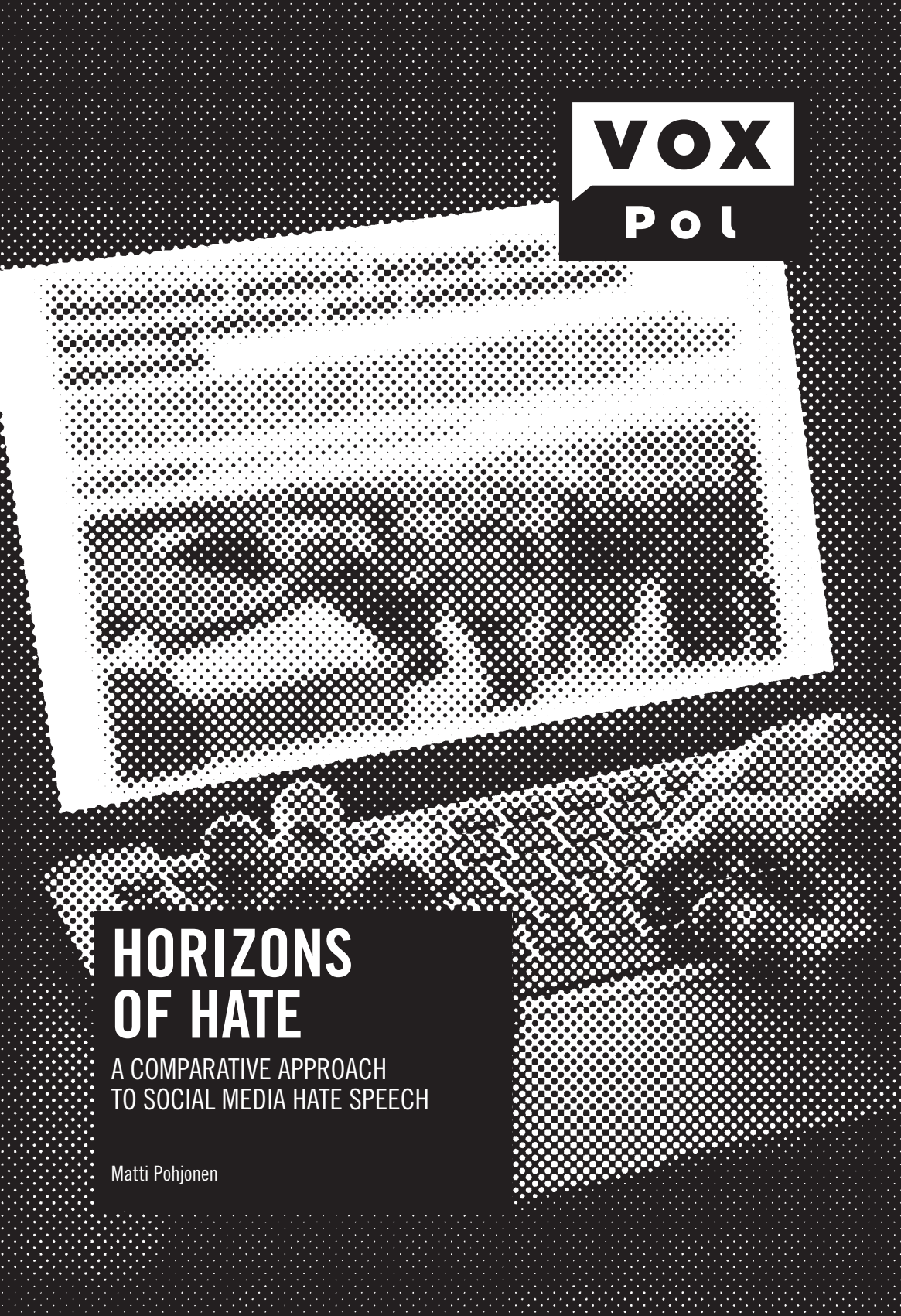


HORIZONS OF HATE

A COMPARATIVE APPROACH TO SOCIAL MEDIA HATE SPEECH

Matti Pohjonen

HORIZONS OF HATE

A COMPARATIVE APPROACH TO SOCIAL MEDIA HATE SPEECH

Matti Pohjonen

About the Author

Matti Pohjonen is a Lecturer in Global Digital Media at the School of Oriental and African Studies (SOAS), University of London. For the past ten years, he has developed critical-comparative research approaches for understanding digital cultures globally. This has included work on international news and blogging in India, mobile technology in East Africa, comparative research on online extremism and hate speech in Ethiopia and Europe, and exploring new methods in "big data" analysis for digital media research. He received his PhD from SOAS, where he also worked as a Senior Teaching Fellow and an AHRC-funded Post-Doctorate Research Fellow. He has also worked as a Researcher for the Programme for Comparative Media Law and Policy (PCMLP) at the University of Oxford, a Research Fellow for the VOX-Pol Network of Excellence, a Visiting Research Fellow at the Centre for Media and Communication (ZeMKI) at the University of Bremen, and as a postdoctoral researcher for the University of Helsinki.

Acknowledgements

The author would like to thank the VOX-Pol Network of Excellence for funding this research.

Designed and typeset by Soapbox, www.soapbox.co.uk

TABLE OF CONTENTS

1. INTRODUCTION

My vox-pol fellowship in 2015–2016 coincided with two important developments in Europe. The first was the eruption of social media hate speech that followed hundreds of thousands of refugees arriving from war-ravaged Syria, Iraq, and Afghanistan. The second was the growing buoyancy of the extreme right online, who tried to capitalise on this anger to increase its political power and recruit followers.

I had just finished a research project in Ethiopia, a country with a long history of civil war and conflict. It was a sobering experience to return to Europe to discover how the social media debates on the refugee crisis had become more aggressive and vitriolic than anything I had experienced before as a comparative digital media researcher. A cursory look, for instance, at Facebook pages in Finland (a country that is more commonly known for its peaceful politics, consensus, and social stability) would reveal thousands of comments using the most graphic and violent language possible, such as "Those rats should be exterminated from the world," and "Why don't we shoot the invaders into a hole and burn them with gasoline to warm our feet?"

All the hallmarks of ostensibly the worst kind of 'hate speech' were present: attacking people based on their group identity; dehumanising them by comparing them to animals; and incitement to violence. Moreover, such comments were posted by individuals using their public profiles, visible for anybody to see.

So if social media conversations in what has been called the safest country in the world had become more violent than those I had observed in a country with an ongoing violent ethnic and political conflict, what was going on in these popular social media forums? Research into the socio-psychological dynamics of violent conflict has shown that an increasingly aggressive and polarised style of communication can be one of the telltale signs of escalating conflict

(Hamelink 2011; Buyse 2014). Were these hateful comments possibly a symptom of some underlying social and political tension simmering under the glittering surface of social media screens waiting to erupt into real-world violence? How dangerous were they?

As my research progressed, it also became increasingly clear that this growing visibility of social media hate speech was also somehow related to the resurgent confidence of the extreme right online. A new style of online political tactics had emerged, the significance of which researchers and policymakers were struggling to understand: ecosystems of fake news; bots manipulating social media popularity rankings; and disinformation campaigns orchestrated on social media forums (Benkler et al. 2017; Marwick and Lewis 2017; Wardle and Derakhshan 2017). What was the relationship between these activities by the extreme right online and the emergence of social media vitriol that targeted refugees and the people who supported them? How successful were these groups in exploiting the affordances of social media platforms such as Facebook to advance their political goals?

A few years later, these questions remain as crucial as ever. Social media debates in Europe and the United States are as toxic as ever. The concerns about the political fallout of extreme right disinformation have become mainstream. Signs of these developments were visible in my research. This report outlines its findings.

1.1 EXISTING LITERATURE

Research on violent online political extremism has conventionally focused on the online activities of violent extremist and terrorist groups (Meleagrou-Hitchens and Kaderbhai 2017). This has presupposed a relatively easy-to-define normative division between legitimate forms of political expression and illegitimate forms of political expression such as incitement to terrorist violence. This normative division, however, becomes difficult to ascertain when the question is of online hate speech. Brown and Cowls (2015, p. 29) write, "Beyond the categories of speech already described, which many states have proscribed by law,

there is less consensus on what constitutes online 'extremist' material that should be policed – especially where it does not directly encourage violence."

Gagliardone et al. (2015a, p. 10) define online hate speech as "expressions that advocate incitement to harm (particularly, discrimination, hostility and violence) based upon the target's being identified with a certain social or demographic group." This definition is also often expanded to include expressions that more generally "foster a climate of prejudice and intolerance on the assumption that this may fuel targeted discrimination, hostility and violent attacks" (*ibid*). The vast literature on online hate speech thus broadly agrees on two characteristics: hate speech dehumanises its victims according to their group identity, but it also amplifies the group identity of the perpetrator by attempting to create an antagonistic relationship between 'us' and 'them' (see Gelber 2011; Heinze 2017; see also Butler 1997). Waldron (2012, p. 4), for instance, writes that hate speech creates "something like an environmental threat to social peace, a sort of slow-acting poison, accumulating here and there, word by word, so that eventually it becomes harder and less natural for even the good-hearted members of the society to play their part in maintaining this public good."

Unlike more clear-cut cases of violent extremist activity, however, such indirect effects of hate speech are difficult to pin down analytically or prosecute legally. Outside clear-cut examples of incitements to violence, there is thus no consensus on exactly what kinds of speech acts should fall outside the purview of legitimate forms of political expression. Bartlett et al. (2014, p. 11) write "how to define the limits of free speech is a central debate in most modern democracies. This is particularly true in respect of speech that might be deemed hateful, abusive, or racist. Defining and legislating against this type of speech is extremely difficult, and has spawned a large philosophical, linguistic, theoretical, and legal literature."

Different historical traditions also inform where this normative line between legitimate and illegitimate forms of political expression are drawn. The United States and Europe, for instance, entertain different notions about where this boundary of acceptable forms

of expression lies, with the United States circumscribing stricter protection for freedom of speech (see Post 2009; Brown 2015). Moreover, once we move away from the relatively sheltered purview of Western liberal democracies, debates on hate speech have also been widely misused for political purposes (see Price and Stremlau 2017). Gagliardone et al. (2015a, p. 10) write, "accusations of fomenting hate speech may be traded among political opponents or used by those in power to curb dissent and criticism." The freedom of speech organisation Article 19 (2015, p. 16) similarly cautions against "too readily identifying expressions as 'hate speech' … as its use can also have negative consequences … and can be abused to justify inappropriate restrictions on the right to freedom of expression, in particular in cases of marginalised and vulnerable communities."

Given these controversies around defining what hate speech is – and especially what should be done about it – this report instead uses the term 'aggressive or hateful speech' when referring to instances of online vitriol, aggression, and hate that are broadly targeted at refugees and immigrants. Where possible, I avoid the more commonly used term 'hate speech'. I do this for two reasons. Firstly, using this term allows me to approach the communicative dynamics of social media conversations without having to first ascribe normative value to them. As I have argued elsewhere, global debates on hate speech have become overdetermined insofar as there are more theories in circulation than empirical evidence would perhaps warrant (see Pohjonen and Ahmed 2016). Recoiling from these legal–normative debates around hate speech, even if temporarily, helps to step back from the controversies and focus more on "the situatedness of online speech forms in different cultural and political milieus" (Pohjonen and Udupa 2017, p. 1174) – the complex communicative relationships and media-related practices that inform such speech acts online. Secondly, using a broader term such as aggressive or hateful speech also allows me to explore large-scale social media conversations without being bogged down with the methodological problems involved in accurately classifying what hate speech is, at least in the stricter legal sense of the term (Davidson et al. 2017; see also Saleem et al. 2017). Davison et al. (2017, p. 4) warn

that "if we conflate hate speech and offensive language then we erroneously consider many people to be hate speakers … and fail [to] differentiate between commonplace offensive language and serious hate speech … Given the legal and moral implications of hate speech it is important that we are able to accurately distinguish between the two." Employing a category whose boundaries are less rigorously defined thus allows the research to shift focus away from trying to define what hate speech is and focus instead on the broader communicative dynamics and communicative relationships behind it. Hopefully, this subtle, yet important, difference becomes clear in the report.

1.2 A COMPARATIVE APPROACH TO HATE SPEECH

Underpinning the research is a tension between two academic traditions. The first is the growing body of research on violent online political extremism. The second is approaches in peace and conflict studies that have analysed the role the media has played in situations of violent ethnic, political, or social conflict. Jackson (2012) argues that these two traditions have developed, surprisingly, in isolation from each other despite sharing similar concerns about the role the media have in catalysing offline violence (see also Conway 2017). This disconnect between them is all the more surprising because some of the most striking examples where the media have been linked to widespread violence do not come from the terrorist attacks in Brussels, London, Nice, or Paris. These events, however horrific, are still dwarfed in comparison to the use of community radio in Rwanda to incite genocide (where more than a million people were killed), the use of Twitter in Kenya during the 2007 presidential elections to stir up ethnic hatred (where more than a thousand people were killed), or the recent use of Facebook to incite violence against the Rohingyas in Myanmar (where thousands have allegedly been killed).

The framework used in this research builds on two previous research projects that explored online hate speech and conflict from such a comparative perspective. The first was a pilot project

that mapped out, for the first time ever, online hate speech in Ethiopia (see Gagliardone, Pohjonen, and Patel 2014). The pilot explored the 'dangerous speech' framework, developed initially by Benesch (2012, 2014) and applied by the Umati project (2012-2013) to explore instances of online hatred during the 2014 presidential elections in Kenya. Benesch argues that hate speech is too broad as a conceptual category for identifying those kinds of speech acts that could act as early warning signs for offline violence. She writes, "First, hate speech is common in many societies, unfortunately, including those at minimal risk of genocide. Second, some hate speech does not appreciably increase the risk of mass violence, although it may cause serious emotional and psychological damage" (2012, p.1). Moreover, she proposes that five additional criteria are needed when assessing how dangerous speech acts are (2012, p. 2):

1. A powerful speaker with a high degree of influence over the audience;
2. The audience has grievances and fear that the speaker can cultivate;
3. A speech act that is clearly understood as a call to violence;
4. A social or historical context that is propitious for violence, for any of a variety of reasons, including longstanding competition between groups for resources, lack of efforts to solve grievances, or previous episodes of violence; and
5. A means of dissemination that is influential in itself, for example because it is the sole or primary source of news for the relevant audience.

Our pilot research found that there were indeed abundant examples of such aggressive or hateful speech in Ethiopian online spaces. However, the pilot project also quickly realised that focusing methodologically on the formal content of social media speech acts alone was not enough to assess its risks: this was neither representative of the dynamics of social media conversations nor indicative

of how dangerous such statements were in catalysing offline violence. As Benesch (2013, 2014), and Leader Maynard and Benesch (2016) stress, a formal analysis of extreme content needs to be augmented with a more contextual understanding of the speakers and audiences, socio-historical context, and media channels used to disseminate it. Conversely, in polarised political situations, focusing only on controversial content without first contextualising it can further risk exacerbating political tensions and give justification to governments to implement measures of censorship that might not be commensurable with its actual dangers.

Our follow-up project in Ethiopia tried to address this concern. Instead of focusing only on the formal content of speech acts, it decided to map out the broader communicative dynamics behind such hateful speech online. To do this, it developed a conceptual framework that categorised online discussions based on whether they facilitated or hindered a communicative relationship between interlocutors involved in online or social media conversations. Conversations classified as 'going against' consisted of statements that represented conflict-maintaining behaviour or advocated hatred, incitement or discrimination based on ethnicity, religion, gender, sexual identity or political affiliation. In turn, the types of conversation that we classified as 'going towards' consisted of statements that helped to maintain a communicative relationship by acknowledging the other person's or group's position and by creating engagement with other members in the conversation even if the tone was critical (see Gagliardone et al. 2016).

This conceptual move helped us to provide a different perspective to the underlying communicative dynamics and relationships behind hateful social media conversations. What was surprising about approaching the object of study from this perspective was that social media conversations in Ethiopia seemed to favour positive engagement over more aggressive or hateful forms of communication. In other words, by contextualising these conversations into the broader media environment and communicative relationships in which they were embedded, these conversations did not reflect

the ethnic and political polarisation in the country as much as much as we had anticipated. Instead, they showed promise as a means to mitigate some of the existing tensions by creating spaces of engagement where ideologically opposed participants could communicate and engage with each other (see Gagliardone et al. 2015b, 2016).

These findings are idiosyncratic to the Ethiopian media environment (see Human Rights Watch 2015, 2016) and they cannot, of course, be applied to other countries without first taking into account the different socio-historical contexts and media environments in these countries. Nonetheless, our research in Ethiopia helped us to extrapolate three suggestions for analysis of online and social media hate speech from such a comparative perspective:

1. Hate speech should not be seen as a universal category but rather as a situated practice that always exists in specific cultural and political contexts and media environments;

2. Focusing only on the content of hate speech acts risks sensationalising online and social media conversations in polarised political environments and situations of conflict; and

3. The risks of hate speech cannot be thus understood by focusing only on the content of hateful speech acts. The broader communicative dynamics and relationships behind hateful speech acts also need to be carefully considered.

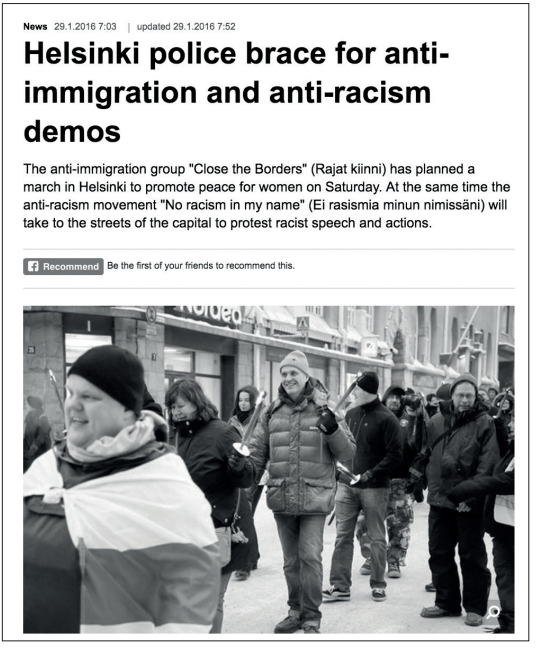
My VOX-Pol research emerged out of an effort to further develop this kind of critical-comparative research for understanding social media hate speech. On a more conceptual level, I wanted to explore what types of insight such a comparative perspective would engender when used in the European context. Could some of the approaches that were developed to understand media and conflict in countries such as Ethiopia, Kenya, and Rwanda also help us to better understand what was going on on European social media during the refugee crisis? What kinds of methods of analysis would help us to research these questions empirically?

2. RESEARCH DESIGN

2.1 CASE STUDY

The so-called refugee crisis refers to a period that began in early 2015 when hundreds of thousands of refugees started arriving in Europe. Around 30,000 arrived in Finland – a nearly ten-fold increase from previous years. This arrival of thousands of people to this relatively homogeneous country led to a heated debate about how they should be received. Rumours about crimes and especially rapes committed by refugees were rife on social media. Soldiers of Odin, an extreme right vigilante group, gained notoriety by patrolling the streets in its self-proclaimed mission to protect vulnerable women from the refugees.[1]

Finnish Broadcasting Company (*Yle*) article on anti-refugee/immigration protest and its counter-protest[2]



News 29.1.2016 7:03 | updated 29.1.2016 7:52

Helsinki police brace for anti-immigration and anti-racism demos

The anti-immigration group "Close the Borders" (Rajat kiinni) has planned a march in Helsinki to promote peace for women on Saturday. At the same time the anti-racism movement "No racism in my name" (Ei rasismia minun nimissäni) will take to the streets of the capital to protest racist speech and actions.

Recommend Be the first of your friends to recommend this.

1 See independent.co.uk/news/world/europe/soldiers-of-odin-the-far-right-groups-in-finland-protecting-women-from-asylum-seekers-a6846341.html.

2 See https://yle.fi/uutiset/osasto/news/helsinki_police_brace_for_anti-immigration_and_anti-racism_demos/8632946.

This polarisation was also reflected in social media. The *Rajat Kiinni* (Close the Borders) Facebook group became a favourite hotbed for anti-refugee/anti-immigration sentiment. In turn, the Facebook group the *Rasmus* (Finland's national network and association working against racism and xenophobia, and promoting equity and human rights) became a popular forum for anti-racist opinions. The two groups routinely engaged each other in online flame wars as well as demonstrations and counter-demonstrations against each other.

The social polarisation that heightened during the refugee crisis was of course not unique to Finland. However, what was perhaps unique about Finland's social media environment was that, when the toxicity of conversations was at its apex at the beginning of 2016, members of these two antagonistic Facebook groups also launched a new group with the explicit purpose of creating a 'civilised conversation' about refugees/immigration. This third group, *Asiallista Keskustelua Maahanmuutosta* (a civilised conversation about immigration), established a strict set of guidelines on what type of speech was tolerated in order to facilitate a less toxic online culture (see Section 3.4).

These three social media communities in Finland thus provided my research with a unique case study to empirically analyse the communicative dynamics and relationships behind hateful online and social media conversations during the refugee crisis. The three groups included: a popular anti-refugee/anti-immigration group criticised for hate speech and its links to the extreme right; an anti-racist group that opposed it; and a group launched in-between as a kind of 'organic' counter-speech aimed at mitigating the toxicity of social media debates (see Bartlett and Krasodomski-Jones 2015; Ferguson 2016). The groups were also public and highly active. In 2016 alone, these groups published close to 100,000 posts and 500,000 comments. The large-scale nature of these public conversations allowed the research to explore methods usually reserved for 'big data' approaches. This allowed me to examine, in particular, three sets of research questions:

- **RQ1**: How prevalent was such aggressive or hateful communication in these three distinctly different types of Facebook

group? Where, and under what circumstances, was it prevalent? What kinds of communicative dynamics and relationships informed them?

- **RQ2**: How did the external information shared in these groups relate to the prevalence of aggressive or hateful communication? What was the relationship between extreme right news sources and disinformation online and the prevalence of such a style of communication?
- **RQ3**: Was there something distinct about the communicative dynamics of the Facebook group that was set up to mitigate toxic social media conversations around refugees/immigration? Was this group successful in fostering engagement and debate?

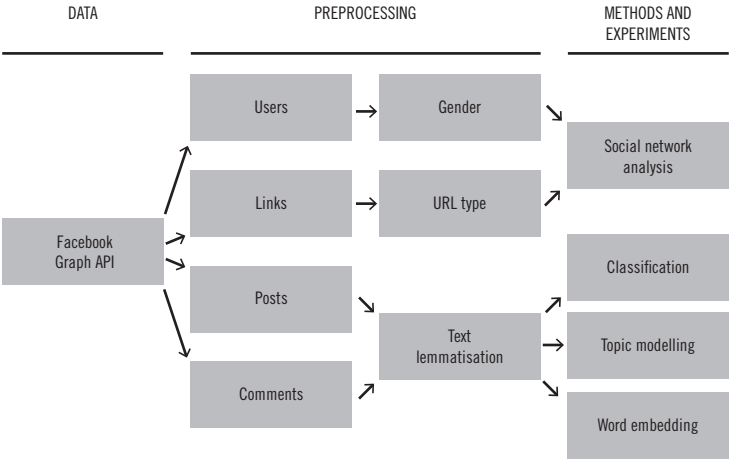
2.2 DATA COLLECTION AND PREPROCESSING

In order to explore these questions, all public data from the three groups was downloaded using the Facebook application programming interface (API).[3] The dataset was then enriched to add gender, the type of information source shared (root URL), and whether the posts and comments contained hateful speech.[4] Figure 1 shows the workflow used for data retrieval, preprocessing and exploration/analysis.

3 I used the Rfacebook package for this purpose. See https://github.com/pablobarbera/Rfacebook.

4 Firstly, the gender of all the users posting and commenting was identified by cross-referencing the names of the users with a list of all male and female names in Finland. This allowed the gender of the speaker to be identified 90–95% of the time. Secondly, the root domain of all the URLs was parsed and then manually labelled. Six categories were used to differentiate the types of news sources shared: (1) news (mainstream); (2) alternative (extreme right); (3) tabloids; (4) entertainment; (5) blog; and (6) social media. The urltools package in R was used to parse the URLs. See https://github.com/Ironholds/urltools. Finally, all the textual content from the posts and comments were morphologically stemmed and lemmatised to facilitate text mining. We used OMORFI, the open-source morphology package for the Finnish language. See https://github.com/flammie/omorfi.

Figure 1. Data collection, preprocessing, and analysis workflow



2.3 METHODS AND EXPERIMENTS

2.3.1 Classification of Aggressive, Offensive and Hateful Speech

One of the biggest challenges in researching social media hate speech is accurately classifying statements that contain aggressive, vitriolic, offensive, or hateful content. The difficulty is both conceptual and methodological. First of all, it is difficult to differentiate what constitutes hate speech conceptually and to determine when it differs from offensive language (Davidson et al. 2017, p. 1). Even with human annotators, it takes a lot of effort to reach coder agreement, especially when working with the stricter legal definition of the term (see Ross et al. 2016). These challenges are compounded when the research uses computational methods on large datasets. The prevalence of offensive keywords can cause algorithms to misclassify statements as hate speech when they should be seen merely as instances of aggressive or hateful communication. Not all hateful speech contains easily identifiable linguistic markers or features that could help to identify it. Instead, the everyday use of language continually changes and is made up of nuanced linguistic forms such as jokes, innuendo,

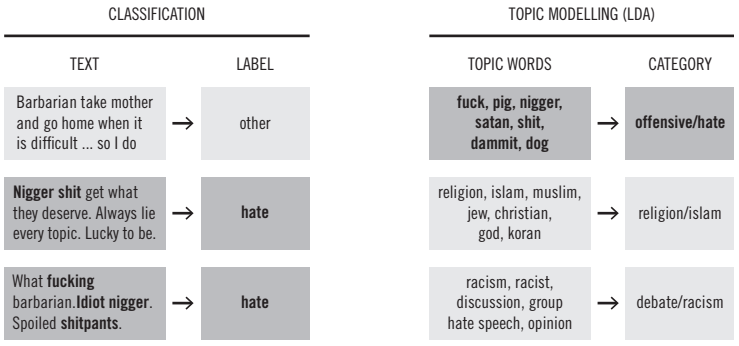
irony, metaphors, and double meanings that obstinately challenge capture by computational methods (Kwok and Wang 2013; see also Burnap and Williams 2015).

Aware of the methodological challenges involved in identifying speech that could be labelled 'hate speech' for analysis, the research focused instead on a category of statements that were more broadly indexical of sentiments of aggression or hate found in Facebook posts and comments. Two machine learning approaches were experimented with to detect such statements. The first explored supervised machine learning classification where a subset of data was labelled to aid the classification of such statements. To do this, I manually labelled 3,000 comments that were randomly selected. From these statements, I identified the types of posts and comments that I was interested in observing in the research. This was done, in particular, by focusing on 'loaded' words, curse words, and expletives that expressed aggression or hate. I then divided the labelled statements into a training and test set (60/40) and used ensemble learning of eight different machine learning classifiers to explore different approaches. The most accurate results were achieved with the classification and regression tree algorithm (with an F-score of 0.87).[5] The second approach experimented with unsupervised machine learning approaches called Latent Dirichlet Allocation (LDA) where topic clusters that contain words commonly associated with such

5 I used the the RTextTools package in R for this purpose. See https://journal.r-project.org/archive/2013-1/collingwood-jurka-boydstun-etal.pdf.

hateful speech were identified.[6] Figure 2 shows the triangulation between these two methods used.

Figure 2. Triangulation of supervised and unsupervised machine learning methods for detecting hateful speech[7]



However, given the exploratory and more qualitative nature of the research (and the lack of resources), I did not carry out intercoder reliability tests with multiple coders on the dataset, but instead checked the results through extensive random checks to

6 For the unsupervised topic modelling, I used the text2vec package in R. I first experimented with a variety of hyperparameters, and number of topics. I ended up using ten topics (document topic prior = 0.5; topic_word_prior = 0.01). I then labelled these topics for the posts and comments based on what I interpreted to be the most relevant topic based on the top 50 words. I finally classified all the posts and comments based on what was given as most probable topic for each post and comment. The topics that were identified in the comments through the method were: (1) cost; (2) debate/racism/speech; (3) English/other; (4) migration/culture; (5) government/politics; (6) news/media; (7) offensive/hate; (8) police/terrorism/border; (9) religion/Islam; and (10) work/welfare. With the exception of offensive/hate, the topics were used for data exploration purposes only.

7 The labelling was done on the original content of the posts and comments. Translations in Figure 2, however, are based on the lemmatised versions of these posts and comments to illustrate the similarities between the different types of words used in the supervised and unsupervised machine learning approaches.

see how relevant the labels were for the analysis in question. Both the supervised classification and unsupervised topic modelling approaches were used thus in a narrower methodological sense to augment the primary qualitative research approach. The supervised and unsupervised machine learning methods should be considered more of an *exploratory research heuristic* that was used to provide insights into the communicative dynamics of these groups together with the *primary digital ethnographic method used by the research.*

What types of comment were then detected by this triangulation of approaches? The statements classified as aggressive or hateful included a wide range of posts and comments explicitly targeted at immigrants/refugees and people who supported them. It is also important to note that due to the comments that were identified as aggressive or hateful, because of a method that emphasised the use of curse words and offensive keywords, a broader range of statements that were aggressive or hateful but that did not target anyone in particular was also detected. This more general category of statements, however, was necessary for the analysis as the focus was to gain a broader understanding of the communicative dynamics behind such conversation during the refugee crisis, and thus not limit itself only to speech acts that fit under stricter definitions of hate speech.

2.3.2 Word Associations

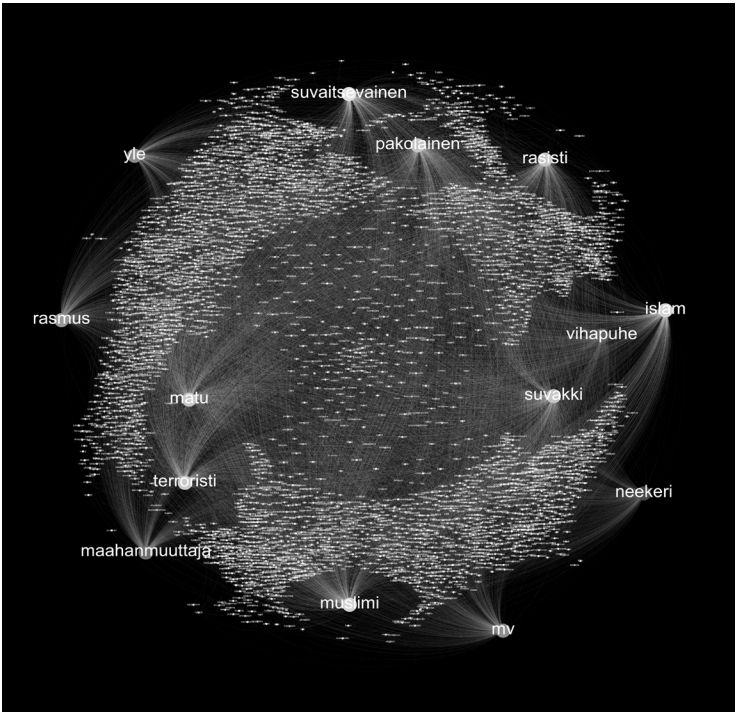
The research also compared word associations of key terms relating to debates around the refugees. This was done by experimenting with a set of deep learning methods called word embedding, which map a vocabulary of words onto a vector of numbers to create representations of the words based on the context in which they occur in the text, and the proximity to other words. Such word embeddings are especially interesting for exploratory research purposes because they have been shown to detect implicit biases in the use of language from large textual datasets. Foulds (2018, p. 2) writes that "word embeddings can encode implicit sexist assumptions," such as the analogy 'man is to computer programmer as woman is to homemaker' (see also Bolukbasi et al. 2016). To explore such associations found within the three different types of group – and especially differences/

biases in the language used – the research thus explored words that were closely associated with terms related to the refugee crisis. The following 'seed' terms were selected for the association analysis:

- Islam – Islam;
- Muslim – Muslim;
- *Maahanmuuttaja* – immigrant;
- *Matu* – a derogatory term for immigrant;
- *Neekeri* – nigger;
- *Pakolainen* – refugee;
- *Rasisti* – racist;
- *Suvaitsevainen* – somebody who is tolerant or liberal;
- *Suvakki* – a derogatory term for somebody who is tolerant or liberal;
- *Terroristi* – terrorist;
- *Vihapuhe* – hate speech;
- *MV* – reference to the popular extreme right online news site *MV-lehti*;
- *Yle* – reference to the mainstream public news channel; and
- *Rasmus* – reference to the anti-racist group.

Figure 3 shows a network representation of some of the exploratory word association network mappings carried out to understand relationships between concepts in the three different ideologically positioned Facebook groups.

Figure 3. A network representation of word embedding based on the cosine similarity of key words related to the refugee crisis



2.3.3 Social Network Analysis

The research also used social network analysis to identify what kinds of social networks and communities were behind these three groups, and how these changed over time. This was done by modelling the conversations into two different types of network. The first network consisted of the relationship between the external news source (URL), and the people who posted and shared this URL. The second network comprised the relationship between people who posted and people

who commented on these posts. The open-source software Gephi was used for network visualisation and exploration. The software ORA was used, where necessary, for statistical analysis and dynamic social network analysis.

2.3.4 Caveats

There were also a number of caveats about the data collection process and methods. Firstly, while these three groups were highly visible in the social media debates during the refugee crisis, they should not be considered a representative sample of the population or social media conversations in general. Rather, these groups provided a non-probability sample that was purposely selected based on the identity of these groups, and their suitability for the research questions. Secondly, the collection of data relied on the Facebook API. This is contingent on the privacy settings of Facebook. While in most cases this does not pose a problem as the groups in question were public, it is nonetheless impossible to verify what proportion of the original conversations were included in the final dataset. Posts and comments are, for instance, sometimes erased after being published; these are not available in the final dataset.

2.4 RESEARCH WORKFLOW

The research used a mixed-method approach combining digital ethnography with data exploration (see Pohjonen 2018; see also Laaksonen et al. 2017). The workflow for analysis consisted of four steps:

1. Longitudinal observation of the three groups over a period of nine months was used to identify what the key themes and topics of interest were, and how these changed over time;

2. Based on this ethnographic engagement, computational text mining was used to identify posts and comments in which aggressive or hateful speech was found, as well as the associations between selected keywords;

3. Social networks analysis was used to identify communities in these groups, how they related to each other, and how they changed over time. This was further extended by overlaying attribute data over the network to visually explore how different types of conversation (such as whether it contained aggressive or hateful speech) or the gender of the speaker related to the social networks and communities; and

4. This data exploration was repeated iteratively until empirically grounded arguments could be formed. The approach of combining qualitative and quantitative insights thus provided both a granular perspective to the conversations as well as helping to identify patterns and trends at a scale usually unavailable for qualitative methods.

3. RESULTS

3.1 DESCRIPTIVE FINDINGS

The final dataset consisted of public data from the three Facebook groups which was published between 1 January and 1 September 2016. The three groups were most active during the first half of 2016 when the debate over refugees was at its most heated in Finland. Table 1 shows that the majority of the members in the *Rajat Kiinni* group and the *Asiallista Keskustelua* group were men. The division between men and women was more evenly divided in the *Rasmus* group: a small majority of members posting were men, and a slight majority commenting were female. The *Asiallista Keskustelua* group had the highest average number of comments per post, in accordance with it being a group that was set up as a space for conversation. This was also reflected in it having the longest average word count. All the groups shared external news sources, with about half the posts containing links to external URLs.

Table 1. Overall description of the dataset

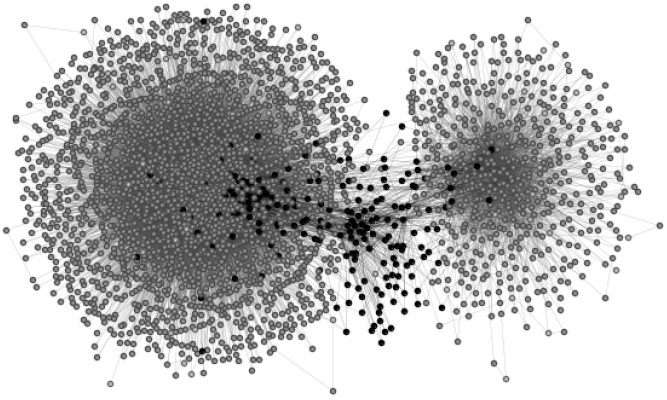
	RAJAT KIINNI	***ASIALLISTA KESKUSTELUA***	***RASMUS***
Members[8]	12,443	1,259	13,787
Number of posts	54,474	1,101	8,308
Number of comments	355,293	16,245	76,010
Unique individuals	Posts: 3,416 Comments: 7,317	Posts: 262 Comments: 642	Posts: 2,263 Comments: 6,106
Gender – posts	Male: 63% Female: 37%	Male: 62% Female: 38%	Male: 54% Female: 46%

8 The number of members are from when the data was downloaded in September 2016, and may have changed significantly since. The *Rajat Kiinni* group has since been shut down by Facebook.

	RAJAT KIINNI	*ASIALLISTA KESKUSTELUA*	*RASMUS*
Gender – comments	Male: 66% Female: 34%	Male: 54% Female: 46%	Male: 46% Female: 54%
Unique URLs shared	2,700	183	2,273
Percentage of posts containing a URL	50%	53%	59%
Mean likes	26.8	5.5	31.6
Mean comments	8.3	15.1	8.4
Mean shares	1.8	0	0.4
Mean number of words per post	15.2	34.3	20.3
Mean number of words per comment	13.1	29.1	24.8

A network visualisation of the dataset in Figure 4 also illustrates how polarised the conversations were: the anti-refugee/anti-immigration (large cluster on the left) and anti-racist groups (cluster on the right) contained only a few individuals who participated in both groups. These separate clusters were bridged by the new discussion group (people who participated in the discussion group coloured in black).

Figure 4. The relationship between people who posted and who commented in the three groups during the month of February 2016 (with isolates and pendants removed)



3.2 THE COMMUNICATIVE DYNAMICS OF HATEFUL SPEECH

How prevalent were such aggressive or hateful conversations in these three ideologically different groups? As Table 2 shows, the *Rajat Kiinni* group contained most of the comments identified as aggressive or hateful speech. The supervised classification method detected 34,501 aggressive or hateful comments (10%). The LDA topic modelling method, in turn, identified 75,840 (21%) aggressive or hateful comments.[9]

9 The discrepancies between the two methods experimented with in the research can be explained by the use of offensive 'feature words' (known offensive and derogatory terms, and curse words) in the supervised classification process. The statements classified as hateful by the LDA method, in turn, identified a broader range of statements that were less tied to these specific offensive keywords, derogatory terms and curse words. Random checks were performed on all the datasets to doublecheck the utility of these classifications to augment the digital ethnographic methods used. In most of the quantitative analysis, the more restrictive machine learning classification method was used, and the topic modelling method was used to verify the results.

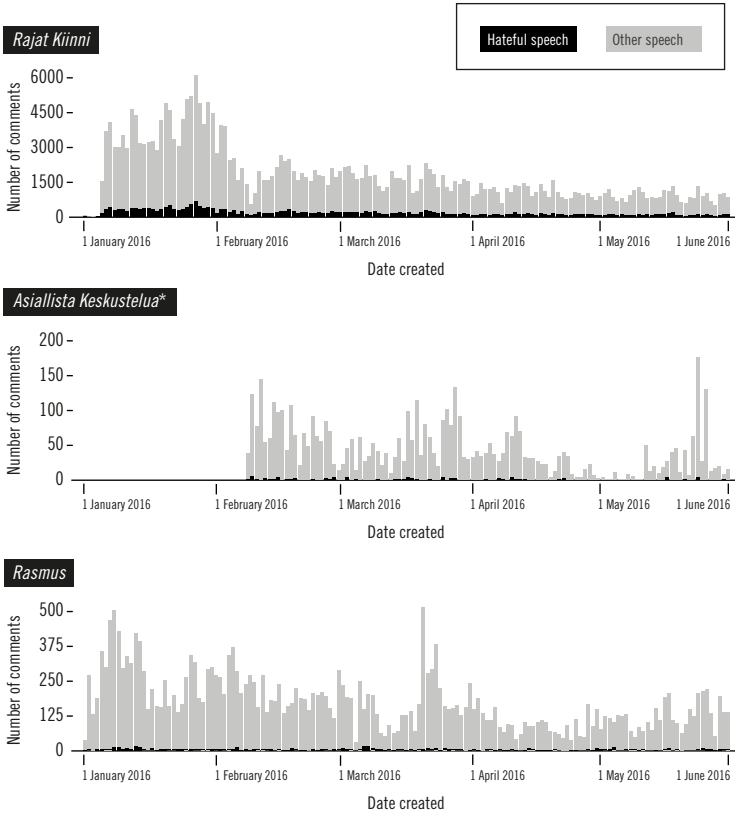
Table 2. The overall distribution of the prevalence of aggressive or hateful comments detected by the two methods used

SUPERVISED CLASSIFIER	*RAJAT KIINNI*	*ASIALLISTA KESKUSTELUA*	*RASMUS*
Percentage of aggressive or hateful comments	10%	2%	1%
Percentage of which are male	77%	37%	41%
Percentage of which are female	23%	63%	59%

TOPIC MODELLING (LDA)	*RAJAT KIINNI*	*ASIALLISTA KESKUSTELUA*	*RASMUS*
Percentage of aggressive or hateful comments	21%	3%	4%
Percentage of which are male	70%	50%	46%
Percentage of which are female	30%	50%	56%

As Figure 5 also indicates, the proportion of these comments identified as aggressive or hateful remained consistent throughout the research period.

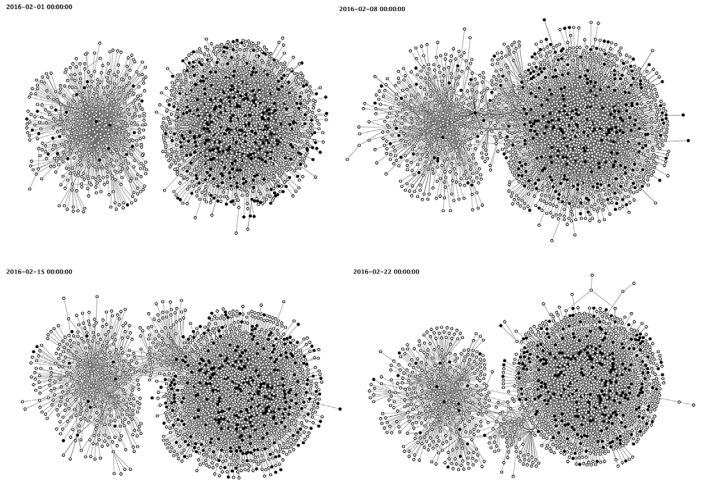
Figure 5. The proportion of aggressive or hateful comments detected over a six-month period (using the classification method), with aggressive or hateful comments in black



* The *Asiallista Keskustelua* group was created at the start of February 2016.

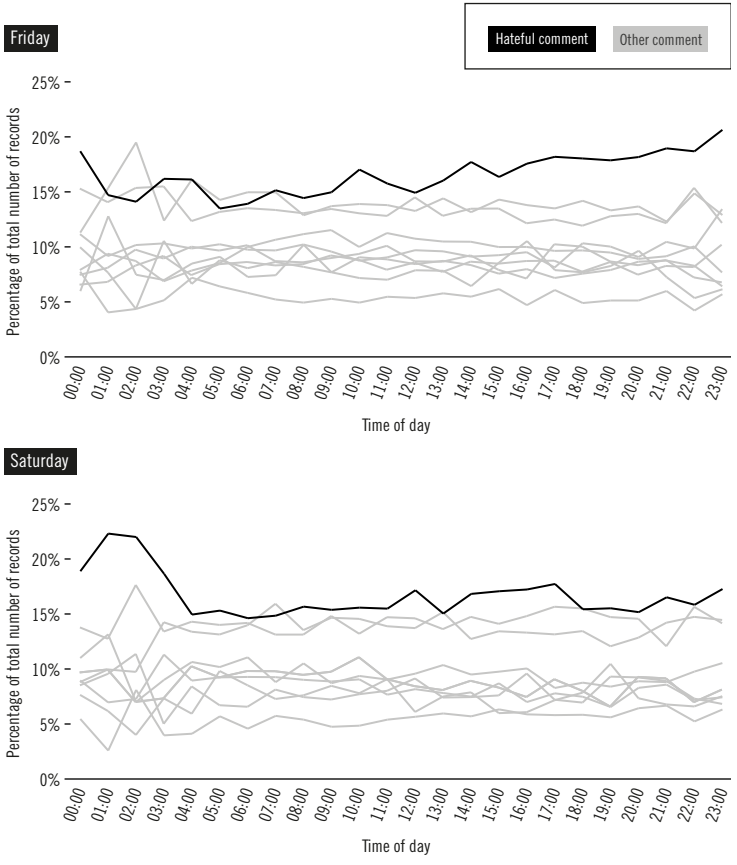
Where were such aggressive or hateful comments then found in these three groups? A network visualisation in Figure 6 shows that there were no distinct patterns, with the exception that they were found predominantly in the *Rajat Kiinni* group. The cluster on the left represents the *Rasmus* group, and the cluster on the right represents the *Rajat Kiinni* group. This holds true for both before and after the anti-refugee/anti-immigration, and anti-racist groups were bridged by the discussion group in-between. The comments that were labelled aggressive or hateful (by the classification method) are coloured in black.

Figure 6. The weekly network evolution of the three groups in February 2016 with aggressive or hateful comments in black



One interesting finding of the research was the time of day such comments were made. Figure 7 shows that during the evenings, there was a significant increase in the proportion of comments identified as aggressive or hateful. Moreover, there also seemed to be a discernible spike in these comments late on Friday nights and into the early hours of Saturday mornings.

Figure 7. The relative percentage of types of comment according to time of day (identified by the topic modelling approach)



These findings fit with other research on antisocial behaviour online, which has argued that such behaviour is usually informed by two factors: an individual's pre-existing mood, and the discussion context in which he/she is writing (see Cheng et al. 2017). One plausible explanation behind these patterns is that members of the *Rajat Kiinni* group who wrote hateful comments later in the evenings were more likely to be intoxicated, and thus predisposed to react aggressively to what they read online. This is also confirmed by a more qualitative observation of the Facebook profiles

of *Rajat Kiinni* members who were the most prolific in writing such comments. They consisted mostly of middle-aged white men whose Facebook profiles combined everyday Facebook activity with the sharing of nationalistic and anti-immigrant memes. Conversely, the discussion context in which these comments were made – a Facebook group where such behaviour is widely accepted and applauded – also provided a fertile ground for this aggressive or hateful style of communication to proliferate.

When combined with a more ethnographic exploration of these conversations, one conclusion can be drawn from the dataset. This is that during moments such as late at night or early on Saturday morning, information read online can trigger a strong response (see Dean 2010). There is a kind of vicious cycle involved whereby individuals who are already predisposed to react aggressively also seek out information that confirms this reaction. Furthermore, when peers in the group widely encourage the use of offensive language, it gradually becomes the new 'normal' – a kind of ritualised opposition to mainstream norms and language that is commonly found in groups associated with the extreme right (See Udupa 2017; Hervik 2019).

3.3 HATEFUL SPEECH AND EXTREME RIGHT DISINFORMATION

If information read online can trigger such a strong response, how then did different types of information shared in these three groups relate to the prevalence of such hateful conversations? And, in particular, how did news shared from extreme right news sources relate to conversations found in the anti-refugee/anti-immigrant *Rajat Kiinni* group? Table 3 shows that members of the *Rajat Kiinni* group overwhelmingly shared more URLs classified as 'alternative (extreme right)'. On the other hand, the *Rasmus* group shared more information from URLs classified as 'news (mainstream)'. The *Asiallista Keskustelua* group was positioned between these groups with a more hybrid media ecology consisting of both mainstream and alternative information sources. What was also noteworthy about the types of URL shared was that this reverse relationship applied

only to news sources classified as alternative (extreme right)/news (mainstream). Other categories such as tabloid, entertainment, blog, or social media were distributed more evenly across the three ideologically opposed groups.

Table 3. Different information sources shared in the groups

	RAJAT KIINNI	***ASIALLISTA KESKUSTELUA***	***RASMUS***
News (mainstream)	36%	50%	58%
Alternative (far right)	26%	13%	2%
Tabloid	24%	25%	20%
Entertainment	8%	7%	9%
Blog	5%	4%	6%
Social media	2%	2%	4%

The same pattern also held across the three most popular online news sources shared: a mainstream Finnish Broadcasting Company (*Yle*), a popular tabloid (*Iltalehti*), and an alternative right news source (*MV-lehti*). Table 4 shows indeed how one of the most popular news sources in the *Rajat Kiinni* group was *MV-lehti*, a controversial online news site with close ties to the extreme right, whereas it was not shared at all by the anti-racist *Rasmus* group.[10]

10 The most visible example of an extreme right website was *MV-lehti*. Originally, a magazine called *Mitä Vittua?* (*What the Fuck?*), this website became controversial during the refugee crisis for publishing stories with an anti-immigrant/refugee slant and personally attacking people supporting refugees. It is also known to have open ties to the Finnish Resistance Movement, a neo-Nazi organisation in Finland. As a result of the activities of *MV-lehti*, the police received dozens of criminal complaints against the website, including accusations of aggravated slander, hate speech, and copyright infringement. An arrest warrant was issued against its founder who had moved its operations to Spain. He is now awaiting trial in Finland. See jacobinmag.com/2017/04/true-finns-finland-timo-soini-nationalists-far-right-xenophobia-elections.

Table 4. The percentage of the three biggest news sources shared in the three groups

	RAJAT KIINNI	***ASIALLISTA KESKUSTELUA***	***RASMUS***
Yle (mainstream)	11%	18%	11%
MV-lehti (extreme right)	9%	3%	0.02%
Iltalehti (tabloid)	11%	7%	9%

A more detailed analysis of all the information shared in the *Rajat Kiinni* group also reveals how influential this alternative/extreme right information ecology online was among the anti-refugee/anti-immigration groups in Finland during the refugee crisis. Table 5 shows the root domains that were shared over 500 times in the *Rajat Kiinni* group. Six of these were linked with the alternative/extreme right news ecology or had sympathies towards it.

Table 5. The top domains that were shared more than 500 times in the *Rajat Kiinni* group

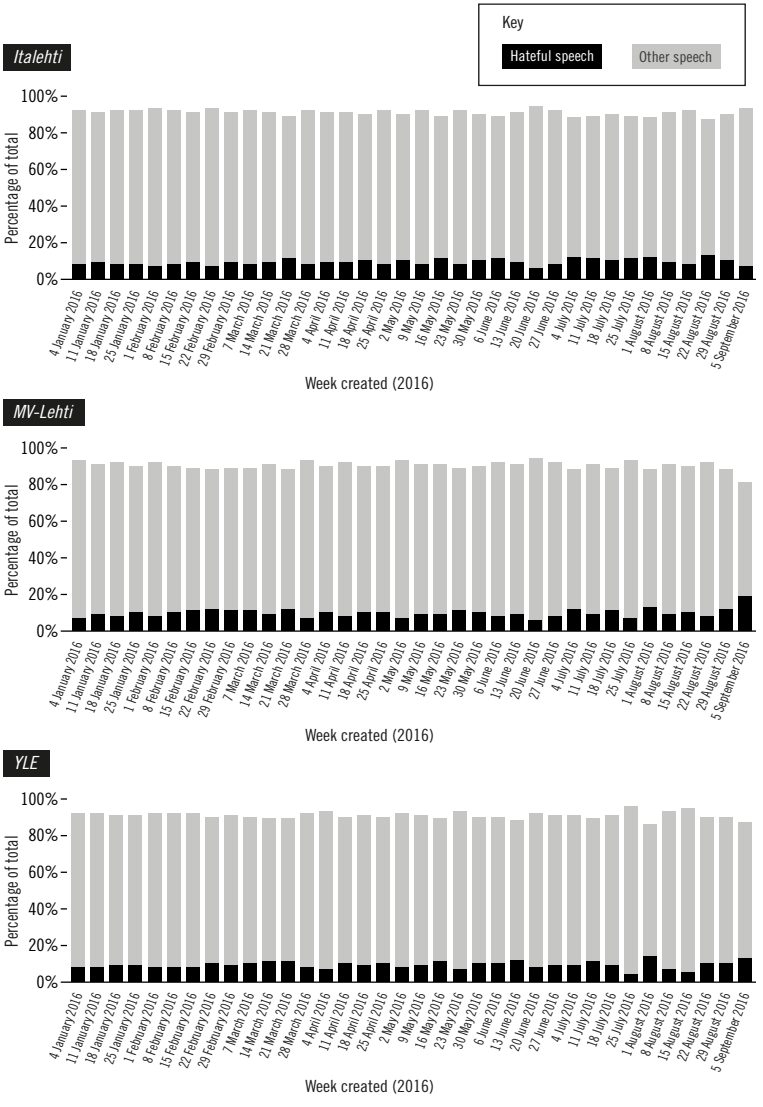
URL	TIMES SHARED	DESCRIPTION
iltalehti.fi	6,240	Tabloid
mvlehti.net	5,504	Alternative/extreme right (news)
yle.fi	5,260	Mainstream news
mtv.fi	3,413	Television channel
verkkouutiset.fi	1,804	News (linked to the National Coalition Party)
hs.fi	1,663	Mainstream news

URL	TIMES SHARED	DESCRIPTION
riippumatonmedia.com	1,325	Alternative/extreme right (news)
uusisuomi.fi	1,281	Mainstream news
paavotajukangas.com	1,274	Alternative/extreme right (blogger)
facebook.com	1,165	Social media
finnleaks.net	1,062	Alternative/extreme right (news)
helsinginuutiset.fi	1,034	Local news
uberuutiset.fi	858	Alternative/extreme right (news)
suomenuutiset.fi	683	News (linked to the True Finns political party)
express.co.uk	623	British tabloid
youtube.com	569	Social media/video
anarkisti.vuodatus.net	522	Alternative/extreme right (blogger)

How did these news sources then relate to the comments that were classified as aggressive or hateful? Contrary to what the research had expected, there was no simple discernible relationship between the information shared and the prevalence of such comments. As Figures 8 and 9 indicate, the comments were evenly distributed across all types of news source shared.

Figure 8. The types of news source shared and their relationship to aggressive or hateful statements in the *Rajat Kiinni* group (classification method)

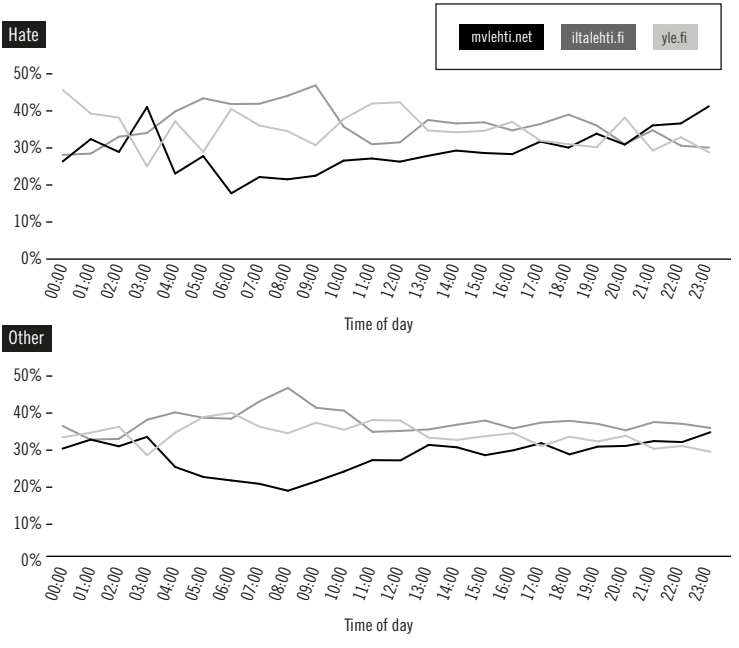
Figure 9. The three most popular news sources and their relationship to aggressive or hateful statements in the *Rajat Kiinni* group (classification method)



However, the data also suggests that where the extreme right information sources differed from other news sources was in the intensity of the conversations they provoked. Posts classified as aggressive or hateful, *and which also shared content from the extreme right MV-lehti*, seemed to incite more spikes in the number of comments and likes.

This also fits with the previous finding that the style of communication in the *Rajat Kiinni* group can be partially explained as a strong reaction to the information read online. As Figure 10 further illustrates, members of the *Rajat Kiinni* group who shared alternative extreme right news sources such as *MV-lehti, and who also wrote aggressive or hateful comments in response to this*, again did so proportionally more late at night and in the early hours of the morning.

Figure 10. The relative percentage of aggressive or hateful comments according to time of day from the three most popular news sources (using the classification method)



Finally, even if there was no clear relationship between the type of information source and the prevalence of aggressive or hateful comments, there were, nonetheless, clear differences in the kinds of comment that were made in response to the type of content shared. A qualitative examination of the conversations in the *Rajat Kiinni* group shows how comments on information shared from extreme right news sites (such as *MV-lehti*) were predominantly targeted at the content that was shared. On the contrary, members who responded aggressively to news shared from mainstream news sites (such as *Yle*) were responding to how they believed the mainstream news was misrepresenting the issue. A comparison of two popular articles demonstrates this distinction.

On 20 January 2016, *MV-lehti* published an article which (falsely) claimed that the entrance requirements to the Finnish police training school were made easier so that refugees and immigrants would have an easier chance of getting in. This article was shared by a member of the *Rajat Kiinni* group, who prefaced the post with the following commentary:

> There is no fucking point any more with selection criteria of quality, when all kinds of 'hairy wrists' and niggers can pass the queue to become police. Soon it will not matter whether somebody has killed a person, as long as they have not killed 10 persons.
>
> This society is sick and extremely unwell![11]

11 After much deliberation with colleagues, I decided to leave these (translated) posts and comments largely unedited even though they contain offensive language. I think it is important to correctly represent the texture and tone of the conversation as it is relevant to the argument that I am presenting here.

This post was commented on 128 times. The comments indicate how the aggression was directed primarily at the content of this article:

> Of course, the entrance criteria have to be made easier; an imbecile's intelligence cannot pass even elementary school. I wonder who came up with the idea to train retarded incest monkeys to become police. It is sure they will not be preventing Muslims rioting.
>
> ... Now we are collecting and weaponising immigrants to protect the government and other shit.
>
> ... What fucking sense does this have? An idiot sells his country to Islam? I think I will pick up my stuff and leave this shit country of losers if this passes?

When information was shared from the mainstream news source *Yle*, the reaction was distinctly different. Another popular news article was shared in the *Rajat Kiinni* group on 22 February. The article was about Soldiers of Odin, the extreme right vigilante group notorious in Finland in the first half of 2016. The member of the *Rajat Kiinni* group who shared this article was angry at the way *Yle* was referring to Soldiers of Odin as a 'radical ultra-nationalist group'. This post received 74 comments. Unlike the case with news shared from *MV-lehti*, however, this time the comments were targeted not at the content of the article, but at its source – that is, how the mainstream media was allegedly framing the topic. The following comments illustrate this clearly:

> This is shit propaganda by *Yle*, something that Kim Jong Un (sic!) of North Korea is jealous.
>
> Fuck *Yle* is cancer.
>
> Again the same *Yle* shit propaganda. Dammit! Fuck what a 'suvakki' [derogatory term for liberals] retard company.
>
> 'Suvakki' propaganda! Don't become disillusioned. You are needed! I don't trust the dickless and understaffed police

> anyway. Immigrant gangs are growing and organising like happened in Cologne and in Berlin. Europe is drowning in shit.

The comments indicate how influential these alternative extreme right new sites had become in anti-refugee/anti-immigration social media groups during the refugee crisis. When the members of the *Rajat Kiinni* group reacted aggressively to news shared from the extreme right *MV-lehti*, this was primarily a reaction to the content that was shared. Conversely, when the news was shared from mainstream news sites such as *Yle*, this reaction was targeted instead at how the mainstream media was discussing the issue. The results, therefore, strongly suggest that proliferation of the extreme right (or extreme right associated) news sources online cannot be understood from the perspective of a simplistic causal or quasi-causal effect that the content produced on these online news sites has on its readers. Instead, more research is needed to know how these alternative news sources are able to act as such authoritative sources of information. In this way, they provide a more extreme framing of the debates that finds resonance in audiences who are already predisposed to react aggressively or hatefully towards this kind of content in the first place (see Archetti 2015a, 2015b). This kind of radical 'frame alignment' through which groups drift away from a shared understanding has also been identified as one of the socio-psychological characteristics of escalating conflict. Therefore, antagonistic parties cease to not only communicate with each other but also begin to articulate critical events and issues in often incommensurable ways using different sources of information to support their conflicting viewpoints (see Hall 1973; Laclau and Mouffe 1985; Aly 2017; see also Della Porta and Diani 2006; Desrosiers 2012).

3.4 HATEFUL SPEECH AND SPACES OF ENGAGEMENT

If the antagonistic groups articulate critical issues in such different ways, how successful then was the group set up to bridge these

polarised ideological echo chambers through fostering a 'civilised discussion' about immigration? The *Asiallista Keskustelua* Facebook group prefaces its purpose by saying, 'this group has been created to be a space for a true encounter between people, not a venue for frustration'. To facilitate this kind of discussion, the group admins enforced a set of guidelines about what kind of conversation was permitted. It was forbidden, for instance, to engage in the following types of behaviour:

- Provoking the other discussants on purpose;
- Attacking another person (for instance, by questioning the intelligence of that person, or asking if they live on welfare);
- Using disrespectful terms about immigrants/refugees;
- Using hate speech (e.g. all Muslims are terrorists, they should be shot, raped, etc.);
- Name calling (e.g. idiot, racist, fascists, 'suvakki', redneck, etc.);
- Whining about what another Facebook group has said or how it has been moderated;
- Shouting, that is, expressing ideas using only capital letters and/or using numerous exclamation or questions marks;
- Constantly questioning the moderation rules of the group (tips for improvement should be in private messages); and
- Comparing humans to animals (e.g. parasite, monkey, etc.).

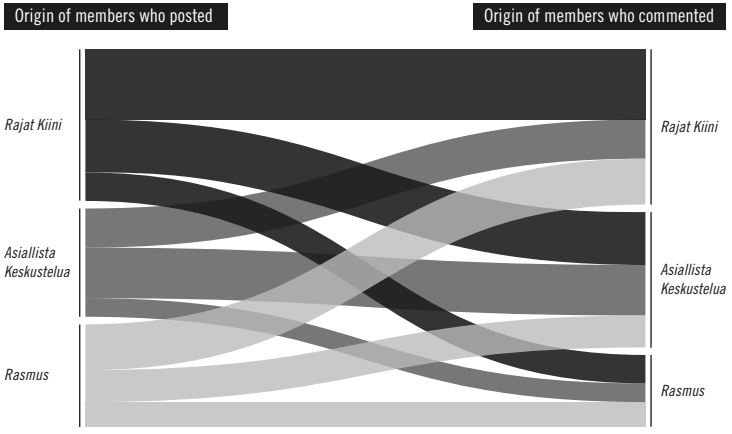
To explore the cross-group dynamics between these ideologically opposed groups, the research first identified the *group origin* of the members most active in the *Asiallista Keskustelua* Facebook group. Table 6 shows how almost half of all the posts and comments were by members from the *Rajat Kiinni* group. Less than a fifth of the posts and comments were from members of the *Rasmus* group, and around one-third of the members had not participated in either group before (Neither).

Table 6. The group origin of the individuals who posted and commented in the *Asiallista Keskustelua* group

	RAJAT KIINNI	*RASMUS*	NEITHER
Posts	51%	19%	30%
Comments	43%	19%	38%

The research also explored how much cross-group dialogue took place between the members of the different factions. Similarly, Figure 11 shows how there was abundant cross-group dialogue between the members who posted (on the left) and who commented (on the right).

Figure 11. The group origin of the members who posted and who commented in the *Asiallista Keskustelua* group



Despite active engagement, however, the research also found that members of the opposing groups in the *Asiallista Keskustelua* discussion group still largely framed key events and issues in antagonistic ways. The different responses to an article shared by a member of the anti-racist *Rasmus* group from the mainstream news site *Yle* on 16 March 2016 clearly illustrate this difference. This article was about

the internal communications of Soldiers of Odin, which had revealed that the members of this vigilante group routinely shared Nazi memorabilia and pictures of weapons, and maintained direct contact with the editors of *MV-lehti*. The member of the *Rasmus* group who shared this article prefaced it with the following question: 'Do we really want these guys to patrol our streets?' Some 160 comments were left in response. A total of 55% of these responses were from the *Rajat Kiinni* group, 17% from the *Rasmus* group, and 28% were from members who did not belong to either group.

The members of the *Rajat Kiinni* group described Soldiers of Odin as a patriotic group made up of normal Finnish people volunteering to keep the streets safe. The comments from the *Rajat Kiinni* group also complained about a smear campaign that mainstream media news sites such as *Yle* were conducting against Soldiers of Odin. Some of the comments included:

> But Soldiers of Odin is not a violent group. They are fathers and mothers as well, they go to work, and they volunteer to do this.
>
> Yes! These words remind us that we have a nation we need to defend. Odin does not cause trouble or get provoked easily. They give safety to people on the streets. And what best: they activate by their example other to react positively when people are mistreated.
>
> Every smart person can figure out that *Yle* has a witch-hunt going on. I do not comment on the Odins but I wish the best to the *MV-lehti* in its battle against a biased and problematic *Yle*.

The members of the *Rasmus* group, on the other hand, expressed the revelations about Soldiers of Odin together with ongoing debates on racism, and the broader rise of fascism in Europe. They also dismissed *MV-lehti* as an authoritative source of news, criticising people who shared content from it for lack of media literacy. Some of the comments included:

> If these comments represent critical viewpoints on immigration as a whole, I am not surprised why people in contemporary Finland are afraid of the rise of fascism.
>
> In my opinion the comment 'two parties fighting each other' and 'we cannot know which side is trustworthier' fit better to other contexts rather than trying to compare *Yle* and the *MV-lehti*. A two-year-old Internet magazine and a 90 years old state organisation who employs thousands of people cannot be seriously compared with each other.
>
> Why is Soldiers of Odin so childish that they avoid on purpose any connection to liberals? Why, if they claim to protect everybody, they cannot give *Yle* an interview? Why do nationalistic people in this country hate Muslims[?] It is futile to claim that Soldiers of Odin would not be anti-immigration. I am sure they are people who want to just protect the streets. But as an organisation the agenda is clear.

Similar antagonistic framing was also found in the exploratory analysis of the different words that were closely associated with key terms relating to the refugee crisis. For instance, words associated most closely with the term *pakolainen* (refugee) in the *Rajat Kiinni* group included words with negative connotations such as 'parasite', 'invader', and 'welfare refugee'. Conversely, words associated with this term in the *Rasmus* group included words with more positive connotations, such as forced movement of people, and the need for help, such as 'departure', 'escape', 'help', and 'poverty'. Similarly, when the research looked at the words associated with the term *rasisti* (racist) in the *Rajat Kiinni* group, its members associated this term with words connoting the expression of opinions, accusation or stigmatisation. Meanwhile, the members of the *Rasmus* group associated it with words such as 'True Finns' (the populist party opposed to refugees and immigration start), '*porukka*' (a word meaning 'group', often used in reference to members of the *Rajat Kiinni* group), or '*maahanmuuttokriittinen*' (a person who is critical of immigration), stressing again the common links between racism, the extreme right, and anti-immigration groups.

Furthermore, what was interesting about this exploratory analysis were the words related to the term *vihapuhe* (hate speech) itself. For members of the *Rajat Kiinni* group, this term was closely associated with words connoting opinions, accusations, and being judged. Conversely, for the *Rasmus* group, this term was closely associated with words connoting violence, xenophobia, incitement, discrimination, and zero tolerance. Table 7 illustrates word associations extrapolated using the word embedding method.

Table 7. Word association of the term *vihapuhe* (hate speech) in the comments of the three groups (based on cosine similarity)

RAJAT KIINNI	***ASIALLISTA KESKUSTELUA***	***RASMUS***
Vihapuhe (hate speech)	*Vihapuhe* (hate speech)	*Vihapuhe* (hate speech)
Rasismi (racism)	*Rasismi* (racism)	*Rasismi* (racism)
Rasisti (somebody who is racist)	*Puhe* (speech)	*Kaikenlainen* (all kinds of)
Rasistinen (something that is racist)	*Syrjintä* (discrimination)	*Rasistinen* (racist)
Kauhea (horrible)	*Rasistinen* (racist)	*Syrjintä* (discrimination)
Kohdistua (targeted towards something)	*Puhua* (to discuss)	*Sallia* (to allow)
Mielestä (in one's opinion)	*Tykätä* (to like)	*Väkivalta* (violence)
Syyttää (to accuse)	*Turha* (pointless)	*Muukalaisviha* (xenophobia)
Kohtaan (against something)	*Musta* (black)	*Selkeästi* (clearly)
Viha (hate)	*Väittää* (to argue/claim)	*Sananvapaus* (freedom of speech)
Määritellä (to define)	*Viha* (hate)	*Uhkailu* (threats)

RAJAT KIINNI	***ASIALLISTA KESKUSTELUA***	***RASMUS***
Termi (a term)	*Tuntua* (to feel)	*Kiusaaminen* (trolling)
Määritelmä (a definition)	*Yleisesti* (in general)	*Nollatoleranssi* (zero tolerance)
Rikos (a crime)	*Asenne* (attitude)	*Täyttää* (fills)
Väittää (to claim)	*Rasmus* (The *Rasmus* group)	*Puuttua* (intervene)
Viharikos (hate crime)	*Kohdistua* (targeted towards something)	*Kiihottaminen* (incitement)
Tuomita (to judge)	*Rasisti* (racist)	*Kansaryhmä* (group of people)
Uhkaus (a threat)	*Vastainen* (against something)	*Hyväksyä* (accept)

Using word embedding to underline differences and biases in how the various social media communities frame events and issues is exploratory. Nevertheless, these findings suggest that even the fundamental concepts associated with the refugee crisis debate were articulated in radically different ways by members of the ideologically opposed groups.

Moreover, these differences were present in the language used to describe the terms of the debate. This difference is also confirmed by the more ethnographic observations of the groups. Members in the *Rajat Kiinni* group routinely articulated terms such as 'racism' and 'hate speech' as attempts by the mainstream and the 'liberals' to silence their opinions and hide the 'real truth' about the costs of immigration. Members of the *Rasmus* group, on the contrary, articulated these terms according to a more mainstream criticism of racist speech and the anti-immigration extreme right. The *Asiallista Keskustelua* group, in turn, held a more ambiguous position, given the strict rules set for conversation in the group, and the language that was allowed in this conversation.

How successful, then, was the *Asiallista Keskustelua* group in creating a 'civilised conversation' across these ideological chasms? On the one hand, it was successful in initiating a dialogue between individuals who seldom, if ever, engaged with each other. There was an abundance of debate between people who, before this, lacked a shared space to do so. On the other hand, whether this conversation resulted in the emergence of a shared inter-communicative understanding of debates relating to refugees is unlikely. On the contrary, the research suggests that the differences between polarised groups run deeper than just a simple lack of engagement across social media echo chambers. These differences have, perhaps, more to do with the antagonistic ways in which different political factions in society, and the anti-immigration extreme right in particular, understand the contours of some of the fundamental debates in contemporary Europe. While the *Asiallista Keskustelua* group indeed provided a laudable experiment in creating engagement across this polarised debate, bridging these deeper ideological fissures will perhaps require more work than creating another Facebook page, or promoting engagement or counter-speech.

4. DISCUSSION AND CONCLUSION

What can this exploratory analysis tell us more broadly about the ongoing debates on hate speech and violent online political extremism? If, as comparative research into situations of violent conflict has suggested, this kind of polarised style of communication is a symptom of some underlying conflict, how can we approach these social media conversations from such a critical-comparative perspective?[12]

Coleman (2003, 2004, 2006) has argued that such conflicts are always framed in different ways by their participants. This also includes the perspective of the researcher who hopes to understand it, or the policymaker who wants to change it. These frames are both implicit (hidden and often unconscious presuppositions about the object of study) as well as explicit (formal methods used to produce knowledge about it). Coleman (2004, pp. 202–226) further identifies five such 'meta-paradigmatic frames' through which conflicts have been historically imagined:

- REALISM: The first frame understands conflicts as the struggle between groups of people in a world where resources are scarce.
- HUMAN RELATIONS: The second frame understands conflicts as the outcome of destructive relationships caused by fear, distrust, misunderstanding, hostile interactions, and lack of constructive engagement between the participants involved.
- THE MEDICAL MODEL: The third frame understands conflicts as the outcomes of some malignant and pathological processes in society that, like disease, can be diagnosed and cured.
- POSTMODERNISM: The fourth frame sees such conflicts from the perspective of narratives though which people make sense of the world and interact with each other based on these frames of understanding.
- SYSTEMS THEORY: The final frame sees conflicts as the outcomes of many interacting levels in a system, where each

12 For an earlier version of the argument see Pohjonen and Ahmed 2016.

of the component parts needs to be understood in a complex relationship with the others.

More importantly, Coleman argues that such 'meta-paradigmatic' frames used to understand conflict also prescribe the methods that can be used to solve it. Hate speech from the perspective of the realist paradigm is the outcome of a political struggle for power through which antagonistic factions in society try to gain supremacy. The methods used to mitigate it include things such as creating legal and institutional frameworks to counter it, or using counter-propaganda to oppose it. Conversely, if seen from the perspective of the human relations paradigm, hate speech is the outcome of an underlying cycle of destructive social relationships. The solution, then, includes creating new ways to increase mutual understanding by supporting reconciliation, dialogue, interdependence, cooperation, and co-existence. Alternatively, from the perspective of the medical paradigm, hate speech is the outcome of some underlying social pathology. In this case, the processes causing it would need to be diagnosed and treated (such as what is the frustration and disenfranchisement that makes some middle-aged white men behave aggressively on social media). From the perspective of the post-modern paradigm, hate speech is the outcome of the antagonistic frames of meaning and narratives people create to understand each other. In this case, the solution to it foregrounds creating new narratives as a way to make participants see the need for change. Or, finally, hate speech in social media can also be seen from the perspective of complex systems theory that sees it as the outcome of multi-layered processes that defy easy explanation. In order to understand such non-linear processes behind it, more research is, therefore, needed; research that is able to take into account the complex and multi-faceted processes causing it that go beyond a reductionist analysis of textual content: 'one of the more commonly applied principles is the idea that intractable conflicts cannot be traced to one or two essential causes but rather should be seen as the result of multiple, interactive elements' (Coleman 2004, p. 223).

In conclusion, then, what the exploratory findings of the research suggest is that perhaps the most dangerous challenge facing Europe is not the explosion of aggressive or hateful content on social media, however offensive and in bad taste much of it is. Instead, the real challenge facing Europe is perhaps this shrinking horizon of understanding between opposing members of society, whereby even the basic concepts of the debate are not understood in mutually commensurable ways. What is especially disconcerting about this finding is that even the sources of information used to produce meaning about critical issues and debates are drifting apart. The extreme right news ecology online has been able to hijack some of the role previously maintained by the mainstream news, to provide an alternative and more extreme framing of events and themes to the audiences who are already predisposed to react strongly to this kind of information. This political polarisation is perhaps a more intractable problem to solve than merely removing aggressive or hateful comments from Facebook.

A critical-comparative analysis of social media hate speech, and of violent online political extremism more broadly, can, therefore, provide two contributions to existing research on social media hate speech. On the one hand, it can help us to soberly assess the dangers of this speech in peaceful countries such as Finland, which are still a long distance away from erupting into the kind of widespread ethnic or political mass violence that we have seen in situations of violent conflict in other parts of the world. But, on the other hand, stepping back from the ongoing debates, even temporarily, can provide the necessary conceptual distance needed to come up with new ideas and strategies that can help to prevent such violence from ever happening in the first place.

REFERENCES

Aly, A., 2017. 'Brothers, Believers, Brave *Mujahideen*: Focusing Attention on the Audience of Violent Jihadist Preachers'. *Studies in Conflict and Terrorism*, 40 (2), pp. 62–76.

Archetti, C., 2015a. '(Mis)communication Wars: Terrorism, Counter-terrorism and the Media', in D. Welch, (ed.), *Propaganda, Power and Persuasion*. London: IB Tauris, pp. 209–224.

Archetti, C., 2015b. 'Terrorism, Communication and New Media: Explaining Radicalization in the Digital Age'. *Perspectives on Terrorism*, 9 (1), pp. 49–59.

Article 19, 2015. '*Hate speech' Explained: A Toolkit*. London: Article 19. Available from: https://article19.org/data/files/medialibrary/38231/'Hate-Speech'-Explained---A-Toolkit-%282015-Edition%29.pdf

Bartlett, J., Reffin, J., Rumball, N. and Williamson, S., 2014. *Anti-social Media*. London: Demos. Available from: www.demos.co.uk/files/DEMOS_Anti-social_Media.pdf

Bartlett, J. and Krasodomski-Jones, A., 2015. *Counter-speech: Examining Content that Challenges Extremism Online*. London: Demos. Available from: https://demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

Benesch, S., 2012. 'Dangerous Speech: A Proposal to Prevent Group Violence'. World Policy Institute, 12 January. Available from: https://worldpolicy.org/wp-content/uploads/2016/01/Dangerous-Speech-Guidelines-Benesch-January-2012.pdf

Benesch, S., 2013. 'Countering Dangerous Speech to Prevent Mass Violence during Kenya's 2013 Elections'. Dangerous Speech Project working paper, 9 February. Available from: https://dangerousspeech.org/countering-dangerous-speech-kenya-2013/

Benesch, S., 2014. 'Countering Dangerous Speech: New Ideas for Genocide Prevention'. Dangerous Speech Project working paper, 11 February. Washington, DC: Unites States Holocaust Memorial Museum. Available from: https://dangerousspeech.org/countering-dangerous-speech-new-ideas-for-genocide-prevention

Benkler, Y., Faris, R., Roberts, H. and Zuckerman, E., 2017. 'Study: Breitbart-led right-wing media ecosystem altered broader media agenda'. *Columbia Journalism Review*, 3 March. Retrieved from https://cjr.org/analysis/breitbart-media-trump-harvard-study.php

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A., 2016. 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings'. ArXiv.org, Cornell University, 21 July. Available from: https://arxiv.org/pdf/1607.06520.pdf

Brown, A., 2015. *Hate Speech Law: A Philosophical Examination*. First edition. London: Routledge.

Brown, I. and Cowls, J., 2015. 'Check the Web: Assessing the Ethics and Politics of Policing the Internet for Extremist Material'. VOX-Pol, 23 November. Available from: https://voxpol.eu/check-the-web/

Burnap, P. and Williams, M., 2015. 'Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making'. *P&I Policy and Internet*, 7 (2), pp. 223–242.

Butler, J., 1997. *Excitable Speech: A Politics of the Performative*. London: Routledge.

Buyse, A., 2014. 'Words of Violence: "Fear Speech", or How Violent Conflict Escalation Relates to the Freedom of Expression'. *Human Rights Quarterly*, 36 (4), pp. 779–797.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C. and Leskovec, J., 2017. 'Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions'. ArXiv.org, Cornell University, 3 February. Available from: https://arxiv.org/abs/1702.01119

Coleman, P. T., 2003. 'Characteristics of Protracted, Intractable Conflict: Toward the Development of a Meta-framework-I'. *Peace and Conflict: Journal of Peace Psychology*, 9 (1), pp. 1–37.

Coleman, P. T., 2004. 'Paradigmatic Framing of Protracted Intractable Conflict: Toward the Development of a Meta-framework-II'. *Peace and Conflict: Journal of Peace Psychology*, 10 (3), pp. 197–235.

Coleman, P. T., 2006. 'Conflict, Complexity, and Change: A Meta-framework for Addressing Protracted, Intractable Conflicts-III'. *Peace and Conflict: Journal of Peace Psychology*, 12 (4), pp. 325–348.

Conway, M., 2017. 'Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research'. *Studies in Conflict & Terrorism,* 40 (1), pp. 77–98.

Davidson, T., Warmsley, D., Macy, M. and Weber, I., 2017. 'Automated Hate Speech Detection and the Problem of Offensive Language'. ArXiv.org, Cornell University, 11 March. Available from: https://arxiv.org/pdf/1703.04009.pdf

Dean, J., 2010. 'Affective Networks'. *MediaTropes* eJournal, II (2), pp. 19–44.

Della Porta, D. and Diani, M., 2006. *Social Movements: An Introduction.* Second edition. Oxford: Blackwell Publishing.

Desrosiers, M-E., 2012. 'Reframing Frame Analysis: Key Contributions to Conflict Studies'. *Ethnopolitics*, 11 (1), pp. 1–23.

Ferguson, K., 2016. *Countering Violent Extremism Through Media and Communication Strategies: A Review of the Evidence*. Partnership for Conflict, Crime and Security Research, University of East Anglia, 1 March. Available from: https://paccsresearch.org.uk/wp-content/uploads/2016/03/Countering-Violent-Extremism-Through-Media-and-Communication-Strategies-.pdf

Foulds, J., 2018. 'Mixed Membership Word Embeddings for Computational Social Science'. ArXiv.org, Cornell University, 25 May. Available from: https://arxiv.org/abs/1705.07368

Gagliardone, I., et al., 2016. 'Mechachal – Online Debates and Elections in Ethiopia. Final Report: From Hate Speech to Engagement in Social Media'. The Programme in Comparative Media Law and Policy (PCMLP), University of Oxford, and Addis Ababa University. Available from: http://academia.edu/25747549/Mechachal_-_Online_Debates_and_Elections_in_Ethiopia._Final_Report_From_hate_speech_to_engagement_in_social_media_Full_Report_

Gagliardone, I., Pohjonen, M. and Patel, A., 2014. 'Mapping and Analysing Hate Speech Online: Opportunities and challenges for Ethiopia'. The Programme in Comparative Media Law and Policy (PCMLP), University of Oxford, and Addis Ababa University. Available from: http://pcmlp.socleg.ox.ac.uk/wp-content/uploads/2014/12/Ethiopia-hate-speech.pdf

Gagliardone, I., Gal, D., Alves, T., Martinez, G., 2015a. *Countering Online Hate Speech*. UNESCO Series on Internet Freedoms. Paris: UNESCO Publishing. Available from: http://unesdoc.unesco.org/images/0023/002332/233231e.pdf

Gagliardone, I. et al., 2015b. 'Mechachal – Online Debates and Elections in Ethiopia. Report Two: Discussing Politics and History in Social Media'. The Programme in Comparative Media Law and Policy (PCMLP), University of Oxford, and Addis Ababa University. Available from: http://academia.edu/19593354/Mechachal_-_Online_Debates_and_Elections_in_Ethiopia._Report_Two_Discussing_politics_and_history_in_social_media

Gelber, K., 2011. *Speech Matters: Getting Free Speech Right*. Brisbane, Australia: University of Queensland Press.

Hall, S., 1973. 'Encoding and Decoding in the Television Discourse'. Paper for the Council of Europe Colloquy on 'Training in the Critical Reading of Television Language'. Organized by the Centre for Mass Communication Research, University of Leicester,

September. Available from: https://birmingham.ac.uk/Documents/college-artslaw/history/cccs/stencilled-occasional-papers/1to8and11to24and38to48/SOP07.pdf

Hamelink, C. J., 2011. *Media and Conflict: Escalating Evil*. Boulder, CO: Paradigm Publishers.

Heinze, E., 2017. *Hate Speech and Democratic Citizenship*. Reprint. Oxford: Oxford University Press.

Hervik, P., 2019. 'Ritualized Opposition in Danish Practices of Extremist Language and Thought'. Forthcoming in the *Journal of Communication*, special section on extreme speech.

Human Rights Watch, 2015. *Journalism is not a Crime: Violations of Media Freedom in Ethiopia*. New York: Human Rights Watch. Available from: https://hrw.org/report/2015/01/21/journalism-not-crime/violations-media-freedoms-ethiopia

Human Rights Watch, 2016. 'Such a Brutal Crackdown': Killings and Arrests in Response to Ethiopia's Oromo Protests'. New York: Human Rights Watch. Available from: https://hrw.org/sites/default/files/report_pdf/ethiopia0616web.pdf

Jackson, R., 2012. 'Unknown Knowns: The Subjugated Knowledge of Terrorism Studies'. *Critical Studies on Terrorism*, 5 (1), pp. 11–29.

Kwok, I. and Wang, Y., 2013. 'Locate the Hate: Detecting Tweets against Blacks'. *AAAI'13: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pp. 1621–1622. Available from: https://pdfs.semanticscholar.org/db55/11e90b2f4d650067ebf934294617eff81eca.pdf

Leader Maynard, J. and Benesch, S., 2016. 'Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention'. *Genocide Studies and Prevention,* 9 (3), pp. 70–95. Available from: http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1317&context=gsp

Laaksonen, S.-M., Nelimarkka, M., Tuokko, M., Marttila, M., Kekkonen, A., and Vili, M., 2017. 'Working the fields of big data: using big-data-augmented online ethnography to study candidate–candidate interaction at election time'. *Journal of Information Technology & Politics*, 14 (2), pp. 110–131.

Laclau, E. and Mouffe, C., 1985. *Hegemony and Socialist Strategy: Towards a Radical Democratic Politics*. London: Verso.

Marwick, A. and Lewis, R., 2017. *Media Manipulation and Disinformation Online*. New York: Data & Society Research Institute. Available from: https://datasociety.net/pubs/oh/DataAndSociety_ExecSummary-MediaManipulationAndDisinformationOnline.pdf

Meleagrou-Hitchens, A. and Kaderbhai, N., 2017. *Research Perspectives on Online Radicalisation: A Literature Review, 2006–2016. VOX-Pol Network of Excellence*. Available from: https://www.voxpol.eu/download/vox-pol_publication/Research_Perspectives_Lit_Review.pdf

Pohjonen, M. and Ahmed, R., 2016. 'Narratives of Risk: Assessing the Discourse of Online Extremism and Measures Proposed to Counter It'. *S+F.* (34, Jg.), 4/2016. Available from: http://academia.edu/31918088/Narratives_of_Risk_Assessing_the_Discourse_of_Online_Extremism_and_Measures_Proposed_to_Counter_It

Pohjonen, M., 2018. 'Towards a data-driven digital ethnography: Methodologies, theories and arguments'. Forthcoming in the Communicative Figurations Working Paper series, ZeMKI, Centre for Media, Communication and Information Research, University of Bremen.

Pohjonen, M. and Udupa, S., 2017. 'Extreme Speech Online: An Anthropological Critique of Hate Speech Debates'. *International Journal of Communication*, 11, pp. 1173–1191.

Post, R., 2009. 'Hate speech'. In: I. Hare and J. Weinstein, (eds), *Extreme Speech and Democracy.* Oxford: Oxford University Press, pp. 123–138.

Price, M. and Stremlau, N., 2017. 'Introduction: Speech and Society in Comparative Perspective'. In: M. Price and N. Stremlau, (eds), *Speech and Society in Turbulent Times: Freedom of Expression in Comparative Perspective.* Cambridge, UK: Cambridge University Press.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N. and Wojatzki, M., 2016. 'Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis'. In *Proceedings of NLP4CMC III.* Bochumer Linguistische Arbeitsberichte. Available from: https://arxiv.org/pdf/1701.08118.pdf

Saleem, H. M., Dillon, K. P., Benesch, S. and Ruths, D., 2017. 'A Web of Hate: Tackling Hateful Speech in Online Social Spaces'. Proceedings of First Workshop on Text Analytics for Cybersecurity and Online Safety, 26 July. Available from: https://arxiv.org/abs/1709.10159

Umati, 2012–2013. *Umati Final Report.* Nairobi: iHub Research. Available from: https://preventviolentextremism.info/sites/default/files/Umati%20Final%20Report.pdf

Udupa, S., 2017. '*Gaali* cultures: The politics of abusive exchange on social media'. *New Media and Society,* 20 (4), pp. 1506–1522.

Waldron, J., 2012. *The Harm in Hate Speech.* Cambridge, MA: Harvard University Press.

Wardle, C. and Derakhshan, H., 2017. 'Information Disorder: Towards an interdisciplinary framework for research and policy making'. Council of Europe Report, September. Retrieved from https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

The VOX-Pol Network of Excellence (NoE) is a European Union Framework Programme 7 (FP7)-funded academic research network focused on researching the prevalence, contours, functions, and impacts of Violent Online Political Extremism and responses to it.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 312827

Email info@voxpol.eu
Twitter @VOX_Pol
www.voxpol.eu