukaista, oppii Utopia AI Moderator ymmärtämään kommentit, jotka ihmismoderoijat ovat ymmärtäneet.

4.5 Mallinnus

Utopia AI Moderatoria käytettiin projektissa kahteen eri tehtävään: ensin tuottamaan tutkijalle aineistoja, jotka nopeuttavat vihapuhetta sisältävien viestien tunnistamista. Kun tutkija oli annotoinut riittävän määrän viestejä käsin tilastolliselle analyysille, näitä viestejä käytettiin opetusaineistona lopullisen Utopia AI Moderator -tekoälymallin kouluttamiseen.

Utopia AI Moderatorin lopullisena opetusaineistona oli 18 925 tutkijan annotoimaa viestiä, joista 2 471 on tutkijan vihapuheeksi merkitsemää ja loput eivät sisältäneet raportin määritelmän mukaista vihapuhetta. Tämän lisäksi vihapuheluokat oli erillisenä annotointina merkitty 1 902 viestiin, joilla opetettiin toinen koneoppimismalli tunnistamaan näitä luokkia.

5 Tulokset

Tekoäly oppi tutkijan annotoiman opetusaineiston perusteella tunnistamaan vihapuhetta huolimatta siitä, että opetusaineisto oli pienehkö ja viestien luokittelu oli ihmisellekin haastava tehtävä. Käsin annotoidun, opetusaineiston ulkopuolisen testijoukon perustella tunnistustarkkuus oli 98.6 %. Tekoäly myös järjesti viestit järjestykseen sen mukaan, kuinka varmasti se piti viestiä vihapuheena. Joissakin tapauksissa tekoäly merkitsi viestin toisin kuin ihmisannotoija. Korkealla varmuudella vihapuheeksi ”virheellisesti” tunnistetut viestit olivat lähes poikkeuksetta vihapuheen rajatapauksia. Näin ollen havaittiin, että jo kohtuullisen pienellä opetusaineistolla pystytään nopeasti toteuttamaan erittäin tarkka tekoälymalli, joka pystyy automaattisesti reaaliajassa käsittelemään valtavia määriä sosiaalisen median viestejä ja merkitsemään selkeimmän vihapuheen. Projektissa opetettua tekoälymallia voidaan sellaisenaan käyttää tunnistamaan vihapuhetta sosiaalisen median viesteistä. Tekoälymallin tarkkuus on myös riittävä vertailemaan eri sivustoja toisiinsa. Mikäli opetettua Utopia AI Moderator -tekoälymallia käytettäisiin tuotantokäytössä, sen löydöksiä arvioisivat vihapuheen tunnistamisen ammattilaiset. He vertailisivat säännöllisesti pienen määrän tekoälyn tekemiä päätöksiä käyttämäänsä vihapuhemääritelmään ja antaisivat arvionsa tekoälylle uudeksi opetusaineistoksi. Näin tuotetta käytettäessä opetusaineisto lisääntyisi ja tekoälymallin tarkkuus kasvaisi edelleen.

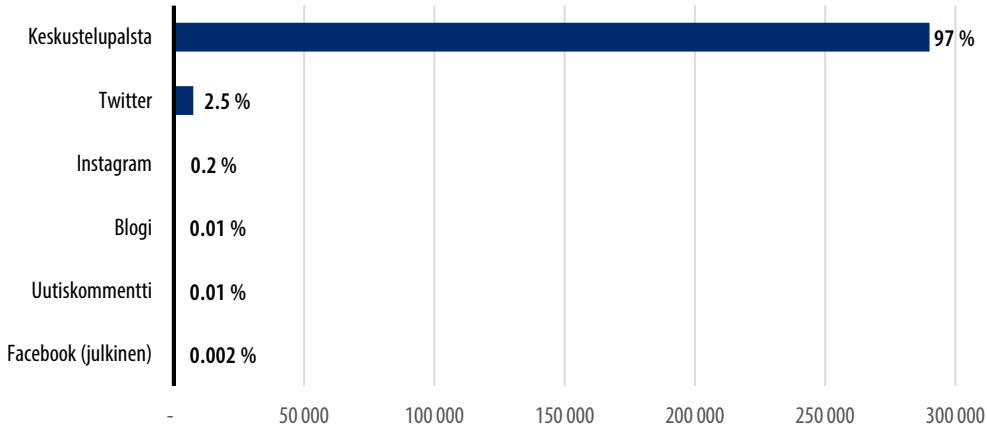
Tässä projektissa tekoälymallin opettamiseen ei käytetty Ylilauta- ja Hommaforum-sivustoilta käsin kerättyä pienempää aineistoa. Jos käsin kerätyn aineiston koko olisi ollut suurempi, olisi myös näiden sivustojen viestejä voitu käyttää osana tekoälymallin opetusta. Tekoälyä ei siis käytetty *Ylilauta- ja Hommaforum -aineiston* analysoimiseen, vaan niiden viestit päädyttiin annotoimaan kokonaisuudessaan käsin samalla tavoin kuin tekoälyn opetusaineiston osuus annotoitiin *sosiaalisen median aineistosta*. Näin varmistettiin myös, että pienestä otoksesta laskettu vihapuhemäärä suhteutettuna suureen aineistoon on mahdollisimman tarkasti oikein.

Ylilaudalta kerätystä 3 205 viestin aineistosta vihapuheeksi tunnistettiin käsin annotoimalla 190 viestiä eli 5.9 %. Hommaforumilla vihapuheeksi tunnistettiin 30 viestiä (8.7 %) aineiston 343 viestistä. Nämä prosenttiosuudet arvioidusta sivuston kokonaisvolyymista tarkoittaisivat arviolta noin 285 000 vihapuheviestiä Ylilaudalla ja noin 1 600 vihapuheviestiä Hommaforumilla tutkittuna ajankohtana 1.9.–31.10.2020. Erityisesti Ylilaudan massiivinen vihapuhemäärä herättää kysymyksiä tulosten vertailukelpoisuudesta *sosiaalisen median aineiston* sivustojen kanssa. Koska sekä Ylilauta että Hommaforum ovat kieltäneet automaattisen sisällön keräämisen, on tällainen manuaalinen otanta käytännössä ainoa tapa arvioida sivustojen vihapuheen määrää. Vihapuheeksi ihmisen arvioimana annotoitujen viestien määrä *sosiaalisen median aineistossa* oli alle 1 %, kun Ylilaudalla se oli 5.9 % ja Hommaforumilla 8.7 %. Vaikka arvioidut määrät eivät olisi täsmälleen oikein, voidaan luotettavasti nähdä, että vihapuheen osuus mainituilla sivustoilla on muuta suomalaista sosiaalista mediaa merkittävästi korkeampi.

5.1 Vihapuheen alustat

Tämän raportin määritelmän mukaista vihapuhetta esiintyy käytettyjen aineistojen perusteella julkisilla suomenkielisillä alustoilla verkossa noin 150 000 viestiä kuukaudessa. Kahden kuukauden tarkasteluajanjaksolla 1.9.–31.10.2020 tunnistettiin 298 032 vihapuheviestiä, joista 97 % esiintyi erilaisilla keskustelupalstoilla. Seuraavaksi yleisimmät alustat olivat Twitter (2.5 %) ja Instagram (0.2 %). Blogit, uutiskommentit ja Facebookin julkinen osuus kattavat alle 0.02 % kaikesta tunnistetusta vihapuheesta. Tuloksia tarkastellessa on hyvä muistaa, että Facebookin suljetut ryhmät ja ei-julkiset tilit eivät ole mukana aineistossa (ks. luku 3). Käytännössä Facebookin yksityisissä ryhmissä julkaistu vihapuhe jää kokonaan katveeseen tässä raportissa, ja oletusarvoisesti Facebookin osuus vihapuheen alustana on huomattavasti suurempi kuin miltä tämän aineiston valossa näyttää. Vihapuheen (298 032 viestiä) osuus kaikista 16.8 miljoonasta viestistä on 1.8 %.

Kuvaaja 2. Vihapuheeksi tunnistettujen viestien jakautuminen eri julkisille sosiaalisen median alustatyypeille. Facebook kattaa vain julkisen osuuden.



Merkittävin vihapuhetta sisältävää keskustelupalsta on Ylilauta (arviolta 285 000 viestiä, 98 % keskustelupalstojen vihapuheeksi tunnistetuista viesteistä, 96 % kaikkien alustojen vihapuheeksi tunnistetuista viesteistä). Määrällisesti seuraavaksi eniten, mutta merkittävästi vähemmän, vihapuhetta sisältävät suomi24.fi (1 931 viestiä), Hommaforum (arviolta 1 600 viestiä) ja vauva.fi (796 viestiä). Hommaforumilla vihapuheen suhteellinen määrä oli 8.7 % kaikista alustan viesteistä, Ylilaudalla 5.9 %, murha.info:ssa 1.3 % ja nyymichan.fi-sivustolla 0.62 %. Keskiarvo keskustelupalstoilla julkaistun vihapuheen määrälle on 3.9 %. Tämä selittyy pääosin Ylilaudalla ja Hommaforumilla esiintyvästä vihapuheesta. Ilman niitä keskustelupalstojen vihapuheosuus olisi 0.13 %.

Kaikki sivustot, joilla vihapuhetta on tunnistettu tässä raportissa vähintään viisi viestiä, on listattu seuraavaan taulukkoon.

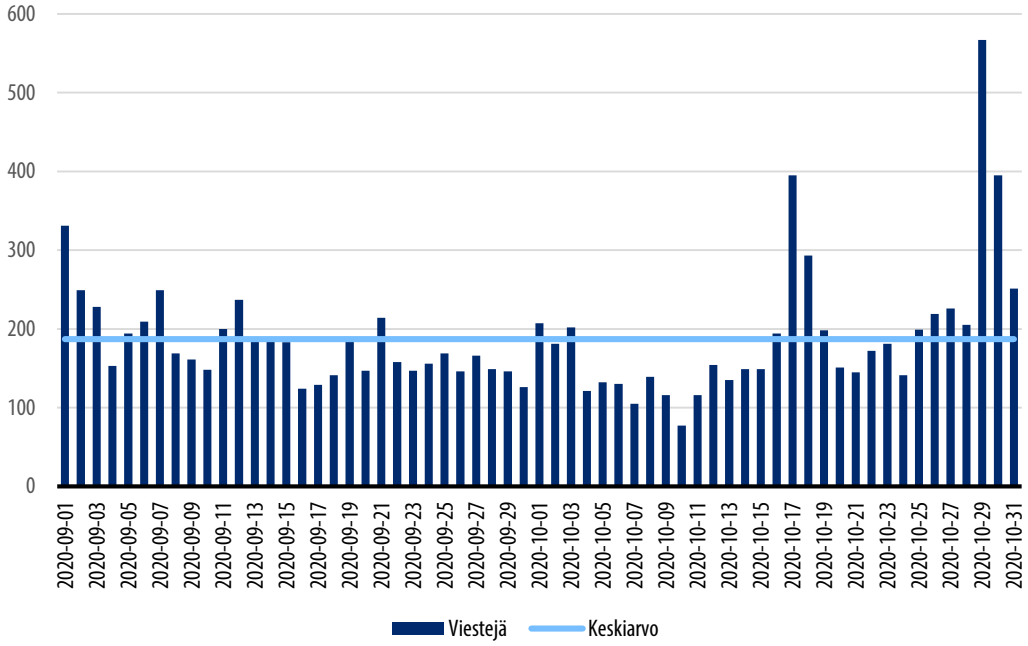
Taulukko 3. Vihapuheeksi tunnistetut viestit eri alustoilla ja osuudet kullakin alustalla. (*) Ylilaudan ja Hommaforumin vihapuheen kokonaismäärät ovat arvioita.

Alustatyyppi ja sivusto	Vihapuheviestejä	Viestejä yhteensä	Viestejä kuukaudessa	Vihapuheen osuus
Twitter	**7 450**	**5 478 192**	**2 739 096**	**0.14 %**
twitter.com	7 450	5 478 192	2 739 096	0.14 %
Keskustelupalsta	**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 %
Muu	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 %
Blogi	**27**	**51 595**	**25 798**	**0.05 %**
puheenvuoro.uusisuomi.fi	20	35 511	17 756	0.06 %
Muu	7	16 084	8 042	0.04 %
Uutiskommentti	**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 %
Muu	6	263 707	131 854	0.002 %
Facebook (julkinen)	**5**	**774 747**	**387 374**	**0.001 %**
www.facebook.com	5	774 747	387 374	0.001 %
Muu	**0**	**401**	**200**	**0.0 %**
Yhteensä	**298 032**	**16 811 302**	**8 405 651**	**1.8 %**

5.2 Vihapuheen jakautuminen ajallisesti

Tekoälyn avulla läpikäytyä *sosiaalisen median aineistoa*, joka sisältää muut alustat kuin Ylilaudan ja Hommaforumin, voidaan tarkastella myös tarkemmin päivätasolla. Raportin määritelmän mukaista vihapuhetta löytyi *sosiaalisen median aineistosta* syys-lokakuussa 2020 aineistosta 11 432 viestiä (0.10 %), keskimäärin noin 187 vihapuheeksi tunnistettua viestiä päivässä. Vihapuhe jakautuu ajanjaksolle seuraavasti:

Kuvaaja 3. Tekoälyn vihapuheeksi tunnistettujen *sosiaalisen median aineiston* viestien jakautuminen aikavälille 1.9.–31.10.2020. Yhteensä aineistossa on 11 432 tekoälyn vihapuheeksi tunnistamaa viestiä. Keskiarvo päivässä vihapuheeksi tunnistetuille viesteille on 187 viestiä.



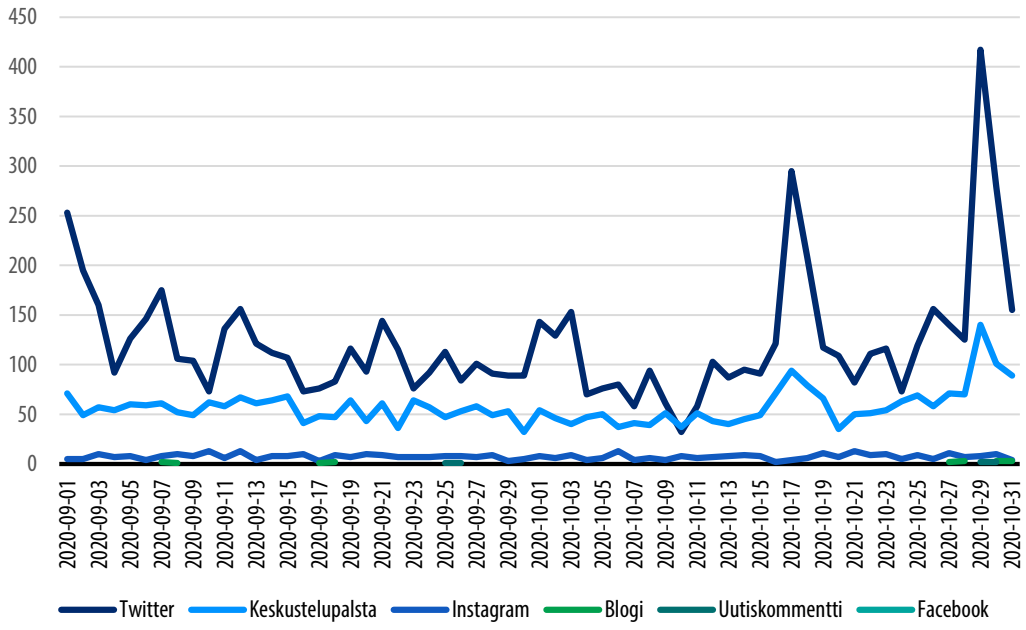
Tarkasteluvälillä oli ainakin kolme tapahtumaa, jotka keräsivät runsaasti huomiota sosiaalisessa mediassa ja joiden yhteydessä on saattanut esiintyä vihapuhetta:

a. Helsinki Pride -viikko 7.9.–13.9.2020 (Pride-kulkue ja hybriditapahtuma 12.9.)
b. 3.10.2020 Elokapina-liikkeen mielenosoitus Helsingissä johtaa tilanteeseen, jossa poliisi käyttää voimakeinona OC-sumutetta.
c. 8.10.2020 Trendi-lehti julkaisee pääministeri Sanna Marinin haastattelun, jonka kuvituksessa Marin poseeraa jakkupuvussa ilman aluspaitaa.

Pride-viikolla, erityisesti sen alkamispäivänä 7.9. ja varsinaisena tapahtumapäivänä 12.9. on nähtävissä kohonneita arvoja tunnistetun vihapuheen määrässä: 249 viestiä (7.9.) ja 237 viestiä (12.9.), noin 30 % yli tunnistetun vihapuheen päiväkeskiarvon. Myös Elokapina-mielenosoituspäivänä 3.10. vihapuheen määrä, 202 viestiä, näyttäisi olevan keskimääräistä suurempi. Toisaalta mielenosoitusta seuraavina päivinä tapausta puitiin runsaasti sosiaalisessa mediassa, mutta vihapuheen määrä näyttäisi noina päivinä olevan jopa keskimääräistä alhaisempi. Niin ikään pääministerin Trendi-haastattelun julkaisupäivänä ja sitä seuraavina päivinä vihapuheen määrä näyttäisi olleen keskimääräistä alhaisempi.

Alla päivätason kuvaaja vihapuheeksi tunnistettujen viestien esiintymisestä *sosiaalisen median aineistossa*.

Kuvaaja 4. Vihapuheeksi tunnistettujen viestien esiintyminen *sosiaalisen median aineiston* ajanjaksolla. Keskustelupalstat eivät sisällä Ylilauta- ja Hommaforum-sivustoja.

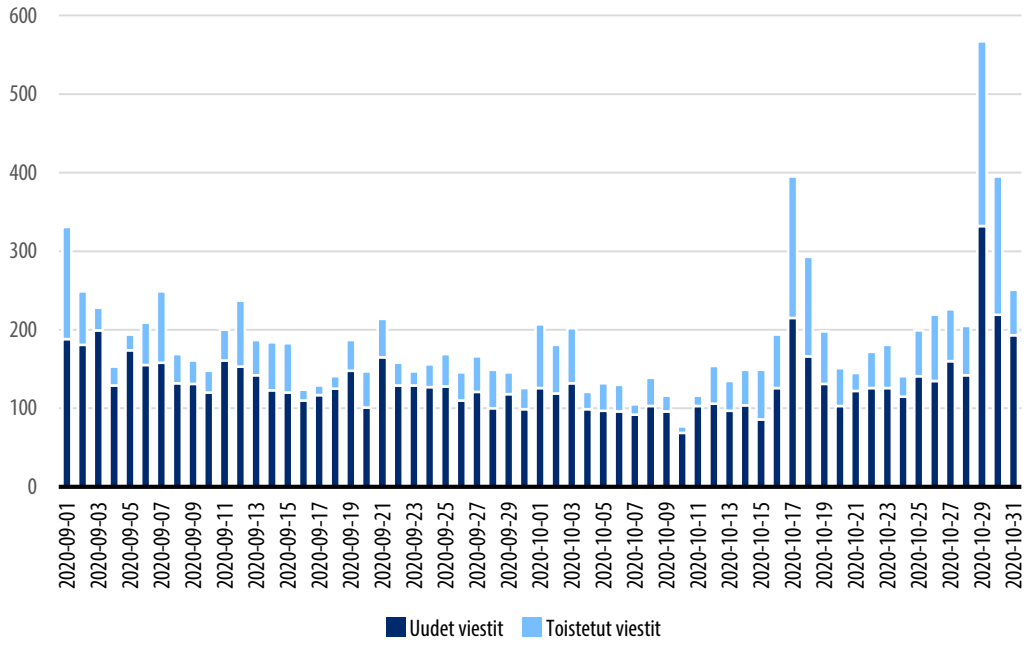


Vihapuheen määrässä on havaittavissa piikit 17.10. ja 29.10. Niitä käsitellään tarkemmin luvussa 6.1.

5.3 Twitter ja uudelleentviittausten voima

Kuvaaja 3 esittelee kaikki tekoälyn vihapuheeksi tunnistamat viestit riippumatta siitä, olivatko ne alkuperäisiä viestejä vai jo julkaistun viestin uudelleen lähetettyjä versioita. Koska uudelleen lähetettyjä viestejä tai duplikaatteja on *sosiaalisen median aineistossa* huomattava määrä, on mielekästä tarkastella, millainen osuus niillä on vihapuheen kokonaisuudesta. Käytännössä lähes kaikki uudelleen lähetetyt viestit olivat Twitterin uudelleentviittauksia. Alla olevasta kuvaajasta käy ilmi, että uudelleentviittausten osuus kaikesta vihapuheesta on suuri erityisesti sellaisina päivinä, kun vihapuhetta julkaistaan paljon. Toistettujen viestien osuus Twitter-viestien joukossa on 39 %.

Kuvaaja 5. Vihapuheeksi tunnistettujen viestien jakautuminen sosiaalisen median aineiston aikavälille 1.9.–31.10.2020, sekä tieto viesteistä, joiden täsmälleen sama sisältö on jo julkaistu aiemmin samassa aineistossa (toistetut viestit).



Tuloksissa on havaittavissa, kuinka Twitter-viestin volyymia voidaan vahvistaa voimakkaasti uudelleentviittauksilla. Uudelleentviittauksiin ei voi luotettavasti suhtautua niin, että ne olisivat erillisten ihmisten tarkoituksellisia päätöksiä toistaa vihaviestejä, mutta toisaalta tätä mahdollisuutta ei voida sulkea poiskaan.

Aineistossa kahdenkymmenen eniten tviitanneen käyttäjätilin joukossa on neljä tiliä, joilla yli 98 % Twitter-viesteistä on uudelleentviittauksia. Kaksi näistä neljästä tilistä tviittaa nimimerkin takaa ja kahdella on oikealta vaikuttava etu- ja sukunimi. Samassa joukossa on myös yhdeksän sellaista tiliä, jotka uudelleentviittaavat vain hyvin harvoin (viesteistä alle 3 % on uudelleentviittauksia). Näistä tileistä neljä tviittaa nimimerkin takaa. Loput seitsemän tiliä sisältävät keskimäärin puolet tviittejä ja puolet uudelleentviittauksia. Näistä seitsemästä tilistä viisi käyttää nimimerkkiä. Näiden kahdenkymmenen aktiivisimmin vihapuhetta tviittaavan käyttäjätilin viestit käsittävät 22 % kaikesta Twitterissä tunnistetusta vihapuheesta.

5.4 Vihapuheen luokittelu

Vihapuhe jaoteltiin raportin vihapuhemääritelmän luvussa 2.3 esitetyn mukaisesti neljään luokkaan:

1. Vihapuhe, jossa syrjitään henkilöä jonkin hänen ominaisuutensa perusteella (ikä, kieli, ulkonäkö, uskonto tai vakaumus, sukupuoli tai sen ilmaisu, seksuaalinen suuntautuminen, etninen tausta tai ruumiin toimintakyky). Esimerkkejä: *”ruma transuhan se on”, ”thaimaan apina”.*

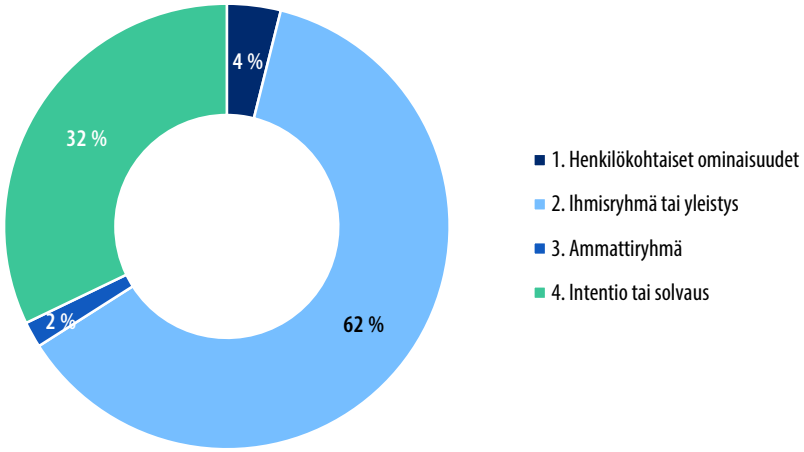
2. Ihmisryhmää leimaava tai yleistävä vihapuhe. *”Raiskaukset ryöstöt väkivalta islam..”*

3. Ammattiryhmään kohdistuva vihapuhe. *”[poliitikon nimi] […] Toivottavasti huora olet saanut koronan ja kuolet”, ”Onkohan [poliitikon nimi] ja näil suvakeilla oikeesti joku poliittinen paritusrinki?”*

4. Muut vihapuheeksi tulkittavat ilmaisut kuten solvaukset ja kehotukset väkivaltaan. Tähän luokkaan merkittiin ne ilmaisut, joiden kohdalla ei ole selvää, kohdistuuko halventava puhe jonkin henkilön todelliseen ominaisuuteen vai halutaanko yleisesti solvata. Solvaukset on tulkittu vihapuheeksi niissä tapauksissa, joissa solvaamisen välineenä käytetään vihapuheen määritelmässä mainittua syytä ja tullaan samalla kohdistaneeksi tätä ryhmää/ominaisuutta kohtaan halveksivaa ja leimaavaa puhetta. Tähän luokkaan merkittiin myös kehotukset väkivaltaan. Esimerkkejä: ”*vitu vammanen*”, ”*homo*”.

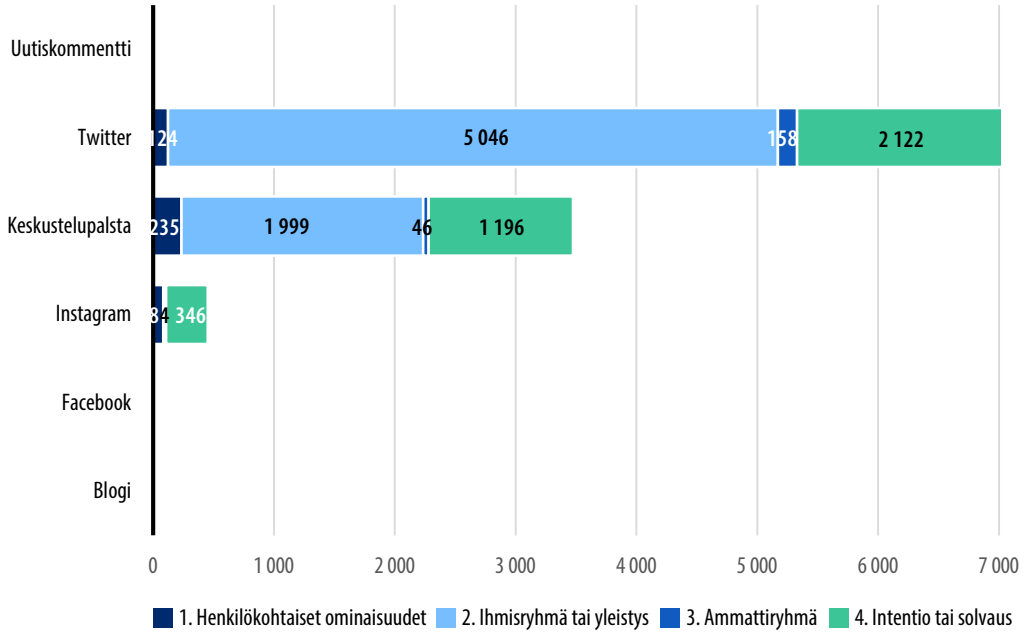
Seuraava kuvaaja näyttää, miten luokat jakautuvat *sosiaalisen median aineistossa*. Tärkeimmäksi tässä tekoälyn avulla käsitellyssä aineistossa, joka ei sisällä Ylilautaa ja Hommaforumia, nousi luokka 2 (Ihmisryhmä tai yleistys), joka käsitti 62 % vihapuheesta. Toiseksi yleisintä ovat luokan 4 (Intentio tai solvaus) viestit (32 %). Henkilökohtaiset ominaisuudet olivat aiheena 4 %:ssa viesteistä ja ammattiryhmä 2 %:ssa.

Kuvaaja 6. Tunnistettujen vihapuheviestien ryhmittely neljään luokkaan *sosiaalisen median aineistossa.*



Määrällisesti suurin osa luokkaan 2 (ihmisryhmää halventava tai yleistävä) merkityistä viesteistä on julkaistu Twitterissä. Toiseksi yleisin vihapuheluokka Twitterissä näyttäisi olevan luokka 4 (halventava intentio tai solvaus). Keskustelupalstoilla samat luokat ovat kärkisijoilla, mutta keskenään melkein yhtä suuria. Instagramissa sen sijaan luokka 4 on ehdottomasti suurin.

Kuvaaja 7. Vihapuheluokkien jakautuminen eri alustatyyppien välillä *sosiaalisen median aineistossa.*



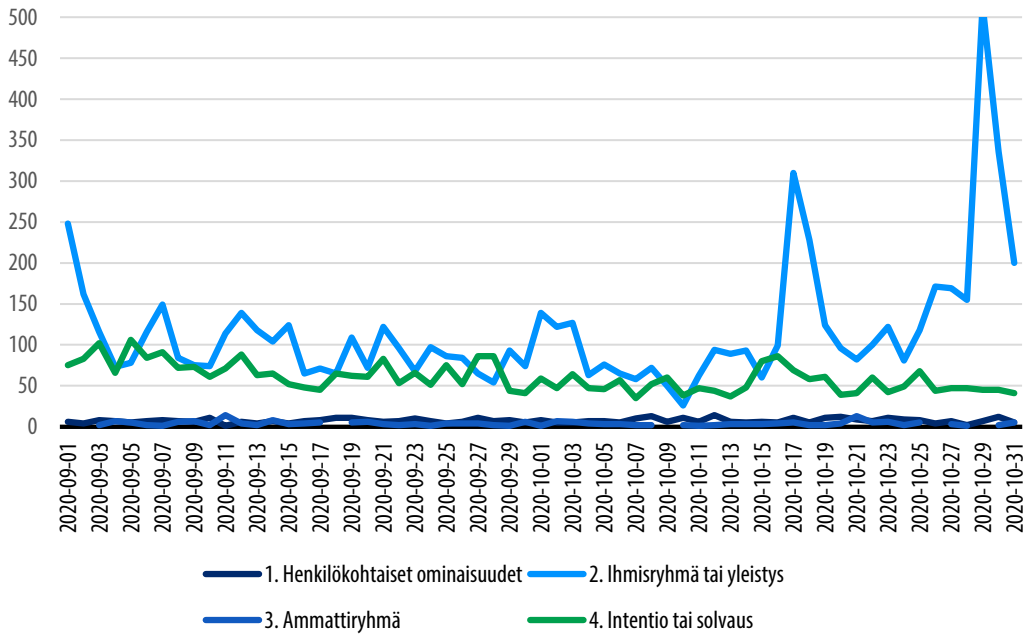
6 Vihapuheen teemoja

Vihapuheen sisältöä tarkastellaan seuraavassa eri näkökulmista: tarkastelemalla tekoälyn tunnistamasta vihapuheesta yksittäisten päivien viestejä, laskemalla, mitkä vihapuheeseen usein yhdistyvät sanat toistuvat aineistossa useimmin, sekä alaluokittelemalla kommentteja raportin määritelmän mukaisiin alaluokkiin.

6.1 Tarkastelussa vihapuheen piikit

Tunnistetun vihapuheen määrässä on havaittavissa huomattavat piikit 17.–18.10. ja 29.–30.10. (ks. Kuvaaja 3). Seuraavassa tarkastellaan lähemmin päivien 17.10. ja 29.10. viestejä. Näiden päivien vihapuheviestit koostuvat ainoastaan luokkaan 2 (Ihmisryhmä tai yleistys) luokitellusta vihapuheesta. Seuraavassa kuvaajassa on esitetty vihapuheluokien jakautuminen *sosiaalisen median aineiston* ajanjaksolle.

Kuvaaja 8. Kunkin päivän vihapuhetyypit *sosiaalisen median aineistossa*.



Twitter-viestien osuus kaikesta tunnistetusta vihapuheesta on näinä kahtena päivänä keskimääräistä selvästi suurempi. Sekä 17.10. että 29.10. Twitter-viestien osuus näissä piikeissä on 75 %, kun muina kuvaajan päivinä Twitterin osuus on 65 %.

6.1.1 Ranskalaisopettajan surman jälkeinen päivä 17.10.2020

Pariisissa tapahtui 16.10.2020 henkirikos, jossa tekijä puukotti opettajan kuoliaaksi ja katkaisi tältä pään. Yle uutisoi tapahtumasta seuraavasti: ”Profeetta Muhammedin pilakuvia luokalle näyttänyt opettaja surmattiin koulun edustalla Ranskassa – ’Islamistinen terrori-isku’” (Yle 16.10.2020).

Vihapuheeksi tekoälyn avulla tunnistettuja viestejä on 17.10. kaikkiaan 395 kappaletta, joista uudelleenjulkaistuja (kahta lukuunottamatta Twitterin uudelleentviittauksia) on 180 kappaletta (46 %). Viestejä manuaalisesti tarkastelemalla selviää, että niissä olevassa tunnistetussa vihapuheessa näyttäisi olevan dominanttina teemana etninen tausta tai uskonto. Kolmessa eniten uudelleentviittauksia saaneessa viestissä mainitaan muslimit ja väkivalta. Kahdessa näistä viitataan eksplisiittisesti pään irrottamiseen.

Tekoälyn 17.10. tunnistamasta vihapuheesta 53 %:ssa esiintyy sana *muslimi* (kaikissa sanayhdistelmissä ja taivutusmuodoissaan). Esiintymistiheys on selvästi suurempi kuin koko aineiston vihapuheeksi tunnistetuissa viesteissä keskimäärin, jolloin se on 26 %. Sana *islam* esiintyy 33 %:ssa viesteistä ja sana *terroristi* 17 %:ssa viesteistä, kun näiden sanojen esiintymistiheys koko aineiston vihapuheeksi tunnistetuista viesteistä on 15 % ja 4 %. Osassa viesteistä esiintyy useampi kuin yksi näistä sanoista. Lisäksi usein esiintyy sana *mielenosoitus* saman Twitter-viestin useissa uudelleentviittauksissa (11 %).

Koneelliseesti tunnistettujen vihapuheviestien teemoja tarkastellessa on tärkeää ottaa huomioon, että vihapuhe on epäeksakti termi. Osa viesteistä on tulkintaa ja harkintaa vaativia tapauksia, eikä vihapuheeksi luokittelu tämänkään jälkeen ole välttämättä yksiselitteistä. Tekoäly muodostaa opetusvaiheessa käsityksen siitä, mihin vihapuhe sijoittuu semanttisessa avaruudessa. Tekoälymallin opetusmateriaalin tärkeän osan muodostavat niin sanotut rajatapaukset ja annotointivaiheessa tehdyt tulkinnat kirjoittajan intentiosta. Myös osa tekoälyn tunnistamasta vihapuheesta on oletusarvoisesti rajatapauksia. Tuloksiin voi vaikuttaa myös se, että tässä hankkeessa tekoälyn opetusaineistossa ei ole huomioitu tekstin ulkoisia tekijöitä kuten esimerkiksi linkkejä, joihin viesteissä viitataan.

6.1.2 Vaikutusvaltaisen tviitin päivä 29.10.2020

Toinen aineistossa havaittava piikki oli 29.10., jolloin tekoälyn vihapuheeksi tunnistamia viestejä julkaistiin kaikkiaan 567 kappaletta. Näistä uudelleenjulkaistuja on 228 viestiä eli 41 %. Myös tämän päivän viesteissä useimmin esiintyvä sana on jälleen *muslimi* (kaikissa taivutusmuodoissaan ja sanayhdistelmissä). Se esiintyy 29.10. peräti 72 %:ssa kaikista vihapuheeksi tunnistetuista viesteistä (koko aineiston vihapuheviestien keskiarvo sanalle on 26 %). Sana *islam* esiintyy 34 %:ssa viesteistä (koko aineiston vihapuheviestien keskiarvo

15 %). Kaikista 29.10. vihapuheeksi tunnistetuista viesteistä vain kuudesosa (17 %) ei sisällä lainkaan sanoja *islam* tai *muslimi*.

Tämän päivän vihapuheeksi tunnistetuista viesteistä neljänneksen (25 %) kattaa yksittäinen tviitti, jota on uudelleentviitattu 138 kertaa. Kyseinen tviitti on samalla *sosiaalisen median aineiston* vihapuheeksi tunnistetuista viesteistä useimmin esiintynyt yksittäinen viesti. Se esiintyy aineistossa kaikkiaan 263 kertaa muodostaen 2.3 % kaikesta *sosiaalisen median aineistosta* havaitusta vihapuheesta.

6.2 Sanastoa

Sosiaalisen median aineistosta tekoälyn avulla 11 432 vihapuheeksi merkitystä viestistä 2 942 viestissä (26 %) mainittiin sana *muslimi* eri taivutusmuodoissa, sanaluokissa ja sanayhdistelmissä. *Islam* eri muotoineen esiintyi 1 665 viestissä (15 %), *neekeri* 1 465 viestissä (13 %) ja *homo* 1 057 viestissä (9.2 %). Lisäksi yli sata kertaa (yli 0.9 %) esiintyivät *suvakki, poliisi, huora, terroristi, somali, ählämi, ja rasisti*. Muita usein esiintyneitä sanoja olivat mm. *allah, apina, feministi, hintti, juutalainen, koraani, maahanmuuttaja, mielenosoitus, monikulttuurisuus, mustalainen, mutakuono, natsi, nekru, pakolainen, rättipää, ryssä, somppu, sosiaalipummi, suvaitsevainen, vammainen,* jotka kaikki esiintyivät vähintään 30 kertaa (0.3 %) tunnistetuissa vihapuheviesteissä. Yksittäisen sanan esiintyminen ei kuitenkaan kategorisesti tee viestistä vihapuhetta, vaan koko viestin sisältö on aina arvioitu. Edellä luetellut sanat ovat pääosin Poliisiammattikorkeakoulun vuosiraportissaan käyttämästä listasta.

Kun nämä yleisimmät sanat yhdistetään vihapuheen luokkiin, saadaan seuraava taulukko.

Taulukko 4. Kunkin vihapuheluokan viesteissä usein esiintyneitä sanoja.

	Usein esiintyneitä sanoja (taivutusmuotoineen)
(1) Henkilökohtaiset ominaisuudet	*huora, homo, vammainen, vähemmistö*
(2) Ihmisryhmän leimaaminen	*muslimi, islam, suvakki, poliisi, somali, ählämi, terroristi, rasisti*
(3) Ammattiryhmä	*terroristi, pakolainen, homo*
(4) Solvaus tai muu vihapuhe	*neekeri, homo, huora, suvakki, poliisi, ryssä, vammainen*

Kun tarkastellaan yleisimpiä viestialustatyyppejä, eri aihepiirit korostuvat. Instagramissa esiintyvässä vihapuheessa yli puolessa (51 %) käytetään termiä *homo*, eikä siellä tämän aineiston perusteella käydä uskontoihin tai maahanmuuttoon liittyvää vihapuheeksi luokiteltavaa keskustelua. Toiseksi yleisin vihapuheviestien sana Instagramissa on *huora* (10 %) *Homo* on myös keskustelupalstojen vihapuheviestien vakiosanastoa (15 %), vaikka vieläkin yleisempi siellä on *muslimi* (17 %). *Islam* mainitaan myös keskustelupalstoilla usein (12 %). Twitterissä *muslimi* on erittäin tärkeä vihapuhetta kuvaava sana: 31 % Twitteristä tunnistetusta vihapuheviesteistä käyttää tätä sanaa. *Islam* on Twitterissä toiseksi yleisin (17 %). Erityisesti Twitterissä suhteessa muihin alustoihin nousee sanan *neekeri* käyttö: se esiintyy jopa 16 % vihapuheeksi tunnistetuista Twitter-viesteistä. Muilla alustoilla yksittäisten sanojen esiintymät olivat niin pieniä, että niillä ei ole tilastollista merkitsevyyttä.

Erillisessä *Ylilauta ja Hommaforum -aineistossa* Ylilaudan vihapuheeksi tunnistetuissa viesteissä olivat erittäin yleisiä sanat *vammainen* ja *huora,* sekä lisäksi *homo* ja *neekeri*. Tämän lisäksi aineistossa korostui nimettyjen henkilöiden terveydentilan yksilöivä arvailu. Hommaforumin viesteissä korostui sana *neekeri,* joka tässä pienessä aineistossa esiintyi 15:ssä (43 %) vihapuheeksi tunnistetuista viesteistä.

6.3 Pieni osa käyttäjistä tuottaa suuren osan vihapuheesta

Kuten luonnollisissa jakaumissa yleensäkin, pieni osa käyttäjistä näyttäisi tuottavan suurimman osan vihapuheesta. *Sosiaalisen median aineistosta* vihapuheeksi tunnistettujen viestien kirjoittajista neljännes (24 %) ei ole kirjoittanut nimimerkillä. Kaikki aineiston anonyymit kirjoittajat ovat julkaisseet viestinsä keskustelupalstoilla. Keskustelupalstojen vihapuheeksi tunnistetuista viesteistä puolestaan kolme neljännestä (76 %) on anonyymien kirjoittamia ja vain 24 % nimetyillä käyttäjätunnuksilla kirjoitettuja.

Sosiaalisen median aineistosta tunnistetusta vihapuheesta löytyy 2 303 eri käyttäjää. Vihapuheviesteissä useimmin esiintyvät 10 käyttäjää vastaavat 11 %:sta kaikesta tunnistetusta vihapuheesta. *Sosiaalisen median aineiston* aktiivisin vihapuheen tuottaja on Twitter-tili, joka julkaisi 352 vihapuheeksi tunnistettua viestiä kahden kuukauden aikana.

6.4 Henkilökohtaiset ja ihmisryhmän ominaisuudet vihapuheessa

Tekoälyn vihapuheeksi tunnistamasta *sosiaalisen median aineistosta* käytiin läpi 500 tekoälyn selkeimmin vihapuheena pitämää tapausta. Ne luokiteltiin tämän raportin vihapuhemääritelmän mukaisiin alaluokkiin käsin. Jos viesti on tulkittu vihapuheeksi useammalla kuin yhdellä perusteella, kaikki perusteet on merkitty. Esimerkiksi kommentin *"vitun homo"* luokitus on c (seksuaalinen suuntautuminen). Kommentin *"Olipa ruma takkutukka!! Kuinka monta ählämiä ja neekeriä sitä on käyttänyt?"* luokitus on b, d, h, koska halveksunnan perusteena ovat b. sukupuoli (oletetusti), d. etninen tausta (*neekeri*-sanan käyttö) ja h. ulkonäkö. Viesteistä 65 %:ssa vihapuhe perustui etniseen taustaan eli ihonväriin, syntyperään tai kieleen. 27 %:ssa viesteistä vihapuhe perustui seksuaaliseen suuntautumiseen ja 20 %:ssa sukupuoleen, sukupuoli-identiteettiin tai sukupuolen ilmaisuun.

Alaluokat jakautuivat seuraavasti:

Taulukko 5. Vihapuheen alaluokitus ja alaluokkien yleisyys aineistossa.

Alaluokka	Esiintymät	% viesteistä
a. ikä	0	0 %
b. sukupuoli, sukupuoli-identiteetti tai sukupuolen ilmaisu	100	20 %
c. seksuaalinen suuntautuminen	136	27 %
d. etninen tausta (ihonväri, syntyperä, kieli)	324	65 %
e. uskonto tai vakaumus	17	3.4 %
f. poliittinen kanta	20	4.0 %
g. ruumiin toimintakyky	22	4.4 %
h. ulkonäkö	18	3.6 %
i. kansallisuus	14	2.8 %
j. muu	4	0.8 %

Yleisin vihapuheen tyyppi näissä viesteissä oli etniseen taustaan, esimerkiksi ihonväriin, syntyperään tai kieleen perustuva vihapuhe (alaluokka d), jota oli 65 %:ssa viesteistä. Esimerkki: *"ei neekereitä Suomeen".*

Toiseksi yleisintä eli seksuaaliseen suuntautumiseen (alaluokka c) kohdistuvaa vihapuhetta löytyi 27 % viesteistä (*"vitun homo"*). Viidennes viesteistä (20 %) sisälsi sukupuoleen (alaluokka b) liittyviä ilmaisuja (*"olet huora"*).

Osa vihapuheeksi tunnistetuista viesteistä olivat moniperusteista vihapuhetta, eli ne kuuluivat useampaan kuin yhteen alaluokkaan: Analysoiduista miltei neljännes (23 %) kuuluu vähintään kahteen alaluokkaan. Vähintään kolmeen alaluokkaan puolestaan kuuluu 6 % viesteistä.

Taulukko 6. Moniperusteinen vihapuhe numeroina: eri vihapuheluokkien lukumäärä kullakin analysoidulla viestillä.

Alaluokkia	Viestejä	Osuus (%)
1	387	77 %
2	83	17 %
3	22	4 %
4	4	1 %
5	4	0.8 %
YHT.	**500**	**100 %**

Yleisin moniperusteisuus on sekä sukupuolta (b) että etnistä taustaa (d) käsittelevät viestit, jota esiintyy 4 %:ssa viesteistä (*”ählämit huumanneet matupatjan”, ”neekerimiesten nussima huora”)*. Toiseksi yleisin moniperusteisuus on sukupuolen (b), seksuaalisen suuntautumisen (c) ja etnisen taustan (d) esiintyminen samassa viestissä (3 % viesteistä). Enimmillään vihapuheen perusteita oli jopa viisi samassa viestissä.

Pienen alaluokitellun aineiston perusteella näyttäisi siltä, että yleisimmät vihapuheen perusteet vaihtelevat hieman eri alustoilla. Twitterissä 91 % tekoälyn varmimmin vihapuheeksi tunnistamista viesteistä kohdistuu etniseen taustaan (d). Keskustelupalstoilla etninen tausta on vihapuheen perusteena kolmessa yli puolessa tapauksista (60 %) ja Instagramissa vain noin yhdessä kymmenestä tapauksesta (12 %). Instagramin vihapuheen perusteena sen sijaan näyttäisi olevan useimmin seksuaalinen suuntautuminen (c), johon kohdistuu 58 % Instagramista tunnistetuista vihapuheviesteistä. Twitterissä seksuaalinen suuntautuminen esiintyy vain 4 %:ssa vihapuheviesteistä. Keskustelupalstoilla esiintyy Twitteriä ja Instagramia reilusti enemmän moniperusteista vihapuhetta (40 % vihapuheviesteistä). Tämä selittynee osaltaan pidemmillä viesteillä.

7 Päätelmiä

Tämän hankkeen yksi keskeisistä tuloksista on tekoälymalli, joka osaa tunnistaa vihapuhetta verkosta. Tekoäly oppi tunnistamaan vihapuhetta, vaikka opetusaineisto oli kohtalaisen pieni. Tekoäly sijoittaa tekstejä lähelle muita semanttiselta merkitykseltään samankaltaisia tekstejä, ja muodostaa vähitellen käsityksen siitä, mikä on vihapuheen alue niin sanotussa semanttisessa avaruudessa. Projektissa syntynyttä tekoälymallia voi sellaisenaan käyttää tunnistamaan vihapuhetta muista viesteistä. Mikäli tekoäly saa jatkossa lisää opetusaineistoa, se oppii yhä varmemmin erottelemaan vihapuheen rajatapaukset raportin määritelmän mukaisesta vihapuheesta.

Yhtenä hankkeen tuloksena voi pitää myös suurta noin 12 miljoonan viestin *sosiaalisen median aineistoa*, josta on koneellisesti tunnistettu vihapuheeksi 11 432 viestiä. Tässä raportissa on tehty tuloksista määrällinen erittely sekä pienempien aineistojen kautta katsaus vihapuheen sisältöihin. Aineisto avaa kuitenkin mahdollisuuden myös syvälliseen laadulliseen analyysiin, jossa voidaan tarkastella valittua osa-aluetta kuten erityisesti tietyllä alustalla julkaistujen vihapuheviestien sisältöjä viestejä tai tietyn vihapuheen aihealueen sisältöjä.

Vihapuhe on epäeksakti termi, jolle voi laatia suppeampia tai laajempia määritelmiä. Tässä hankkeessa vihapuhe määriteltiin laajemmin kuin se esimerkiksi rikoslaissa määritellään. Kirjoitusten merkitseminen vihapuheeksi tai ei-vihapuheeksi vaati usein tulkintaa ja harkintaa. Niin sanotut rajatapaukset sijoittuvat tekoälyn näkökulmasta vihapuheen alueen reunoille semanttisessa avaruudessa, kun taas selvät tapaukset sijoittuvat sen keskelle. Kun tämän hankkeen *sosiaalisen median aineisto* käytiin koneellisesti läpi, tekoälyn poikkeavat päätökset ihmisannotoijaan nähden olivat aina rajatapauksia. Voidaan siis ajatella, että tekoälyn avulla on mahdollista luotettavasti seuloa verkkoaineistoista tämän hankkeen määritelmän mukaisesti selkeimmin vihapuheeksi luokiteltavia kirjoituksia. Koneellisesti tunnistetun vihapuheen teemoja tarkastellessa on kuitenkin hyvä tiedostaa, että osa aineistosta on oletusarvoisesti rajatapauksia.

Arvioomme perustuen Ylilauta näyttäisi toimivan areenana suurelle osalle suomenkielistä verkossa julkaistavaa vihapuhetta. Määrällisessä analyysissä se arvioitiin erityisen suurivolyymiseksi alustaksi 96 %:n osuudella kaikesta tässä hankkeessa tunnistetusta verkkovihapuheesta. Ylilauta on alakulttuurinen kuvalautafoorumi, jolla vallitsee vapaan keskustelun ja nimettömyyden ihanne. Se on tullut tunnetuksi poliittista korrektiutta vastustavasta huumorista, pilailusta ja trollaamisesta eikä keskustelua juuri kontrolloida muuten kuin selvien laittomuuksien osalta. (Vainikka 2019). Tämän hankkeen vihapuheen määritelmä on rikoslain määritelmää laajempi. Ylilaudalla omaksuttu keskustelukulttuuri voi olla osaltaan syynä siihen, miksi niin suuri osa sillä julkaistuista viesteistä on tunnistunut

hankkeessamme vihapuheeksi. Tulostemme valossa voisi kuitenkin pohtia, onko Ylilaudalla omaksuttu puhetapa sellainen, joka näyttäytyy laajemmalle sosiaalisen median käyttäjäkunnalle vihamielisenä ja loukkaavana puheena ja onko alustasta tullut vihapuheelle tuen antava yhteisö? Vihapuheen vastaisessa työssä voisikin olla hyödyllistä kohdentaa toimenpiteitä ja vuoropuhelua myös suoraan tämän kaltaisille alustoille sen lisäksi, että puututaan kansainvälisten digijättien toimintaan.

Suurista sosiaalisen median alustoista puolestaan Twitter oli vihapuheen kannalta merkittävin. Twitter oli julkaisualustana 2.5 %:ssa vihapuheeksi tunnistetuista viesteistä. Mielenkiintoinen tulos oli uudelleentviittausten suuri määrä vihapuheeksi tunnistettujen kirjoitusten joukossa. Tämä kertoo siitä, että vihapuheen tuottajien määrä voi olla pieni, mutta tehokkaalla uudelleenlevityksellä vihapuheilmiö voi näyttää kokoaan suuremmalta. On kuitenkin muistettava, että nämä suhteet odotusarvoisesti muuttuisivat, jos mukana aineistossa olisivat myös Facebookin yksityiset tilit ja ryhmät. Uudelleentviittausten määrä oli aineistossa suurin sellaisina päivinä, kun vihapuhetta oli muutenkin paljon. Vihapuheen piikkien sisällön tarkastelu paljasti, että uudelleentviittauksilla oli niissä merkittävä rooli. Piikki saattoi selittyä pitkälti muutamalla yksittäisellä, runsaasti uudelleentviitatulla viestillä. Tavallisten sosiaalisen median käyttäjien rooli korostuukin vihapuheen levittäjinä ja heidän käyttäytymisellään eli viestien uudelleen lähettämisellä tai lähettämättä jättämisellä on suuri merkitys.

PROJEKTIRYHMÄ

Tämän raportin tuottamiseksi projektiryhmään osallistuivat seuraavat:

Vihapuheen määritelmä, annotointi, raportti: Laura Kettunen, FM
Raportin kommentointi: Reeta Pöyhtäri, YTT
Utopia Analytics:
- Raportti: Mari-Sanna Paukkeri, TKT
- Datan mallinnus: Jaakko Väyrynen, TKT
- Projektinhallinta: Kari Kemppi, DI

Oikeusministeriö

LÄHTEET

Euroopan neuvoston ministerikomitean suositus vihapuheesta 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/. Luettu 10.9.2019.

Knuutila, Aleksi; Kosonen, Heidi; Saresma, Tuija; Haara, Paula & Pöyhtäri, Reeta (2019). Viha vallassa: Vihapuheen vaikutukset yhteiskunnalliseen päätöksentekoon. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja.

Korhonen, N.; Jauhola, L.; Oosi, O. & Huttunen, H. P. (2016). Usein joutuu miettimään, miten pitäisi olla ja minne olla menemättä. Selvitys vihapuheesta ja häirinnästä ja niiden vaikutuksista eri väestöryhmiin. Oikeusministeriön julkaisu, 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 s. 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

Lakimiesliitto (2019). Lakimiesliitto vaatii järeitä toimia oikeudenhoidon ammattilaisten suojaamiseksi vainoamiselta. Suomen Lakimiesliitto, 21.05.2019. https://www.lakimiesliitto.fi/uutiset/lakimiesliitto-vaatii-jareita-toimia-oikeudenhoidon-ammattilaistensuojaamiseksi-vainoamiselta/

Mäkinen ym. (2019). Sanat ovat tekoja: Vihapuhe ja nettikiusaamisen vastaisten toimien tehostaminen. Sisäministeriön julkaisu.

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

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

Rauta, J. (2018). Poliisin tietoon tullut viharikollisuus Suomessa 2017. Poliisiammattikorkeakoulun raportteja 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*, toim. E. Palonen & T. Saresma, 181–198. Tampere: Vastapaino.

Tiedonjulkaisemisen neuvottelukunta (2015). Kysely tutkijoiden asiantuntijaroolissa saamasta palautteesta: Tulosyhteenveto. Tiedonjulkaisemisen neuvottelukunta, 22.12.2015.

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

Valtakunnansyyttäjänviraston työryhmän raportti (2012). Rangaistavan vihapuheen levittäminen Internetissä. https://www.valtakunnansyyttajanvirasto.fi/material/attachments/valtakunnansyyttajanvirasto/vksvliitetiedostot/tyoryhmat/6Jqa1QEsJ/17-34-11_tyoryhmaraportti.pdf

Oikeusministeriö
PL 25
00023 Valtioneuvosto
www.oikeusministerio.fi

Justitieministeriet
PB 25
00023 Statsrådet
www.justitieministeriet.fi

ISSN 2490-0990 (PDF)
ISBN 978-952-259-893-6 (PDF)

OIKEUSMINISTERIÖ
JUSTITIEMINISTERIET