



ISD

Powering solutions  
to extremism  
and polarisation

ONLINE

CIVIL

COURAGE

INITIATIVE

# Cartographie de la Haine en Ligne

Online Civil Courage Initiative France

Cooper Gatewood, Cécile Guerin,  
Jonathan Birdwell, Iris Boyer & Zoé Fourel

## About this paper

**This report presents the findings of a research project investigating the scale and nature of online hateful speech in France. It examines different categories of hateful speech, spanning race, gender, sexuality, religion and disability. Using social media data analytic tools that combine machine learning and natural language processing with qualitative analysis, the report provides a detailed analysis of the dynamics of the most prevalent types of hateful speech across social media platforms in France. This report also recommends some steps to be taken by technology companies, government and civil society organisations to counter hateful speech online. Part of the Online Civil Courage Initiative (OCCI), a strategic partnership between the Institute for Strategic Dialogue (ISD) and Facebook, this report aims to inform both civil society and policymakers' responses to online hate.**

## About the authors

### Cooper Gatewood

Cooper is a manager within ISD's Digital Research Unit, focusing on quantitative research into the spread of hateful and polarising narratives online, and how they are leveraged by extremist actors. Cooper also develops monitoring and evaluation frameworks to measure the impact of a number of ISD's intervention projects. Cooper is currently contributing to ISD's research on disinformation campaigns, particularly those aimed to influence and disrupt election processes. He also manages the OCCI in France, co-ordinating activities to support civil society's response to hate and extremism online. In addition, Cooper conducts ongoing evaluation of a number of ISD's programmes, including Be Internet Citizens and Young Digital Leaders. Cooper holds a Master of International Affairs from Columbia University and a Master in International Security from Sciences Po.

### Cécile Guerin

Cécile Guerin is a co-ordinator at ISD, supporting the organisation's European development and analysis work. She works on the OCCI, a Facebook-funded project which aims to upscale civil society efforts against hate speech and extremism online. Cécile also contributes to ISD's research and policy work, with a focus on social media analysis and network mapping related to hate speech, extremism and disinformation online. She has written for a range of publications, including the Guardian, Prospect and the Independent. Cécile holds an MSc in International History from the London School of Economics and an MA in English from the École Normale Supérieure in France.

© ISD, 2019

London | Washington DC | Beirut | Toronto

This material is offered free of charge for personal and non-commercial use, provided the source is acknowledged. For commercial or any other use, prior written permission must be obtained from ISD.

In no case may this material be altered, sold or rented. ISD does not generally take positions on policy issues. The views expressed in this publication are those of the authors and do not necessarily reflect the views of the organisation.

Designed by forster.co.uk. Typeset by Danny Arter.

## About the authors

### Jonathan Birdwell

Jonathan is Deputy Director, Policy & Research, overseeing ISD's policy work and networks, and ISD's work on education policy and programming. Jonathan supervises ISD's research and primary datasets, programme monitoring and evaluation, and edits all of ISD's written outputs. He currently focuses on building out ISD's unique partnerships and online analytic technology and capabilities to provide up-to-date understanding of extremist propaganda and recruitment tactics, and evaluating online campaigns and ISD's online one-to-one interventions. Prior to joining ISD, Jonathan was Head of Programme at the UK cross-party think tank Demos, where he published over 40 research reports on topics including violent extremism both Islamist (*The Edge of Violence*, 2010) and far-right (*The New Face of Digital Populism*, 2011). Jonathan has also written extensively on education (*The Forgotten Half*, 2011), social and emotional learning (*Character Nation*, 2015), youth social action and attitudes towards politics (*Tune In, Turn Out*, 2014), digital politics and marketing (*Like, Share, Vote*, 2014), trust in government (*Trust in Practice*, 2010) and religion and integration (*Rising to the Top*, 2015), among other topics. Jonathan holds a Master's degree (with distinction) from the London School of Economics and Political Science and Bachelor's degrees in Political Science and Philosophy from Tulane University in New Orleans, Louisiana.

### Iris Boyer

Iris is Deputy Head of Technology, Communications and Education, overseeing a number of programmes supporting and amplifying civil society's efforts against extremism through scaled partnerships with tech companies and grassroots organisations. Iris also co-ordinates multi-sectorial networks spanning government, academia, the media and the non-government organisation (NGO) sector, and regularly advises them with ISD's insights on the latest trends in extremism and the most effective and innovative approaches to tackle its mainstreaming. She is also working on ISD's regional expansion, and particularly leading on thought leadership and development in France. Iris holds a French five-year diploma in social sciences and humanities from Sciences Po, as well as an international Master's degree in public affairs from the Higher School of Economics in Moscow and the London Metropolitan University.

### Zoé Fourel

Zoé is an intern at ISD, working predominantly on the OCCI in France. In addition to contributing to this research project, Zoé helps co-ordinate the activities of the OCCI in France. Zoé also contributes to other ISD programmes focusing on empowering civil-society-led responses to hate and extremism. Zoé holds a five-year diploma from Sciences Po Lyon in International Affairs, which included studies at the School of Oriental and African Studies in London and Georgetown University in Washington, DC.

## Acknowledgements

We would like to express our gratitude to members of the ISD, especially Henry Tuck and Chloe Colliver, for their helpful feedback and revisions, and Hannah Martin, who co-ordinated the production of the report.

We would like to thank the organisations and institutions which have provided advice on the methodology of the report and have shed light on our choice of keywords, including: Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA), Le Refuge, La Voix des Roms, the American Jewish Committee (AJC) and #JeSuisLà.

We would also like to thank our partners at the Centre for the Analysis of Social Media (CASM), in particular Carl Miller, Josh Smith, Chris Inskip and Andrew Robertson for their support in classifier training and data analysis.

This report could not have been published without the financial support of Facebook France, which has funded this research. Within Facebook, we are particularly thankful to Clotilde Briend, Hamida Moussaoui and Sarah Yanicostas for their support.

Any mistakes or omissions are the authors' own.

---

# Contents

<b>Executive summary</b>	<b>7</b>
<b>Glossary</b>	<b>10</b>
<b>Introduction</b>	<b>12</b>
<b>Methodology</b>	<b>17</b>
<b>Analyses of discourses</b>	<b>22</b>
<b>1. Gender &amp; sexuality</b>	<b>22</b>
1.1 Misogynistic discourse	22
1.2 Anti-LGBTQ discourse	25
<b>2. Race &amp; ethnicity</b>	<b>29</b>
2.1 Anti-Arab discourse	30
2.2 Anti-Roma and Anti-gypsy discourse	34
2.3 Anti-black or Anti-African discourse	38
2.4 Anti-white discourse	40
2.5 Anti-Asian discourse	42
<b>3. Religion</b>	<b>43</b>
3.1 Anti-Muslim discourse	44
3.2 Anti-Semitic discourse	47
3.3 Anti-Christian discourse	48
<b>4. Disability</b>	<b>50</b>
4.1 Ableist discourse	51
<b>Intersectionality of Hateful Speech</b>	<b>53</b>

<b>Recommendations</b>	<b>59</b>
<b>Appendices</b>	<b>64</b>
Appendix 1	64
Appendix 2	66
Appendix 3	73
<b>Endnotes</b>	<b>75</b>

---

## Executive Summary

**In the last two years, tackling online hate has emerged as a major public concern in Europe. Following the adoption of the NetzDG law (Network Enforcement Act) in Germany, the French National Assembly adopted the Loi Avia to tackle hateful content online. This imposes new requirements for tech companies to remove illegal content from their platforms. At the time of writing (late 2019), the law is still pending review by the Senate, France's upper house of Parliament.**

Tech companies have also stepped up their efforts to identify, analyse and quantify hate speech on their platforms. The fourth report of the EU Code of Conduct on Illegal Hate Speech Online has shown that tech companies are processing 89% of flagged hate content within 24 hours.<sup>1</sup> In 2018, four more companies (Google+, Snapchat, Instagram and Dailymotion) joined the EU Code of Conduct, with Facebook, Microsoft, Twitter and YouTube having joined as founding members in 2016.<sup>2</sup>

Despite legislative changes to tackle online hate speech, the parliamentary report of the Loi Avia<sup>3</sup> highlighted that rigorous data on the scale of the problem is still scarce. It is within this context that the OCCI, a strategic partnership between the ISD and Facebook, aims to provide insights and campaigning resources to civil society in order to improve the civic response to hateful content online.

As a key element of the work of the OCCI in 2019, this report addresses the existing information gap by providing a data-driven overview of a variety of forms of hateful speech online in France across different social media platforms. It aims to give decisionmakers in the public sector, tech companies and civil society organisations insights into the scope and nature of online hate speech to inform policy, technological and civic responses.

The findings of this report draw on data analysis using social listening tools and natural language processing software, combined with qualitative analysis. Covering a period of five months, we produced datasets of different types of hateful discourses, using sets of keywords drawn from ISD's ongoing research of extremist milieus, as well as consultation with French civil society actors working to counter diverse types of hate. We trained

algorithms to identify the proportion of hateful content for the four categories of hateful speech whose keywords returned the greatest number of posts, and conducted qualitative analysis for the other types of hateful speech.

### Key Findings

This report endeavours to outline the most prominent types of hateful speech online in France (drawn from publicly available data), in order to inform civic, policy and technological responses to these divisive discourses. Eleven datasets were created by querying social listening tools with terms frequently associated with content targeting groups based on gender and sexuality, ethnicity and race, religion and disability that had the intention of inciting hatred, violence or discrimination. The four largest datasets were analysed with natural language processing algorithms to identify the scale of hateful speech. Each of these discourses has their own specificities, yet there are key takeaways that are relevant across the spectrum of hateful speech. We found that:

- **Using machine learning, we were able to identify confidently just under 7 million instances of online hateful speech against women; lesbian, gay, bisexual, transgender and queer (LGBTQ) communities; people with disabilities; and French Arab communities.**

This included approximately 5.4 million instances of misogynistic hateful speech, over 1 million instances of anti-LGBTQ hateful speech, 265,000 instances of ableist hateful speech and 131,000 instances of anti-Arab

“ The findings of this report draw on data analysis using social listening tools and natural language processing software, combined with qualitative analysis ”

19%

Of the accounts most frequently posting hateful speech exhibited automated or bot-like behaviour

13%

Of the accounts most frequently posting hateful speech showed affiliation with far-right groups or ideology

hateful speech. Our ability to identify hateful speech confidently and algorithmically for these groups was facilitated by the very large scale of the initial datasets of hateful keywords related to these groups, demonstrating the widespread use of slurs and insults that originally targeted them. It was not possible to identify hateful speech algorithmically for the other groups, requiring more manual and qualitative analysis. The scale of hateful speech online is likely to be significantly higher, but cannot be determined owing to lack of access to public communications on many social media platforms.

- **Most hateful speech online comprised the generalised use of slurs and insults based on attacks of protected categories.** The normalised use of misogynistic language (like *sale pute*, *dirty bitch*), homophobic slurs (like *pédé*, *fag*) and ableist insults (like *mongol*, pejorative slang for someone with Down Syndrome) figured prominently across our sample. We identified over 4 million posts using hateful misogynistic language, just under 1 million posts using hateful anti-LGBTQ language and just under 250,000 posts using hateful ableist language. There was often significant overlap in accounts using hateful speech. Hateful misogynist, anti-LGBTQ and ableist slurs were used most often to attack politicians and footballers.

- **A small percentage of hateful speech was made up of targeted attacks on individuals.** Around 5% of our dataset was made up by direct misogynistic attacks on users who appeared to be women on the basis of their gender. While it was not possible to identify all targeted attacks across the different types of discourse using machine learning, qualitative analysis suggests that the groups that had the highest scale of targeted hateful online speech were women, Arabs and Muslims.

- **Of the accounts most frequently posting hateful speech, roughly one in five (19%) exhibited automated or bot-like behaviour.**<sup>4</sup> Around 13% demonstrated affiliation with a far-right group or ideology, 4% were associated with the Yellow Vest Movement and 4% had been deleted by the time of writing. French alternative news sources also made up a small portion of these accounts. All ten of the top ten most active accounts on Twitter sharing anti-Muslim keywords were affiliated with the far-right.

- **Events like International Women's Day and the announcement of the nomination of Bilal Hassani to the Eurovision Song Contest drove spikes in hateful speech.** Other events that drove hateful speech against Muslims and Arab users included the start of Ramadan and the Christchurch attack.

- **Our keyword-based approach revealed a small amount of organic and co-ordinated counterspeech efforts on social media.** For example, roughly 1% of the dataset built from misogynistic keywords was made up of users pushing back against the use of misogynistic language and slurs. This type of speech was also notable in the anti-LGBTQ, anti-black and anti-Roma datasets. Counterspeech efforts included the re-appropriation of hateful slurs.

- **There was significant overlap between the different types of hateful speech, demonstrating the need for an intersectional analysis of hateful speech online.** This was particularly true of the hateful misogynistic and anti-LGBTQ speech, as well as hateful anti-Arab and anti-Muslim speech.

- **This research demonstrated the possibilities and limits of working with natural language processing algorithms to identify hateful speech online.** Our researchers were able to train algorithms to be 85% accurate in identifying hateful speech online – a high level for this type of research.<sup>5</sup> However, the diversity in terminology and usage of hateful terms posed significant obstacles, demonstrating the need for a holistic approach to identifying and moderating hateful speech online.

## Recommendations

1. **Online platforms should increase transparency on public content on their platform by providing open application programming interface (API) access to provide better understanding of the scale of hateful discourses.** Online platforms should provide greater data access to vetted research organisations to enable better understanding and ability to challenge hateful speech online. Twitter's current approach provides a model for balancing transparency and research access with privacy protections.
2. **Government regulators and online platforms must consider the limits of machine learning algorithms when identifying hateful content.** Artificial intelligence should not be seen as a panacea, and approaches to moderating content that may be hateful or designed to provoke hatred must include human review. Image and video content pose further challenges to this approach.
3. **Online platforms should work closely with staff in civil society organisations to tackle hateful content that is legal but nonetheless problematic and harmful.** The Loi Avia will require online platform operators to remove manifestly illegal hateful content. Partnerships with civil society organisations should be pursued to tackle the much larger body of hateful content that will not meet this threshold. This should include trialling and testing counterspeech approaches that involve a range of direct engagement techniques. Civil society organisations can help online platforms decide what the appropriate responses are to different types of hateful content.
4. **Online platforms should provide increased transparency on moderation policies and approaches, and the role of algorithms and automated accounts in spreading hateful content.** Transparency on content moderation processes and greater oversight from a government regulator, as outlined in the inter-ministerial mission's interim report,<sup>6</sup> is needed to ensure that moderation is appropriate, well-resourced and accurate. This should include transparency about the scale and nature of user complaints relating to hateful content and the actions taken in response. It is also vital that there is greater transparency on the role of algorithms in spreading content that may be hateful, particularly during those events where we saw increased scale of hateful content.
5. **Online platforms, government and civil society organisations need to collaborate on effective campaigns to tackle the widespread, normalised use of slurs in society.** These efforts should draw from best practice understanding of behaviour change campaigns, including those that have addressed normalised slurs or casual racism successfully. They should focus on those communities where these slurs are widespread, including football fans and gamers. Campaigns must avoid appearing to be 'politically correct' in order not to be counter-productive.
6. **Online platforms, government and researchers need to pay greater attention to the intersectional nature of hateful speech.** This should include undertaking research into the experience of victims or groups who are frequently subjected to online hateful content. Online platforms should use this research to inform moderation policies and product updates with greater emphasis on safeguarding. Civil society should attempt to build coalitions to address the intersectional aspects of hate more effectively.
7. **Media organisations, government, local authorities, police and online media platforms should try to create a co-ordinated mechanism for responding to events that tend to cause spikes in hateful speech.** This could be modelled after or take inspiration from the Global Internet Forum for Countering Terrorism (GIFCT) Content Incident Protocol<sup>7</sup> and leverage the OCCI to mobilise counterspeech in the wake of such events.
8. **Greater attention should be given to the relationship between online hateful content and offline hate crimes or incidents (such as attacks).** This should include further research on these phenomena to determine if there is a correlation between them. French police statistics should also include a category of online hate crimes.

## Glossary

**Ableist discourse** Hateful forms of speech which prejudice and discriminate against 'persons with impairments and attitudinal and environmental barriers that hinder their full and effective participation in society on an equal basis with others', based on the definition of the UN Convention on the Rights of Persons with Disabilities (2006).<sup>8</sup>

**Anti-LGBTQ or homophobic discourse** The Council of Europe<sup>9</sup> describes anti-LGBTQ discourse as expressions which promote or justify homophobia or transphobia defined as:

**Homophobia** 'Irrational fear of, and aversion to, homosexuality and to lesbian, gay, bisexual persons based on prejudice'

**Transphobia** 'As an irrational fear of, and aversion to, transgender persons' gender non-conformity based on prejudice'.

**Extremism** ISD defines extremism as 'the advocacy of a system of belief that posits the superiority and dominance of one "in-group" over all "out-groups", propagating a dehumanising "othering" mind-set that is antithetical to the universal application of human rights. Extremist groups advocate, through explicit and more subtle means, a systemic change in society that reflects their world view.'

**Harassment** Repeated attitude or action which has an impact in the deterioration of condition of life which triggers an alteration of physical or mental health.<sup>10</sup>

**Hate speech** This research was based on the Council of Europe's definition of hate speech: 'all forms of expression which disseminate, incite, promote or justify racism, xenophobia, anti-Semitism or other forms of intolerance based on hate, including intolerance'.

**Hateful speech** Beyond a strictly legal understanding, hateful discourse encounters a more normalised uses of hateful slurs on the one hand, to aggressive, violent and potentially illegal hate speech on the other hand.

**Misogyny** Ging and Siapera define misogyny as hatred or fear of women, 'which may not involve violence but almost always entails some form of harm; either directly in the form of psychological, professional, reputational, or, in some cases, physical harm; or indirectly, in the sense that it makes the internet a less equal, less safe, or less inclusive space for women and girls'.<sup>11</sup>

---

**Re-appropriation (of terms)** Re-claiming of offensive terms by groups which were originally targeted, this mechanism aims to empower those who were stigmatised.

**Racism** According to the European Commission Against Racism and Intolerance (ECRI), 'racism shall mean the belief that a ground such as "race", colour, language, religion, nationality or national or ethnic origin justifies contempt for a person or a group'.<sup>12</sup>

**Sexist hate speech** According to the European Council,<sup>13</sup> sexist hate speech 'is one of the expressions of sexism, which can be defined as any supposition, belief, assertion, gesture or act that is aimed at expressing contempt towards a person, based on her or his sex or gender, or to consider that person as inferior or essentially reduced to her or his sexual dimension. Sexist hate speech includes expressions which spread, incite, promote or justify hatred based on sex. The true extent of sexist hate speech is partly hidden by the fact that many targeted women do not report it.'

**Slur** As defined by the Anti-Defamation League, 'an insulting, offensive or degrading remark, often based on an identity group such as race, ethnicity, religion, ethnic, gender/gender identity or sexual orientation'.<sup>14</sup>

---

## Introduction

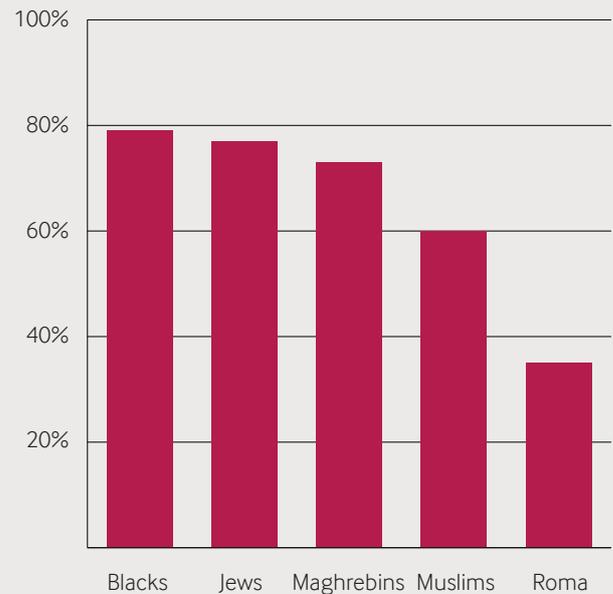
The issue of online hate, and its real world consequences, frequently makes headlines in France. Two incidents from earlier this year serve as good examples. First, on 8 February 2019, the newspaper *Libération* revealed the existence of 'Ligue du LOL', a Facebook group created by journalist Vincent Glad in which a number of male journalists abused female colleagues and co-ordinated online harassment campaigns against female journalists, activists and politicians. Then, in late March, a series of punitive anti-Roma attacks took place on the outskirts of Paris. They were sparked by false rumours of child kidnappings by the Roma population going viral on Twitter, in a case that highlighted both the persistence of prejudices against Roma people and the relationship between online communication and offline hate.

According to an OpinionWay survey published in December 2018, 59% of French citizens have experienced hateful attacks on social media at some point in their life. Research points to a persistence of prejudicial attitudes against ethnic and religious minorities in France. In April 2019, the Commission Nationale Consultative des Droits de l'Homme (CNCDH) published its 28th report on racism, anti-Semitism and xenophobia. This report, which covers the 2018 calendar year, includes a tolerance index for five minority groups: black people, Jews, Maghrebins, Muslims and Roma. The tolerance index is based on public opinion surveys carried out by Ipsos, where a low index score indicates generally low levels and a high index score generally high levels of tolerance towards a minority group.<sup>15</sup> As seen in Figure 1, Roma people and Muslims experience the lowest levels of tolerance of these five minority groups in France.

While prejudice is by no means a new phenomenon in society, social media provides a platform for hate speech and prejudice to be exposed in a way that is visible and can be measured. Social media platforms also provide an opportunity for ideologically motivated individuals and groups to spread misinformation and highly sensationalised stories in order to increase hate and antipathy towards certain sectors of society.

Increasingly, there is concern that online hate can have real world consequences. While causal links between online hate and offline attacks are difficult to establish,

**Figure 1**  
Tolerance levels towards five minority groups in France, 2018<sup>16</sup>



some research has identified correlations between them. For example, researchers from the University of Warwick found an association between support for the xenophobic Alternative for Deutschland party and anti-refugee sentiment on Facebook and violent crimes against refugees in Germany.<sup>17</sup>

Offline violence often has trails online. A recent string of far-right-inspired attacks in the US and New Zealand took shape in online platforms and emerged against a backdrop of anti-immigrant and anti-Muslim rhetoric online. A report by Amnesty International analysing social media messages directed at female public figures in the US and UK in 2017 found that there was online abuse of at least one female UK or US politician every 30 seconds on Twitter, only months after the murder of British MP Jo Cox by a far-right extremist.

Increasing public and political pressure to tackle hate speech online has culminated in the adoption of new legislation in a number of European countries. In June 2017, the German Bundestag passed the NetzDG law (or Network Enforcement Act), which legally obliges large

social media platforms to remove illegal content within 24 hours of it being reported to them. Coming into effect in October 2017, this law can be used to impose fines of up to €50m in cases of systematic breaches. In the UK, the government's Online Harms White Paper has raised concerns over the impact of online hate speech on British citizens and highlighted the need for greater transparency from social media platforms on content moderation, as well as proposing regulatory oversight for how algorithms contribute to the spread of harmful content.

Similarly, the French government has taken new steps to regulate content moderation by adopting the Loi Avia against online hate on 9 July 2019. The law is set to be discussed by the Senate in the autumn of 2019. In line with the German NetzDG law, the Loi Avia, put forward by La République en Marche (LREM) MP Laetitia Avia, aims to regulate hate speech online. The key provision of the law requires platforms to remove content that is 'manifestly illegal' (as defined under the 2004 law) within 24 hours of being notified by users. Companies which fail to meet this provision expose themselves to fines which could be equivalent to 4% of their global annual turnover. More broadly, President Macron's government has asked for more transparency and collaboration from online platforms, for instance through the reinforcement of judicial co-operation.

### **Gaps in Understanding Hate Speech Online: The Need for More Research**

The Loi Avia aims to combat online hate; however, there remain many gaps in the understanding of the nature, volume and dynamics of this 21st-century phenomenon. The parliamentary report of the Loi Avia emphasises 'the unacceptable proliferation of hateful content online'.<sup>19</sup> However, the report also raises difficulties in reaching a full understanding of this phenomenon, highlighting for instance that few rigorous studies about online hate exist. As a result, the report – which sets the context for an online regulation law in France – largely relies on European statistics<sup>20</sup> when looking at hate speech online, only referring to French statistics<sup>21</sup> for incidents which occurred in the offline space.

While the Loi Avia emphasises the lack of systematic studies of online hate, there have been several attempts to fill the gap. In May 2018 the content management

and moderation company Netino presented to the Secretary of State for Digital Affairs its Panorama de la haine en ligne, which analysed thousands of Facebook comments on large French media outlets' pages. The report showed that one in ten comments analysed could be classified as hateful. Throughout 2017, the monthly Baromètre des manifestations de la haine en ligne from IDPI (Idées, Pratiques, Innovations) analysed the amount of hateful speech on Twitter, examining different categories (sexism, xenophobia, homophobia, anti-Semitism and anti-Muslim hate). The Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA) regularly monitors hate speech online. These initiatives follow in the footsteps of similar international initiatives. Elsewhere in Europe and North America, non-profit organisations and research institutions have attempted to analyse and quantify hate speech online. The Anti-Defamation League (ADL) in the US, for instance, conducted a one-year study of anti-Semitic hate speech on Twitter.<sup>22</sup>

These attempts at mapping online hate speech have made valuable contributions to the understanding of this growing threat, but gaps remain.

First, many reports previously conducted did not include misogynistic or ableist speech – categories that are often excluded from legal definitions of hate speech.<sup>23</sup> Meanwhile, other reports included categories that did not correspond to protected categories often found in definitions of hate speech, for example 'aggressive speech', which does not target a minority.

The reports mentioned above also neglect to examine the intersectionality between different forms of hate speech. Intersectionality encourages a multi-level identity analysis (based on race, gender, sexual orientation, disability, etc.), while a single-issue framework can possibly limit the understanding and experience of victims of hate speech.<sup>24</sup> A recent study by the European Commission has acknowledged the importance of an intersectional perspective<sup>25</sup> when examining discrimination, recognising nonetheless a lack of 'data disaggregated by both sex and race, still less by other sources of intersectional discrimination, such as ethnicity and disability'.<sup>26</sup>

Finally, previous analyses tend to adopt a one-platform approach (e.g., Twitter or Facebook). This is often due

to restricted access to data from online platforms, as well as the growing number of alternative platforms that make research on hate methodologically difficult (given the diversity of sources and inconsistently indexed data). There is a need for more rigorous and comparative methods to quantify and classify hate speech online across different platforms over time, though this is severely limited by lack of data access to some online platforms.

### This Report

This report seeks to add to the research base described above and provide a systematic understanding of hateful speech on online platforms in France that are open to big data analysis. Our aim is for this research to act as a resource for French civil society organisations challenging hate and extremism. These organisations are at the core of ISD and Facebook's OCCI, and fight on the front lines against hateful and divisive speech online and offline. This report will also provide important information to policymakers as they shape policy responses to hate speech and polarisation in the digital age. The authors also hope to help tech companies improve protections against online hate for targeted communities.

These are two key research questions that this report seeks to answer:

- What is the scale of hate as part of online discourse in France looking across a wide range of potential groups of people who may face hateful speech, especially relating to ethnicity, religion, gender and sexuality and disability?
- What are the characteristics of online hateful speech in France, including key themes, influencers and online or offline trigger events?

While the report sheds some light on the scale, nature and targets of hateful speech online, adding to existing research, there remain a number of obstacles to obtaining a true understanding of hateful speech online. This is largely due to lack of data access. Twitter remains the platform that is most open to research and thus remains the dominant platform for these types of analyses. While the research sought to analyse hateful speech on Facebook and other platforms and forums, the lack of API access prevents full and comprehensive analysis.

### Defining Hateful Speech

The concept of 'hate speech' is notoriously difficult to define. In France, the French legal framework for hate speech was first set by the freedom of press law of 29 July 1881,<sup>27</sup> completed by the 1972 Pleven law,<sup>28</sup> which condemns 'insults, defamation and provocations encouraging discrimination hate and violence towards a person or a group because of their national origin or their ethnicity, nation, race or religion'. This legal definition was later amended by the 21 June 2004 law, which added 'gender, sexual orientation, gender identity and disability'.<sup>29</sup> The key issue with this definition from a legal standpoint has been the understanding of terms such as 'insults',<sup>30</sup> 'defamation'<sup>31</sup> and 'provocations encouraging discrimination'.

Judges' interpretation of these three terms has been inconsistent<sup>32</sup> and experts have emphasised contradictions in French jurisprudence.<sup>33</sup> For instance, in a case where Michel Houellebecq was tried for declaring, 'the stupidest religion, it's Islam',<sup>34</sup> the defendant was found not guilty as the courts considered the language to express opposition to an ideology and belief system. On the other hand, comedian Dieudonné was found guilty of hate speech for saying, 'to me Jews are a sect, a fraud',<sup>35</sup> which was considered by the courts to be an injury against a protected category of people based on their origin.

It was not our aim in this report to identify illegal hate speech or to determine whether certain instances of online hate speech in France was legal or illegal. Our aim was instead to identify and analyse the wide range of contexts in which 'hateful discourse' may or may not take place online: from the more normalised uses of hateful slurs on the one hand to aggressive,

“ There remain a number of obstacles to obtaining a true understanding of hateful speech online ”

# 53%

Of insults or aggressive comments online are made against other internet users\*

\*Source: Netino

violent and potentially illegal hate speech on the other. In order to inform the definition of 'hateful discourse' used in this report, we undertook a review of existing research and studies that sought to identify and analyse hate speech online.

As mentioned above,

Netino examined 15,000 comments randomly selected among 15 million comments on the public Facebook pages of 25 established media outlets to identify hateful trends online.<sup>36</sup> In the first quarter of 2019, they identified four main forms of hateful speech:

- insults or aggressive comments<sup>37</sup> against other internet users (comprising 53.1% of hateful content);
- attacks against politicians (30.1%);
- attacks against personalities (15.5%);
- attacks against the media or journalists (15.1%).

Overall, in the first quarter of 2019, 14.3% of randomly selected posts were found to be aggressive or hateful.

The think tank IDPI, which published the Monthly Barometer of Manifestations of Hate Online cited above, identifies five categories: anti-hate, neutral, ordinary racism, hateful speech and 'hijacking' (co-opting hashtags or other online campaigns to spread hateful messages). The Barometer includes a randomly selected sample of tweets coded manually into these five categories.<sup>38</sup> The most recently published Barometer, from January 2018, categorised 24% of the sample as ordinary racism and 11% as explicitly hateful, based on a sample of 2,950 tweets.<sup>39</sup> The former is defined as 'an expression based on implicit hate (such as stereotypes and prejudices), corresponding to an intentionally hateful or insulting expression', whereas the latter is 'an intentionally hateful or insulting expression'. The Barometer looks at five types of hate in particular: homophobia, anti-Semitism, sexism, xenophobia and anti-Muslim hate.

A 2018 report from the Quaker Council for European Affairs on anti-migrant hate speech focused on the

sharper end of hate speech.<sup>40</sup> While the report based its definition of hate speech on the Council of Europe's, it did not necessarily consider 'comments... which were unpleasant, injurious, or anti-migrant' as hate speech. Instead, it established two main criteria to categorise content as hate speech: calls for violence and dehumanisation.

On the other end of the spectrum, the ADL's Online Hate Index takes a much broader definition than typical legal definitions of hate speech.<sup>41</sup> It defines hate speech as

"Comments containing speech aimed to terrorize, express prejudice and contempt towards, humiliate, degrade, abuse, threaten, ridicule, demean, and discriminate based on race, ethnicity, religion, sexual orientation, national origin, or gender... Also including pejoratives and group-based insults, that sometimes comprise brief group epithets consisting of short, usually negative labels or lengthy narratives about an out group's alleged negative behaviour."<sup>42</sup>

The ADL's approach also established several labels by which to classify hateful posts, including insult, profanity, conspiracy theory, sarcasm and threat.

Similarly, the British think tank Demos established different categories to classify misogynistic hate: serious, non-offensive, colloquial, casual, generally misogynistic, abusive and other (which includes ambiguous categories such as subversive and pornographic content).<sup>43</sup> In another study on anti-Muslim content on Twitter, Demos separated hateful content into three categories: insults, derogatory statements which linked Muslims or Islam to terrorism, and statements which more broadly claimed Muslims are socially and culturally destroying the West.<sup>44</sup>

### Our Approach

These different approaches highlight the variety of definitions that can be adopted in the study of online hateful speech. For this report, we chose the Council of Europe's definition of hate speech as 'cover[ing] all forms of expressions that spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance'. However, classifying content according to this definition raised a number of challenges, as the boundary between

normalised slurs and hateful speech proved to be difficult to establish in many cases.

As the types of discourses of interest in this report were not simply illegal hate speech, but also broader hateful discourse, this definition was interpreted to include normalised slurs, given how these can contribute to social division and polarisation (similar to 'ordinary racism' as defined by IDPI and 'colloquial', 'casual' and 'generally misogynistic' from the Demos research). Therefore uses of slurs that have been normalised, but are founded on attacks of protected categories (e.g., pute [bitch] and pédé [fag]), have been included in hateful subsets. While individuals who use these terms may not have the intention of spreading, inciting, promoting or justifying hate, the terms themselves are founded on prejudices which have the impact of normalising these types of hatred. Where possible, this report distinguishes between this type of normalised hateful discourse and more targeted hateful speech.

Finally, while informing our own approach adopted in this report, there is an important difference between our approach and those of previous studies. Previous studies were able to develop more sophisticated categories because they largely relied on qualitative analysis of small samples of social media content. The largest sample among the studies cited above was 15,000 posts from Twitter in the Netino study. By contrast, our aim was to use natural language processing and machine learning algorithms to analyse much larger datasets of online content with a view to providing a more complete and comprehensive picture of online hateful discourse.

While our analysis was able to provide some more fine-grained insights into the different types of 'hateful' speech identified, next steps for this research should focus on investigating the extent to which natural language processing and machine learning can be developed to parse 'hateful' datasets into more detailed categories. In time, the aim would be to build a system that is capable of confidently identifying hateful speech, analysing it in real time and then providing insights directly to NGOs who work to support and protect victims of online hate speech and produce counterspeech.

“ The types of discourses in this report were not simply illegal hate speech, but also broader hateful discourse ”

---

## Methodology

As stated earlier, the research questions to which this report sought to respond were:

- What is the scale of hate as part of online discourse in France looking across a wide range of potential groups of people who may face hateful speech, especially relating to ethnicity, religion, gender and sexuality and disability?
- What are the characteristics of online hateful speech in France, including key themes, influencers and online or offline trigger events?

We also sought to understand the extent to which natural language processing and machine learning can be used to identify hateful discourse confidently at speed and scale.

In order to answer our research questions, we adopted a mixed-methods approach, combining qualitative analysis of large datasets with natural language processing techniques to identify and analyse hateful speech online.

### Data Collection

We began by attempting to identify a comprehensive list of keywords to capture hateful speech categorised by the group or minority they target. Selection of the different categories of hate speech was informed by the Council of Europe's definition of hate speech, the Délégation Interministérielle à la Lutte Contre le Racisme, l'Antisémitisme et la Haine anti-LGBTQ (DILCRAH; Inter-ministerial Delegation on the Fight Against Racism, Anti-Semitism and Anti-LGBTQ Hate), as well as the protected categories covered by the terms of service of Facebook and Twitter. Consultation with members of the OCCI Steering Committee in France and our desire to include categories often under-represented in studies of online hate speech informed some of the categories in this report.

The types of hateful speech and groups we identified included:

- ableist discourse
- anti-Arab or anti-Maghrebin discourse
- anti-Asian discourse
- anti-black or anti-African discourse
- anti-Christian discourse<sup>45</sup>
- anti-LGBTQ discourse
- misogynistic discourse
- anti-Muslim discourse
- anti-Roma or anti-gypsy discourse
- anti-white discourse.<sup>46</sup>

In order to identify potential instances of these kinds of discourse, we developed lists of relevant keywords that are often used in a hateful way or in conjunction with hateful speech. These lists are based on previous research by ISD as well as consultation with a number of anti-hate organisations working in France, including LICRA, Le Refuge, the AJC, its initiative #JeSuisLà and La Voix des Roms. The full lists of keywords can be found in Appendix 1.

We queried two commercial social listening tools using these keywords to build our initial datasets. Posts containing at least one of the keywords in our lists were then included in our initial datasets. The social listening tools are:

- **Crimson Hexagon**, a social listening platform that aggregates data from a number of public sources, including Twitter, YouTube comment sections, Reddit and other forums and blogs
- **CrowdTangle**, a tool that aggregates Facebook data from public pages and groups (only posts made by these pages and groups, no user-level data).

The breakdown of sources for all datasets is displayed in Figure 2. Owing to data access restrictions that vary from platform to platform, the sample is skewed heavily towards Twitter, which has the most open access. Other platforms either restrict access or are difficult to index, leading to their under-representation in this report. While we tried to include as many platforms as possible, data restriction still poses a barrier to this type of research. See 'Limitations' below and 'Recommendations' later in this report for further discussion of this.

**Figure 2 Breakdown of sources of data by platform**

Platform	%
Blogs	1%
Facebook	1%
Forums	4%
Reddit	0%
Tumblr	1%
Twitter	89%
YouTube	4%
Other	1%

**Table 1 Initial datasets identified using at least one hateful keyword before running relevancy and hateful classifiers**

Discourse	Posts
Misogynistic discourse	7,948,332
Anti-Arab/anti-Maghrebin discourse	1,752,405
Anti-LGBTQ discourse	1,695,268
Ableist discourse	565,662
Anti-Roma or anti-gypsy discourse	299,157
Anti-black or anti-African discourse	223,483
Anti-Muslim discourse	168,324
Anti-Christian discourse	156,047
Anti-white discourse	125,654
Anti-Semitic discourse	79,289
Anti-Asian discourse	45,804

In the analysis below, we do not break down the hateful datasets by platform because the perception of a problem or lack of one on a specific platform will be skewed by the differing levels of data access across social media platforms. In other words, because Twitter provides the most open access to data for research, it may appear as if they have the biggest scale of hateful speech, when in fact that may not be the case.

The date range for each query was 1 January 2019 to 31 May 2019, and queries were restricted to French-language posts geo-located to France only. The total size of each of these initial datasets is shown in Table 1.

Commercial software tools like Crimson Hexagon and CrowdTangle provide in-built analytics including volume over time metrics, and surface other terms and content frequently associated with keywords. They also identify users most frequently using the keywords. However, they do not provide the ability to test these datasets to ensure that the content they are returning is in fact relevant to hateful speech. With a topic as sensitive as hateful speech, it is vital that researchers can test the data further to ensure relevancy and accuracy. This is particularly important given the very large initial datasets that are returned for discourse that contained keywords that could be misogynistic, anti-Arab or anti-LGBTQ, in order to ensure that all of the content identified by keywords is in fact hateful speech.

To address some of these issues with commercial analytic tools, ISD partnered with CASM, to leverage its proprietary tool for natural language processing, Method52. This tool allowed us to train machine learning algorithms to recognise specific types of speech or to organise very large datasets to highlight patterns.

First, in order to ensure we did not exclude any important keywords, we analysed all of our datasets with a 'surprising phrase detector' in Method52. This compared the gathered datasets of hateful activity with a reference corpus of general French-language text, in order to identify any words or phrases that occurred more frequently in our sample than in average online discussion. These surprising phrases were manually appraised to determine that no additions to our keyword lists were necessary.

Because of time and resource constraints, we selected the four largest datasets (misogynistic discourse, anti-Arab or anti-Maghrebin discourse, anti-LGBTQ discourse and ableist discourse) for further quantitative analysis using Method52 in order to confirm the relevancy of the content in the datasets, and identify and analyse specifically 'hateful' content. Being the largest datasets, the requirement for automated analytical methods rather than dip-sampled or manual approaches was considered to be greater. Next iterations of this research should focus on similar in-depth analysis for the remaining datasets.

### Analysis of the Four Largest Datasets

We further analysed the four largest datasets following a four-step process.

#### Step 1

We trained a natural language processing algorithm to separate relevant from irrelevant posts within the four datasets. Irrelevant posts were defined as cases in which the use of the keywords did not refer to the protected category concerned. Each of these algorithms were trained to be at least 85% accurate at identifying relevant posts. For more information on algorithm training, see Appendix 2. The number of relevant posts from the four largest datasets (using the natural language processing algorithm) are presented in Table 2.

**Table 2** Number of posts used in analysis following relevancy classifier

Dataset	Relevant posts	%
Anti-women or misogynist	6,669,862	84%
Anti-Arab or anti-Maghrebin	1,003,668	57%
Anti-LGBTQ or homophobic	1,705,196	99%
Ableist or anti-disability	344,078	61%

These results show that the keyword lists returned varying levels of relevant posts. For example, anti-LGBTQ and misogynistic keyword lists returned a majority of relevant posts. Within the ableist dataset, terms like mongolien frequently surfaced posts related to Mongolia, which were categorised as irrelevant. In the anti-Arab dataset, a majority of the irrelevant posts

“ We selected the four largest datasets for further quantitative analysis using Method52 in order to confirm the relevancy of the content ”

were cookery-related, given the inclusion use of beur as a keyword (which returned posts including the word beurre [butter]). In the LGBTQ section, the keyword pédale produced some irrelevant content, as it is a slur that refers to gay men whose original meaning is pedal. In the misogyny section, examples of irrelevant posts included the use of terms which are derived from misogynistic slurs but do not refer to people, such as saloperie (filth or junk).

#### Step 2

We trained a second algorithm to identify hateful speech within the remaining data. To determine what speech was hateful, we referred to the Council of Europe's definition of hate speech. As discussed above, this definition was interpreted to include slurs defined by the ADL as 'an insulting, offensive or degrading remark, often based on an identity group such as race, ethnicity, religion, ethnic, gender/gender identity or sexual orientation'. All of these algorithms were trained to be at least 85%<sup>47</sup> accurate at identifying hateful posts (Table 3).

**Table 3** Number of posts from the 'relevant' datasets that were identified as 'hateful speech'

Dataset	Hateful posts	%*
Anti-women or misogynist	5,453,603	82%
Anti-Arab or anti-Maghrebin	131,731	13%
Anti-LGBTQ or homophobic	1,007,034	59%
Ableist or anti-disability	265,126	77%

\*Percentage of relevant data set

### Step 3

We trained a third algorithm to provide greater granularity to each of the subsets of hateful posts. It aimed to separate hateful speech into the following categories:

- **generalised hateful slurs**, referring to messages in which slurs were used to belittle or demean a person or group of persons (including normalised uses of terms like pute and pédé)
- **targeted hateful speech**, referring to messages in which a user who appeared to be a member of a protected category was directly attacked on the basis of that category in a personalised message or where hateful language was used to attack a person or group of persons on the basis of (perceived) membership in that category
- **other hateful speech**, which included hateful speech in unclear contexts (e.g., use of hateful language by pornographic accounts and quotes of hateful speech)
- **counterspeech**, including speech that was both part of co-ordinated campaigns and in the form of organic push-back against the use of hate or hateful terms.

### Step 4

We conducted a community-based analysis on the four largest datasets to identify distinct networks of users, pages or groups that use hateful speech (as determined by the classification in Step 2). The analysis used clustering to identify communities that employ specific hateful narratives, based around terms that were most frequently used in the datasets. This allowed for more granular analysis of the types of hateful speech present online and the different centres of discussion. The findings of this analysis are presented in network graphs of key terms in the sections below.

#### Analysis of Accounts that Appeared in More than One Dataset

Method52 was also used to identify accounts that appeared in more than one of our datasets – accounts that used multiple types of hateful speech. This allowed us to examine instances where multiple types of hateful speech were used by the same accounts and where different types of hateful speech intersect.

For further details on the Method52 and the classification process, see the Appendix 2.

#### Analysis of Smaller Datasets

The smaller datasets were analysed using Crimson Hexagon and CrowdTangle. While commercially available tools such as Crimson Hexagon present limitations, it nonetheless allowed us to draw some interesting observations.<sup>48</sup> Three ISD researchers analysed data samples with the different functionalities offered by these platforms. For each category of discourse, ISD's researchers identified spikes in activity and examined which content and keywords were most often mentioned during the relevant timeframe to determine which event had driven an increase in conversation. Using the tool to identify the most shared pieces of content allowed us to identify key themes and trends in the data sample. Crimson Hexagon and CrowdTangle also made it possible to identify the most active accounts that mentioned the keywords from our lists. This enabled us to gauge to what extent keywords were predominantly used in a hateful context by active users.

#### Limitations

While this report aims to be as comprehensive as possible, there are a number of limitations to this type of study.

First is the issue of data access. Our datasets have been drawn from publicly available data from the main social media platforms (e.g., Facebook, Twitter, YouTube) and other sources indexed by the social listening tools with which we work. However, there exist many private, closed and encrypted channels to which we do not have consistent or computational access. Therefore, this report, its findings and recommendations have been limited to data that is publicly available and consistently indexed. It does not speak to the wider ecosystem of private channels and smaller sites that may also host hateful content.

Second, we gathered data using a keyword-based approach. Given the dynamic nature of language and of online language in particular, it is possible that our datasets have missed some hateful speech that does not contain the terms in our keyword lists. Conversely, as demonstrated by the relevancy analysis discussed above, keyword-based approaches can also draw in irrelevant data. Additionally, hateful speech targeting

---

specific groups can differ significantly, not just in the terminology but also in the number of terms and/or slurs that exist in society to refer to a particular group. This can lead to inequity in the number of keywords used to collect data on each type of speech, which can influence the number of results returned and the depth of the data. While we took multiple measures to minimise the impact of these limitations (by consulting colleagues in other expert organisations and employing surprising phrase detection analysis, as detailed above), some omissions may persist.

Third, the social listening tools with which we work remove from their databases posts that have been deleted from online platforms for violating community guidelines, terms of service or local laws. Therefore, our datasets will likely not include some of the most heinous posts that may have been removed by the platforms before or during data collection. This may explain, for example, the relatively small amount of anti-Semitic content in our datasets despite a spike in offline hate crime in France, according to official statistics.<sup>49</sup>

Finally, as mentioned above, our algorithms were trained to be at least 85%<sup>50</sup> accurate at identifying hateful posts. Natural language processing and machine learning are not an exact science and – much like humans – there will always be a degree of bias and error in the decisions the algorithms make. An overall accuracy of 85% is broadly consonant with the equivalent performances presented in recent peer reviewed work in the same area.<sup>51</sup> However, as with all machine learning, it is impossible to be 100% accurate in recognising the multi-faceted nuances that exist in natural language, particularly in an area that is as fraught and context-specific as hateful speech.<sup>52</sup> Indeed, even human review is not 100% accurate. This represents another limitation and challenge to this study, and to automated content moderation in general, as will be discussed later in this report. The confidence scores produced by Method52 provide this report with greater transparency and confidence than other, more opaque, machine learning algorithms (see Appendix 2).

“ Our algorithms were trained to be at least 85% accurate at identifying hateful posts ”

## Analyses of Discourses

This section presents the analyses conducted on the 11 different datasets described in the previous section. These are divided into four primary categories:

- gender and sexuality
- race and ethnicity
- religion
- disability.

Analysis of each of these discourses includes key findings from our analysis, including:

- the scale of 'hateful' content or discourse containing 'hateful terminology'
- events or drivers of spikes in conversation
- key themes
- analysis of the ten most active accounts.

Some categories of hateful discourse delivered surprisingly low results, which could be the result of a number of factors. These include, but are not limited to, content moderation by social media platforms, limitations or biases in the keywords used and limited access to data on the different platforms, as highlighted in the methodology section.

### Gender and Sexuality

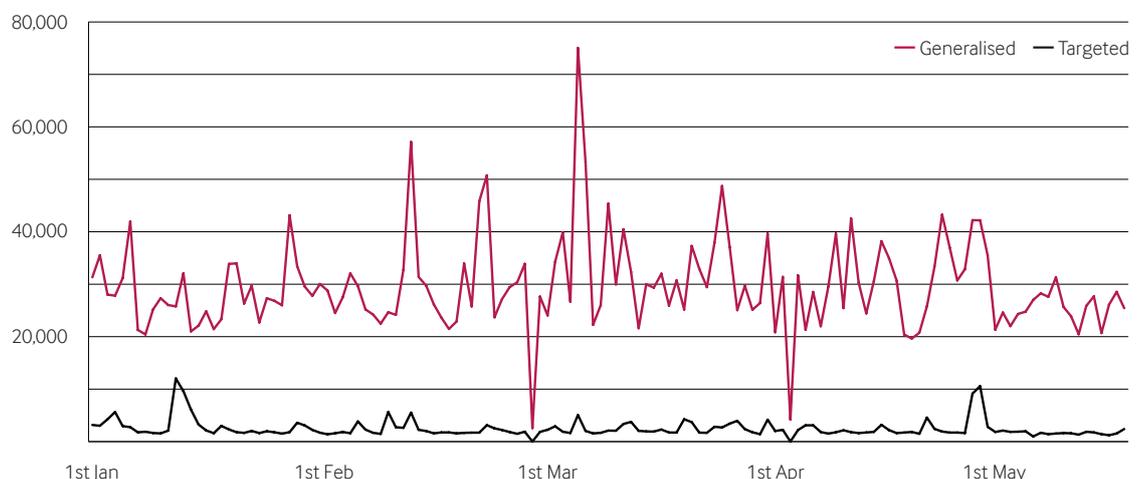
This section includes misogynistic discourse and anti-LGBTQ discourse. These two datasets were characterised by their high volume and the normalised usage of misogynistic and homophobic slurs in diverse contexts, especially terms like pute and pédé. Hateful speech from these two categories also showed substantial overlap, indicating a general conflation of gender and sexuality.

## Misogynistic Discourse

### Key Findings

- Out of the 6.6 million relevant posts identified, our algorithm classified **5.5 million posts as hateful misogynistic speech** (Figure 3).
- This included **4 million posts containing generalised misogynistic slurs, and around 300,000 posts that were targeted hateful attacks** against users who appeared to be women. Over 1 million posts in the 'other' category of hateful speech predominantly related to accounts that were sharing pornographic content.
- An algorithm trained to recognise counterspeech identified **100,000 posts (only 1% of relevant posts) containing both organic and co-ordinated counterspeech**.
- Hateful misogynistic speech was characterised by its consistent volume over time in comparison with the other discourses analysed in this report. **The volume of hateful misogynistic speech exceeded 200,000 tweets almost every week.**
- **International Women's Day** in March corresponded with the largest spike in generalised hateful speech. The two smaller spikes in targeted hateful speech were caused by misogynistic attacks on individuals, which have been deleted by the platforms.
- **Generalised misogynistic slurs** were frequently used to express anti-government views; for example, attacking Macron with terms like *fil de pute*. Attacks against female politicians and influencers was a key trend. This is particularly the case for Marlène Schiappa, the French Secretary of Equality, who attracts substantial abuse online.
- **Most active accounts and pages** in propagating hateful misogynistic speech included five bot or semi-automated accounts with general or one-issue interests, such as football, three pro-Yellow Vest accounts and some individual users.
- **Accounts most frequently mentioned in hateful misogynistic speech belonged to political and sports personalities** – many of them male, though some belonged to private individuals who may have been victims of harassment.
- Misogynistic content showed overlaps with other types of hateful content, including anti-Arab content and anti-LGBTQ content.

**Figure 3**  
Posts containing hateful misogynistic discourse between 1 January 2019 and 31 May 2019



## Misogynistic Discourse Examples of hateful speech



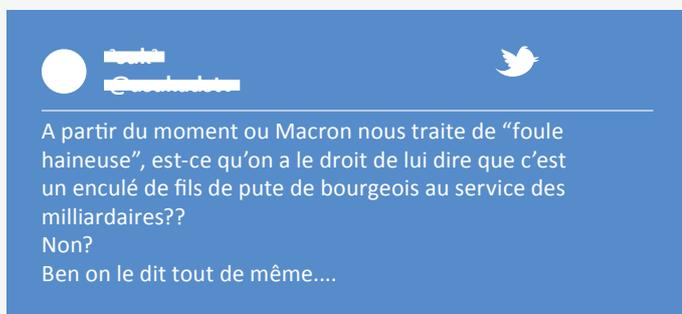
**Figure 4** Hateful content posted on International Women’s day (source: Twitter)

**Translation** ‘Woman’s day? Hmm, slut’s day, I’d rather say’



**Figure 5** Generalised use of misogynistic slurs (source: Twitter)

**Translation** ‘For fuck’s sake, make this fat bitch shut up’



**Figure 6** Misogynistic language used against Macron (source: Twitter)

**Translation** ‘If Macron calls us a “hateful mob” can we say he is a bourgeois son of a bitch working for billionaires? No? We’ll say it anyway’



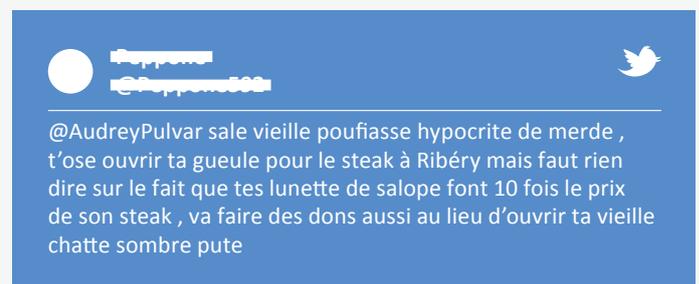
**Figure 7** Misogynistic abuse targeting journalist Audrey Pulvar (source: Twitter)

**Translation** ‘hey dirty slut, I’m talking to you @AudreyPulvar’



**Figure 8** A post targeting Marlène Schiappa, which includes violent threats (source: Twitter)

**Translation** ‘You’re giving money to whoever you like you fat bitch, but when you get a bullet in your head, will it be against Republican values or not????’



**Figure 9** Misogynistic abuse targeting Audrey Pulvar after she called on footballer Frank Ribéry to give his money to charitable causes (source: Twitter)

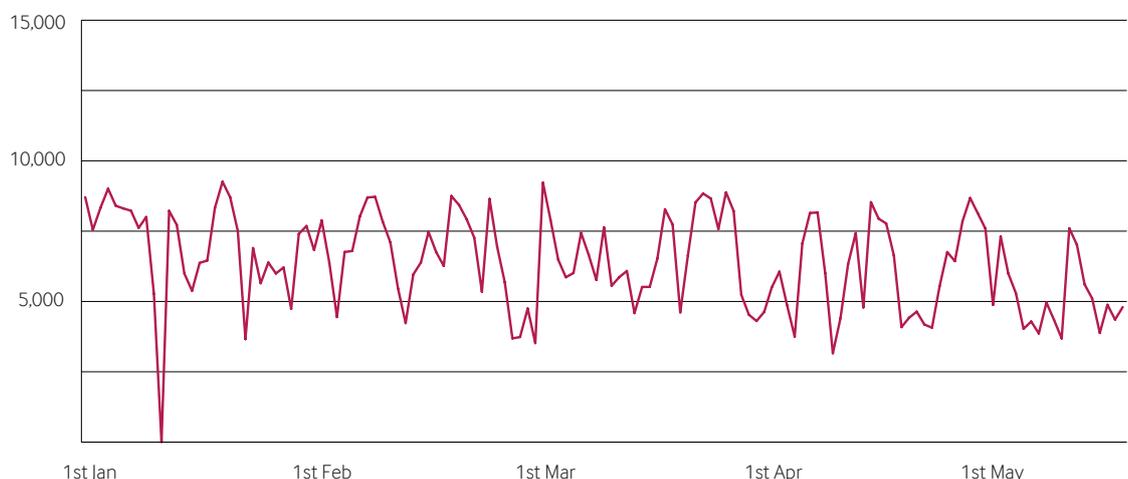
**Translation** ‘@AudreyPulvar, you dirty old skank, fucking hypocrite, you dare open your mouth about Ribéry but no one can say a word about your slutty glasses which cost ten times the amount of his steak, go and give your money to charity instead of opening your old pussy, you fucking slut’

## Anti-LGBTQ Discourse

### Key Findings

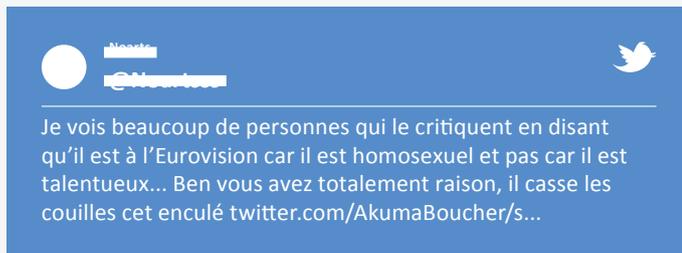
- Out of the 1.7 million relevant posts identified, our algorithm classified **1 million posts as hateful anti-LGBTQ** speech during the period studied (Figure 10).
- **Almost all hateful anti-LGBTQ speech is used in a generalised fashion.** It was not possible to train an algorithm to identify targeted hateful attacks confidently because of the relatively low volume in the sample.
- **Hateful homophobic and transphobic speech have become normalised in many contexts.** The abbreviated form of pédé (pd), itself an abbreviated form of pédéraste (referring to homosexuality and derived from the same root as pederasty), was associated with the highest proportion of hateful speech.
- We identified a **small subset of accounts re-appropriating homophobic slurs**, which could be considered a form of organic counterspeech.
- A key focus of hateful anti-LGBTQ speech over the period was **Bilal Hassani, the French entrant to the Eurovision Song Contest.**
- **The most active accounts and pages** included six exhibiting bot-like behaviour. Four of these accounts had 'bot' in their user name and posted high volumes of content referencing video games and using anti-LGBTQ slurs in their posts. The remaining four were individual users' accounts, two of which had been suspended. The most active pages on Facebook all appeared to be either official accounts of media platforms or newspapers, or general interest pages that did not seem to be using anti-LGBTQ keywords in a hateful manner. This underscores the challenge of identifying hateful pages using a keyword-based approach on commercial analytic software.
- **French political figures** were the most common targets of hateful anti-LGBTQ speech.

**Figure 10**  
Posts containing hateful anti-LGBTQ discourse between 1 January 2019 and 31 May 2019



## Anti-LGBT Discourse

### Examples of hateful speech



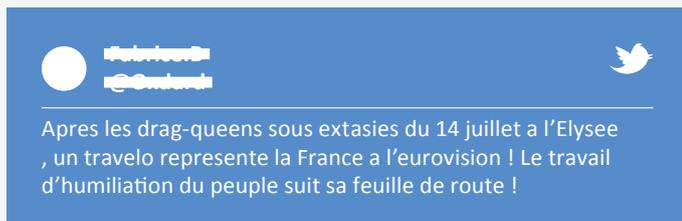
**Figure 11** A hateful tweet targeting Bilal Hassani (source: Twitter)

**Translation** 'I see a lot of people criticising him because he is gay, and not because is talented... Well, you are completely right, this fag is breaking our balls'



**Figure 12** Post demonstrating the normalisation of anti-LGBTQ slurs in online conversations as well as the intersectionality of anti-LGBTQ and misogynistic discourse (source: Facebook)

**Translation** 'Please tell me, today is International Women's Day, are fags concerned?'



**Figure 13** Hateful discourse targeting Hassani (source: Twitter)

**Translation** 'After drag queens on ecstasy on the 14 July at the Elysée [French presidential palace], a tyranny is representing France at the Eurovision! The humiliation of the French population is on its way'

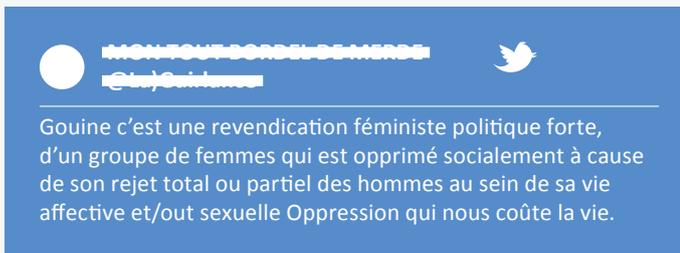


**Figure 14** A tweet targeting Hassani (source: Twitter)

**Translation** '#BilalHassani is openly taunting the French. His princess dress and stupid song are insults to France, and our 2000-year-old culture. He is the singer of the #LGBTQ and... our Minister of Culture's pet! Shame!!! #Eurovision2019'

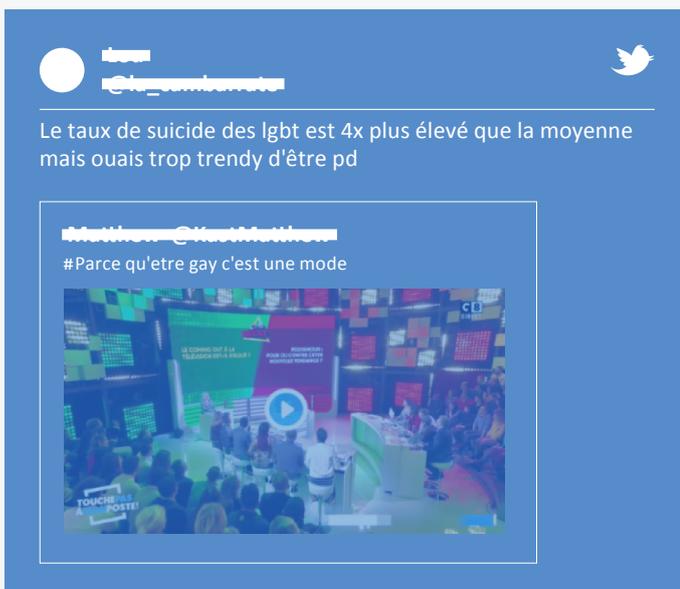
## Anti-LGBT Discourse

### Examples of Counterspeech



**Figure 15** Tweet re-appropriating the word gouine (equivalent to dyke in English, source: Twitter)

**Translation** 'Dyke is a strong feminist political statement by a group of women who are socially oppressed because of their partial or total rejection of men from their sentimental or sex life. This oppression is costing us our lives'



**Figure 16** Tweet condemning an initial Twitter post stating that 'being gay is trendy' and sharing a picture of TV show Touche pas à mon poste!, presented by Cyril Hanouna, known for his controversial statements about the LGBTQ community

**Translation** 'LGBTQ people's suicide rates are four times higher than average, but yeah, it's super trendy to be a fag'; 'because being gay is trendy'



**Figure 17** Counterspeech message by SOS Homophobie, tweeted following a transphobic attack in Paris on 2 April 2019 (source: Twitter)

**Translation** 'Can we please praise the courage of #Julia, the victim of an attack, who is breaking the silence about #transphobic violence, exposing the truth and advocating for a more open and inclusive society. We fully support you. #transphobia #LGBTphobia'

## Case Study



 **RayneGarcia**  
@RayneGarcia

Depuis 2014, en 5 évènements, la seule personne ayant gagné pour sa musique date de 2015. LE HASARD On l'aime pas, non pas parce qu'il est gay, mais parce qu'il se comporte comme une pute Oublie pas de me bloquer car tu n'as pas d'arguments

 **Rayne**  
@RayneGarcia

mdr "queen Hassani" quel fiotte

**Figure 18** Abusive tweet targeting Hassani (source: Twitter)  
Translation 'lol "queen Hassani" what a poof'

 **ROBESON**  
@Robeson

Le pd d'hassani est une humiliation pour notre peuple

 **Hassani@Eurovision**  
Une pensée pour tout les fdp de lgbt ,la maladie ne triomphera jamais #Eurovision

**Figure 19** Abusive tweet targeting Hassani (source: Twitter)  
Translation 'This Hassani fag is a humiliation for our nation'

## Attacks on Eurovision candidate Bilal Hassani

The abuse faced by 2019 Eurovision candidate Bilal Hassani emerged as a key theme in our analysis, with a substantial amount of hateful speech directed at him. Two small peaks in hateful discussion during the first half of the year both related to Hassani: the announcement of his nomination to the competition in January and the announcement of his new album in March. The video of the song with which he competed, *Roi*, focused on acceptance and empowerment of homosexual and gender-diverse individuals, which attracted significant abuse.

Hassani's Twitter account featured in the top ten most mentioned in posts including our list of anti-LGBTQ slurs. The tweets below show the types of harassment that the singer received, including slurs such as *enculé* (fucker, derived from slang for sodomise), *travelo* (slang for transvestite) and *fiotte* (derogatory term for homosexual). His name was also significant within the hateful anti-LGBTQ dataset, as can be seen in figures 11, 13, 14, 18 and 19.

## Race and Ethnicity

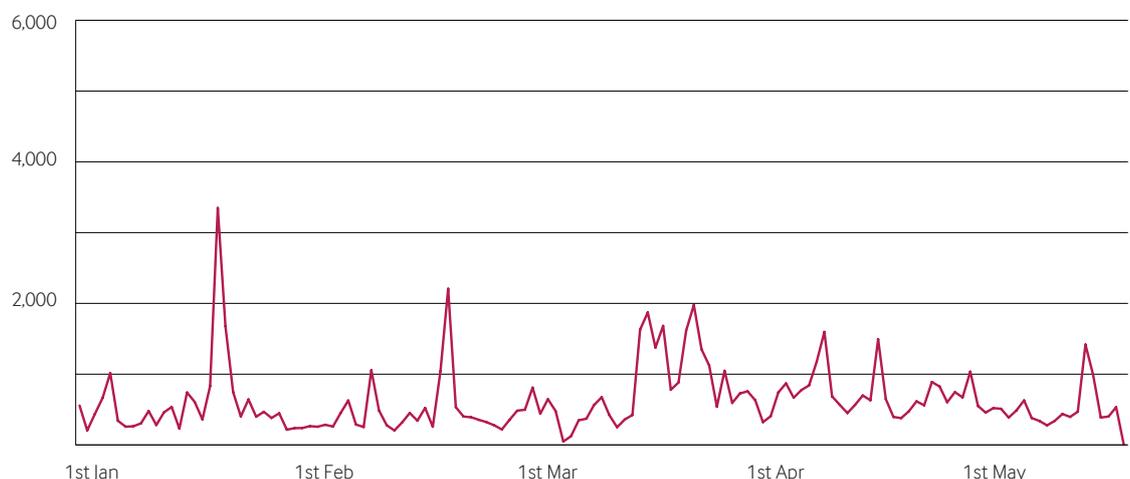
This section includes anti-Arab, anti-Roma, anti-black, anti-white and anti-Asian discourses. We used natural language processing to analyse the anti-Arab dataset because of the size of the initial dataset generated through keywords. This dataset exhibited significant overlap with anti-Muslim discourse, as will be discussed further in subsequent sections. Both anti-Roma and anti-black datasets included a significant portion of accounts fighting stereotypes and hateful speech. The anti-white dataset was dominated by the debate over whether anti-white racism exists. Anti-Asian keywords returned some debate around anti-Asian racism, though with little hateful speech.

## Anti-Arab Discourse

### Key Findings

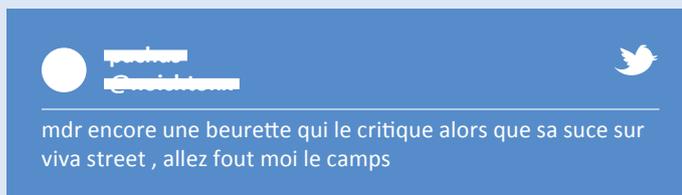
- Out of the 1 million relevant posts identified, **our algorithm classified 132,000 posts as hateful anti-Arab speech** during the period studied (Figure 20). The diverse ways in which anti-Arab keywords were used made it difficult to train an algorithm to be accurate. This subset is certainly not representative of the scale of hateful content within the dataset. The size of the dataset did not allow for further classification of these hateful posts to differentiate between targeted and generalised hateful speech.
- **The ways in which people online used anti-Arab keywords were diverse, and at times account holders used generic insults to attack Arabs, while others used anti-Arab slurs to target other groups.** The latter situation was encountered with the word *racaille*, which has a long tradition of being used by far-right and nationalist groups to refer to people of Arab origin. During the period of study, however, this term was often applied to the Yellow Vests and the police.
- **Offline events, including the attack in Christchurch, led to an increase in hateful anti-Arab speech, driven particularly by posts mentioning the Great Replacement theory;** however, non-hateful posts containing anti-Arab keywords also spiked following the attack, many of which opposed the idea of a 'great replacement' and those who propagate such views. Our algorithm classified 6% of the discussion in the week following the Christchurch attack as hateful.
- **The most active accounts in propagating hateful anti-Arab speech included some from known far-right accounts and those affiliated with Identitarian groups.** The Great Replacement theory and anti-migrant sentiment emerged as recurring themes in the hateful dataset.
- **The most mentioned accounts also belonged to far-right figures** who were frequently mentioned in posts containing anti-Arab hateful language.
- There was significant overlap with misogynistic language, most clearly seen in the use of the term  *beurette*, the feminine form of  *beur*  (a slang term used to refer to people of North African descent, or Arabs more generally). This term was used in roughly 9% of the posts in the hateful anti-Arab dataset. We also identified some efforts to re-appropriate the term.

**Figure 20**  
Posts containing hateful anti-Arab discourse between 1 January 2019 and 31 May 2019



## Anti-Arab Discourse

### Examples of hate speech



**Figure 21** A tweet conflating anti-Arab and misogynistic abuse (source: Twitter)

**Translation** 'lol, another beurette criticising him, although she is sucking dick on viva street, fuck off'



**Figure 22** A tweet using the word racaille (source: Twitter)

**Translation** : 'while UNEF [left-wing student trade union] is calling for the massacre of white people and the banlieues are ruled by Islamist riff-raff, the power in place is dissolving an organisation whose goal was to help French homeless people #BastionSocial'

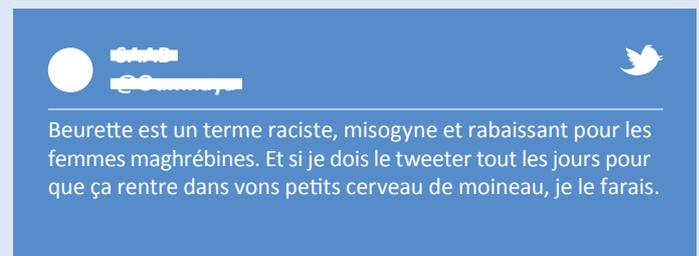


**Figure 23** Tweet by far-right platform conflating Arab women with Islamists (source: Twitter);

**Translation** 'the beurette Bouchera Azzouz, or the unveiled Islamist'

## Anti-Arab Discourse

### Examples of Counterspeech



**Figure 24** Tweet condemning the degrading use of beurette (source: Twitter)

**Translation** ' beurette is a racist, misogynistic and degrading term for women from the Maghreb. And if I need to tweet this everyday so that it fits in your little heads, I will'

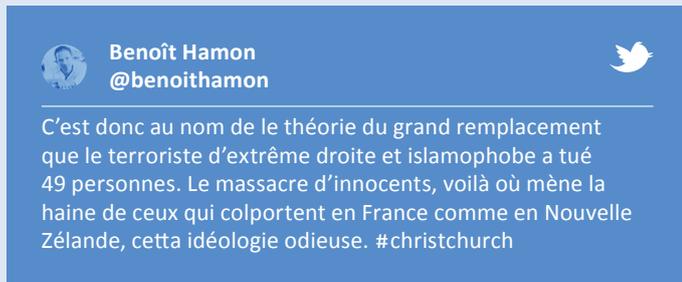


**Figure 25** Tweet condemning the Christchurch attack and the Great Replacement theory (source: Twitter)

**Translation** 'The "great replacement" theorised by Renaud Camus, turned into fiction by Michel Houellebecq and promoted in the media by Eric Zemmour, is not an opinion that should be debated, but a murderous ideology. My analysis via @Mediapart #Christchurch'

## Anti-Arab Discourse

### Examples of Counterspeech



**Figure 26** Tweet condemning the Christchurch attack and the Great Replacement theory (source: Twitter)

**Translation** 'the Islamophobic and far-right terrorist killed 49 people in the name of the Great Replacement theory. The massacre of innocents: that what you get out of the hatred of those in France, like in New Zealand, who spread this hateful ideology. #christchurch'



**Figure 27** Tweet condemning the degrading use of the term beurette and response, which argues that the word is not degrading (source: Twitter)

**Translation** 'so for starters, beurette means someone born on French soil whose parents are from North Africa, lol, nothing degrading in the word, it's just been twisted by people by Osman the revenant'

## Case Study

### Discussions of the Great Replacement after Christchurch

One of the primary themes that emerged within this dataset was discussion about the Great Replacement theory.<sup>53</sup> ISD trained algorithms to identify instances in which the term 'great replacement' was used in a manner that sought to provoke or incite hatred against Arab individuals and communities. These posts constituted roughly 4% of the overall sample and over half (55%) of the hateful subset. The roots of this theory can be traced back to Renaud Camus<sup>54</sup> with his book *The Great Replacement*, published in 2011. The theory asserts that white European populations are being deliberately replaced at an ethnic and cultural level through migration and the growth of minority communities.



**Figure 28** Tweet by Renaud Camus advocating for fighting against the great replacement (source: Twitter)  
**Translation** 'Great replacement, Islamisation, Deculturation: resist this civilisational threat'



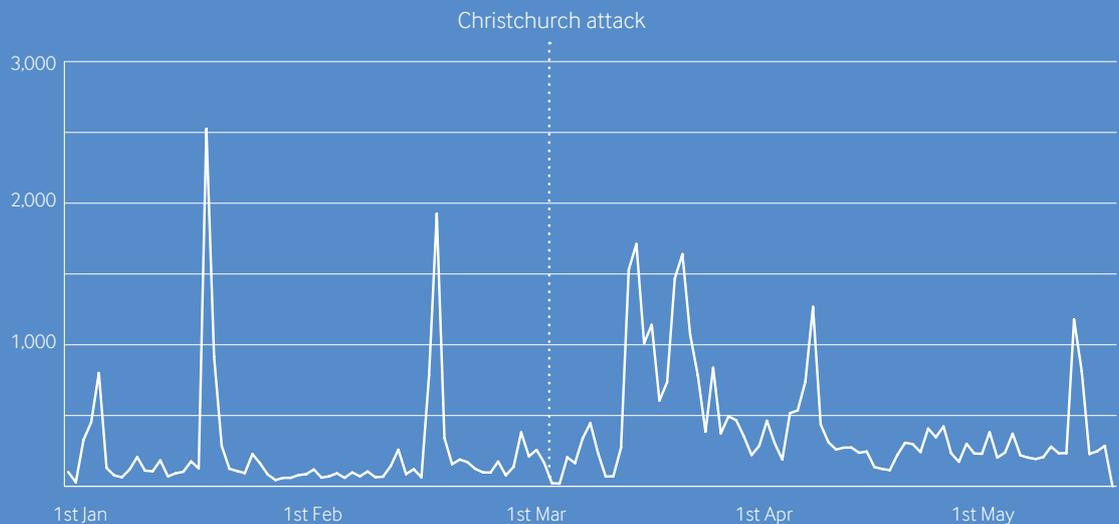
In the hateful dataset, in the week following Christchurch the number of mentions of the great replacement increased, as illustrated in Figure 29.

Recent events, especially the Christchurch attack, highlight the role played by great replacement rhetoric in driving extreme-right activities and violent attacks. ISD's latest report on the great replacement emphasised, 'it is clear that the theory lends itself to calls for radical action against minority communities – including ethnic cleansing, violence and terrorism.'<sup>55</sup>

The promotion of this ideology by public figures has contributed to the normalisation of this rhetoric and can be tied to the spread of hateful content at the community level. The network map of our hateful dataset demonstrates the centrality of the idea of the great replacement to this online community. Users who discuss this idea also refer to Arabs using slurs like rats, porc (pig) and arabe de service (service Arab).

← New Zealand Prime Minister Jacinda Ardern visited the Phillipstown Community Centre on 16th March 2019, less than 24 hours after an attack on two mosques in Christchurch

**Figure 29**  
Mentions of the term 'great replacement' in the hateful anti-Arab dataset between 1 January 2019 and 31 May 2019

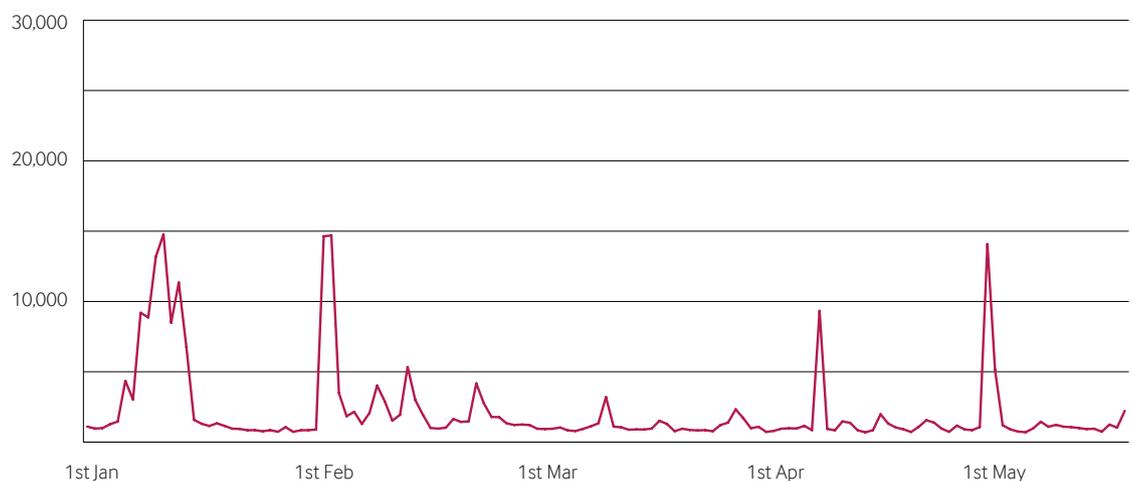


## Anti-Roma and Anti-gypsy Discourse

### Key Findings

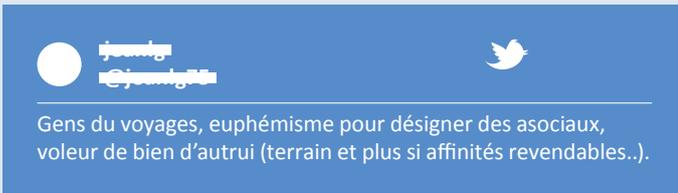
- Anti-Roma or anti-gypsy keywords selected for this report returned 299,157 results (Figure 30).
- Discussions connected with anti-Roma and anti-gypsy keywords spiked as a result of particular news developments, including the **arrest of Yellow Vest activist Christophe Dettinger** in January and **controversial comments by former Minister of European Affairs Nathalie Loiseau**. As a result of the incident involving Dettinger, **accounts related to the Yellow Vest Movement supported the inclusion of Roma in the movement**. These types of posts made up a majority of the dataset.
- **Hateful posts largely focused on stereotypes that associate Roma with crime and characterise them as drains on society**, as exemplified in the posts featured below.
- **The most active accounts associated with our list of keywords largely belonged to members of the Roma or gypsy community**, suggesting that terms were appropriated by the community.
- While the level of hateful speech appeared to be low in the sample, **a series of anti-Roma and anti-gypsy attacks in late March 2019 shed light on how disinformation can be used online to incite hatred and violent attacks**. This case study also illustrates the challenges of identifying hateful content online.

**Figure 30**  
Posts containing anti-Roma or anti-gypsy keywords between 1 January 2019 and 31 May 2019



## Anti-Roma and Anti-gypsy Discourse

### Examples of Hate speech



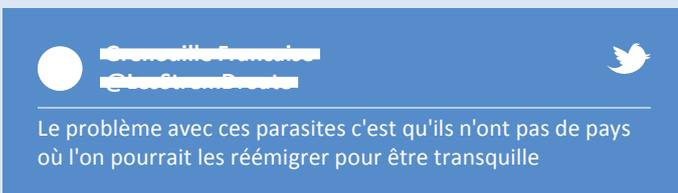
**Figure 31** A tweet stereotyping Roma people as thieves (source: Twitter)

**Translation** 'travelling people, a euphemism to describe anti-social folks who steal people's property (land and more if it can be sold..)'



**Figure 32** A Facebook post in response to Emmanuel Macron's comments on the Dettinger case (source: Facebook)

**Translation** 'Manu is criticising and showing contempt for the way Christophe speaks'



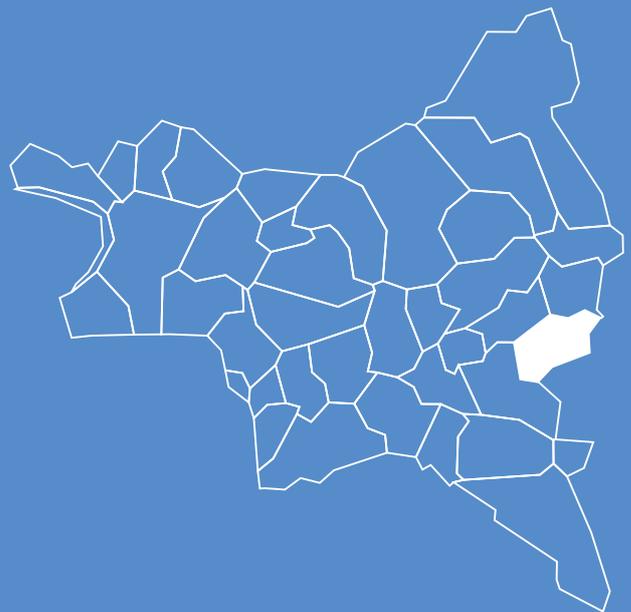
**Figure 33** A hateful anti-Roma tweet (source: Twitter)

**Translation** 'the problem with these parasites is that they don't have a country we could re-migrate them to in order to live in peace'

## Case Study

### The Roma Kidnapping Disinformation Campaign

During the night of 26 March 2019, a series of assaults against Roma people took place on the outskirts of Paris. The attacks, which received extensive coverage in mainstream media and on far-right platforms, were sparked by false rumours on social media alleging that Roma individuals were responsible for kidnappings in the area.



**Figure 34** Map of Seine-Saint-Denis, with Montfermeil district highlighted

The rumour became viral with a tweet posted on 24 March which reported an alleged kidnapping attempt by an organ trafficking Roma network in Montfermeil. It was retweeted over 12,000 times and received over 8,000 likes. An analysis by ISD of Facebook posts about Roma people in the 24 hours that followed the attacks showed that over-performing content were mainly pieces by mainstream media, as well as posts from the anti-racism NGO SOS Racisme

## Case Study

### The Roma Kidnapping Disinformation Campaign

condemning the attacks, which showed that mainstream media coverage of the attacks and counterspeech generated most engagement.

Particularly active in covering the events on Facebook where Russian state media outlets, including RT France and Sputnik. Comments by users below a number of these pieces claimed that rumours of kidnappings and organ trafficking by Roma people were in fact true (Figure 35). Some expressed anti-Roma views.

The rumour which led to the series of attack in March started days before the events, when local football clubs posted messages warning parents against risks of kidnappings (Figure 36). The football magazine Panamefoot published an article about kidnappings of children on 23 March, sharing a post by Football Club Aulnaysien. Football clubs started sending parents text alerts about risks of kidnappings.

The rumour began when, on 24 March, a video of a white van, allegedly belonging to child kidnappers, emerged on Snapchat. The rumours of kidnappings then started spreading on Facebook and Twitter. One tweet from 24 March amplified the disinformation widely, sparking hateful anti-Roma attacks. The post originated from a Twitter account, whose holder described himself a 17-year-old Turkish student fighting disinformation. On 25 March, another Snapchat post spreading false information about kidnappings by Roma people led to several attacks against the Roma community.

In the days that followed the attack, a Facebook page emerged which shared testimonials of parents whose children had allegedly been kidnapped, and stories were also widely circulated on websites dedicated to fighting

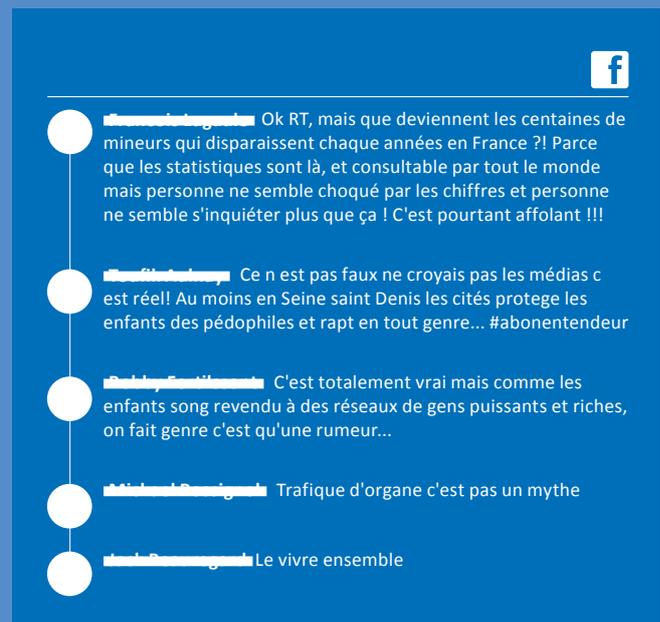


Figure 35 Facebook comments sharing conspiracy theories after anti-Roma attacks (source: Facebook)

#### Translation

- Comment 1 "ok RT, but what happens to the hundreds of under-age kids who disappear each year in France?! Because the statistics are there, and available to everybody, but no one seems shocked by the figures and no one seems to worry! It's frightening though!!!";
- Comment 2 "It's not fake don't trust the media it's real! At least in Seine Saint Denis, people protect kids from paedophiles and kidnappings of all kinds";
- Comment 3 "It's completely true but because kids are sold to networks of powerful and rich people, everybody pretends it's just a rumour...";
- Comment 4 "organ trafficking is not a myth";
- Comment 5 "that's living together in harmony for you"

paedophilia, some of which shared conspiracy theory narratives of a mainstream media cover-up of kidnappings by paedophiles. An article on 26 March shared the same view. It was posted by the anti-paedophilia website 'Wanted Pedo', which is known for sharing disinformation and whose Facebook page was shut down in 2016. Despite mainstream media's coverage of the attacks and attempts at exposing the false rumours, disinformation continued to spread.

Libération's fact-checking team Checknews traced the rumour back to 2012 when three Roma individuals were accused of kidnaping a child. Judges dismissed the case as DNA testing did not confirm the accusations. Since then, several rumours, which were all denied by local authorities, have emerged across several platforms (e.g., Facebook and WhatsApp). In France in 2018, police investigated two stories of attempted kidnappings by men driving a white van.

Surprisingly, we found no significant increase in hateful anti-Roma discourse during the attacks. Several factors can explain this. Snapchat was instrumental in spreading the disinformation campaign and hateful narratives about the Roma community. The nature of the platform makes content from Snapchat extremely difficult to analyse, as videos are deleted within 24 hours. Further, our choice of a keyword-based approach did not allow us to analyse video content.

On platforms such as Twitter and Facebook, the content related to these attacks did not use the keywords selected for our research. Users regularly used the terms kidnappeur, fdp (son of bitch) and Roumain (Romanian). While the terms kidnappeur and Roumain were initially considered, they had to be discarded as they resulted in too many false positives.



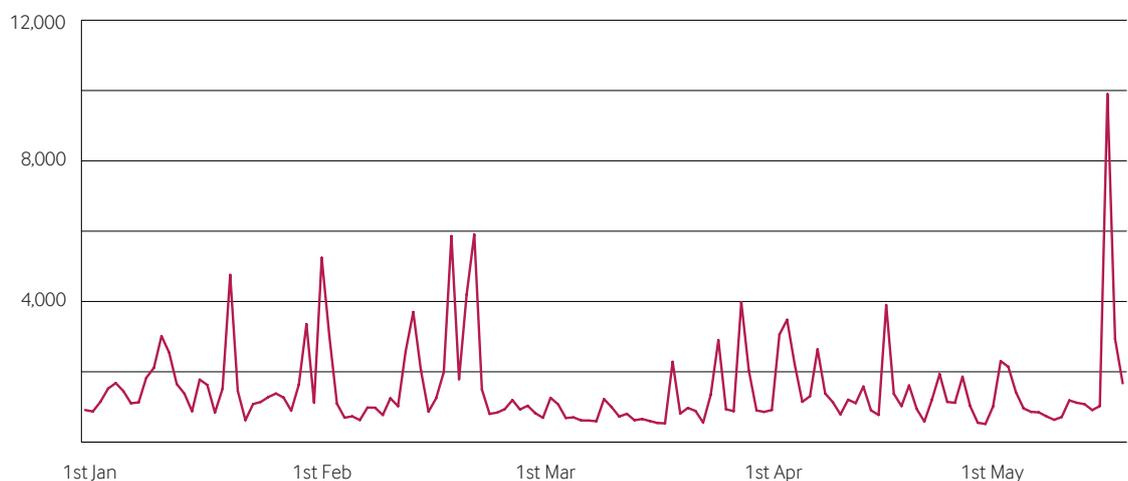
Figure 36 A message sent by a football club alerting against the risk of child kidnappings

## Anti-black or Anti-African Discourse

### Key Findings

- Anti-black or anti-African keywords selected for this report returned 223,483 results (Figure 37).
- **Conversation using anti-black or anti-African keywords was driven by specific events**, including racist attacks against stand-up comedian Donel Jack'sman, the death in police custody of Ange Dibenisha and social controversies related to the depiction of black or African people in French society.
- **The most significant spike in Anti-black or anti-African keywords took place between 12 and 16 May**, as a result of the introduction of a painting at the Assemblée Nationale, to celebrate the anniversary of the first abolition of slavery in France in 1794. The painting sparked a debate around its alleged racism and representation of blackface in the painting.
- **We also observed spikes in conversation from 17 to 20 February following the controversial commercialisation by French bakeries of pastries called tête de nègre (Negro head) and bamboula (pejorative for a black person).** The public backlash against the pastries led to some stores removing them from their shelves.<sup>56</sup> During our research, we identified several accounts which hijacked the incident, arguing that a similar incident involving anti-white terms would not lead to the same level of outrage.
- Conversations about institutionalised racism emerged as a key theme from the keywords selected.
- The social media accounts which used anti-black and anti-African keywords most frequently show a split between explicitly anti-immigration accounts and accounts held by members of the black or African communities in France.

**Figure 37**  
Posts containing anti-black or anti-African keywords between 1 January 2019 and 31 May 2019



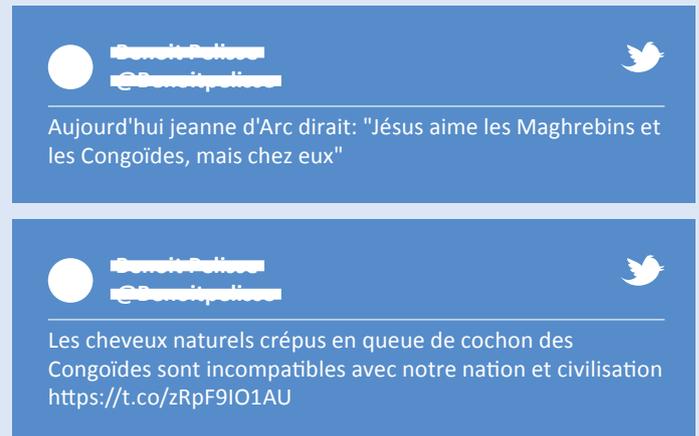
## Anti-black or Anti-African Discourse

### Examples of discourse containing anti-black or anti-African keywords



**Figure 38** A post following the commercialisation of pastries called tête de nègre (source: Facebook)

**Translation** ‘mamadaou’s MATE, THE WORST OF FRANCE IS DROPPING HIS PANTS IN FRONT OF THEM this anti-white hate??? Long live negroes’ heads, and off [with] these idiots’ heads – The Black African Defence League managed to get bakers to withdraw their negroes’ heads pastries: baker is forced to apologise’



**Figure 39** Tweets from Benoit Pélisse, the second most active account among those using keywords associated with anti-Black hateful speech (now suspended), who was president of Fraternité & Réémigration, an organisation that promotes the remigration of non-European populations (source: Crimson Hexagon)

#### Translation

**Tweet 1** ‘today Jeanne d’Arc would say “Jesus loves the Arabs and the Congoïdes, but in their land”’

**Tweet 2** ‘the Congoïdes’ natural frizzy pigtail hair is incompatible with our nation and civilisation’

“ Conversations about institutionalised racism emerged as a key theme from the keywords selected ”

## Anti-white Discourse

### Key Findings

- Anti-white keywords selected for this report returned 125,654 results (Figure 40).<sup>57</sup>
- **We did not identify posts using the terms babtou or toubab with hateful intent or to direct abuse at white people.** Discussions generated by anti-white keywords seemed to focus on discussing stereotypes about white people and prompted debate about the existence of anti-white racism.
- **There was one substantial peak in conversation over 20–26 January**, caused by a post playing stereotypes of black and white people against each other (using the term babtou), but which was not overtly hateful.
- **Most active accounts were in discussions of anti-racism and questioned the existence of anti-white racism.** Four accounts use anti-white keywords to discuss white oppression of black people and French colonisation. The most active pages on Facebook did not share terms in a hateful manner, but rather appear to be false positives as babtou and toubab originate from a Wolof word meaning white.

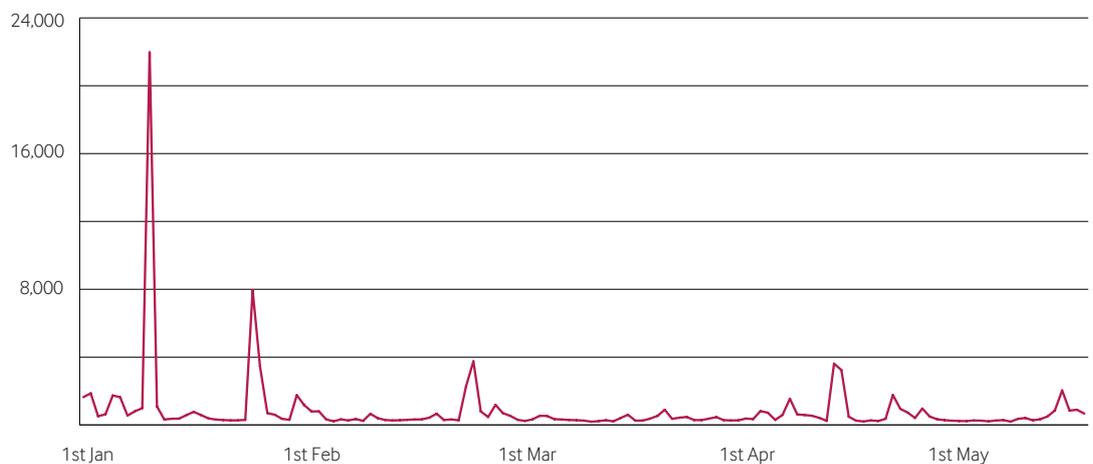
## Anti-white Discourse

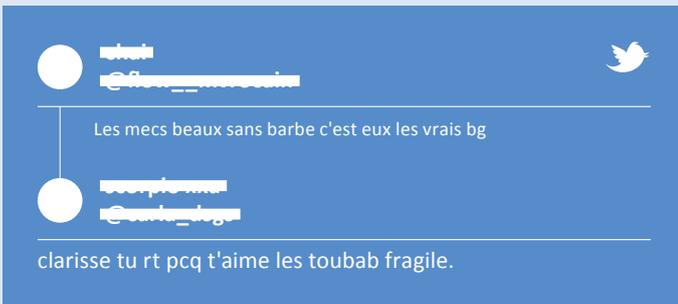
### Examples of conversations with anti-white keywords



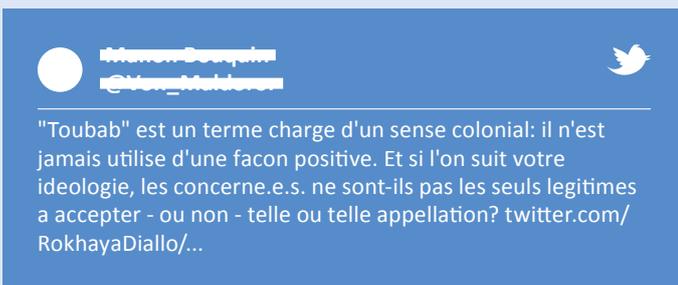
**Figure 41** A tweet using the word toubab (source: Twitter)  
**Translation** 'lol, the white guy is in love'

**Figure 40**  
**Posts containing anti-white keywords between 1 January 2019 and 31 May 2019**

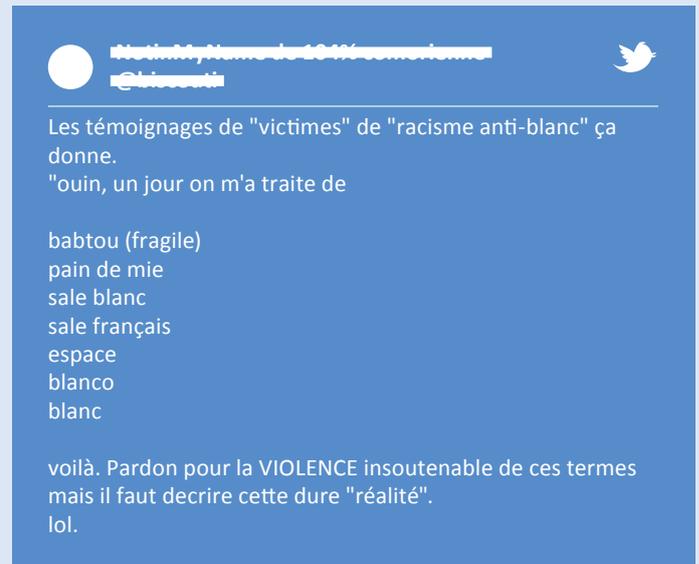




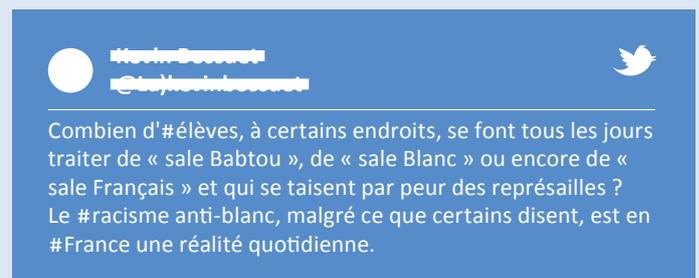
**Figure 42** A tweet using the word toubab (source: Twitter);  
**Translation** ‘Clarisse, you RT cos you like vulnerable white guys’



**Figure 43** Tweet arguing that toubab has negative connotations (source: Twitter)  
**Translation** ‘Toubab is a term loaded with colonial meaning; it is never used in a positive way. And if I follow your ideology, are those who are concerned not the only ones who are legitimate to decide whether they accept this or that term?’



**Figure 44** Tweet arguing that anti-white racism is not comparable to other forms of racism (source: Twitter)  
**Translation** ‘The testimonies of “victims” of “anti-white racism” show: “yeah, one day I was called... fragile white guy, bread crumb, dirty white, dirty Frenchman, space, blanco, white”. That’s it. Sorry for the unbearable VIOLENCE of these terms but we have to describe this hard “reality”. Lol’



**Figure 45** Tweet arguing that anti-white racism exists (source: Twitter)  
**Translation** ‘How many students in some places are called dirty babtou, dirty white, or even dirty Frenchman every day and remain silent for fear of reprisals? Anti-white racism, despite what some may say, is a daily reality in France’

### Anti-Asian Discourse

#### Key Findings

- Anti-Asian keywords selected for this report returned 45,804 results (Figure 46).
- **Anti-Asian keywords were deployed in a variety of contexts, including in established French expressions** such as c’est du chinois (‘It’s Chinese’, used in a similar way to the English expression ‘It’s all Greek to me’). **Researchers did not identify any specific examples of targeted hateful content.**
- A qualitative analysis of a sample of messages using the expression bridés showed that it was used in a variety of contexts without a clear unifying theme, with the majority of posts containing the expression yeux bridés (slanting eyes). This ranged from posts about beauty and make-up to posts asking questions about slanted eyes (examples below).

### Anti-Asian Discourse

#### Examples of online content with anti-Asian keywords

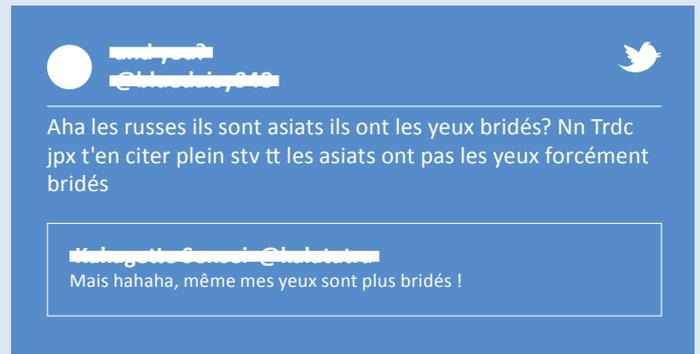


Figure 47 A tweet about slanted eyes (source: Twitter)

**Translation**

**Tweet 1** ‘Aha, the Russians are Asians, they have slanting eyes? No asshole, I can give you loads of examples, often all Asians don’t have slanting eyes’

**Tweet 2** ‘But hahaha, even my eyes are more slanted!’

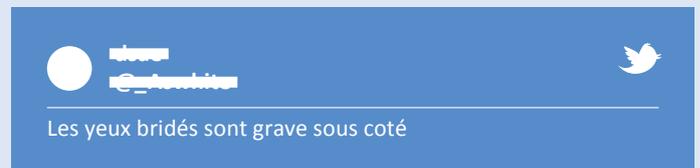
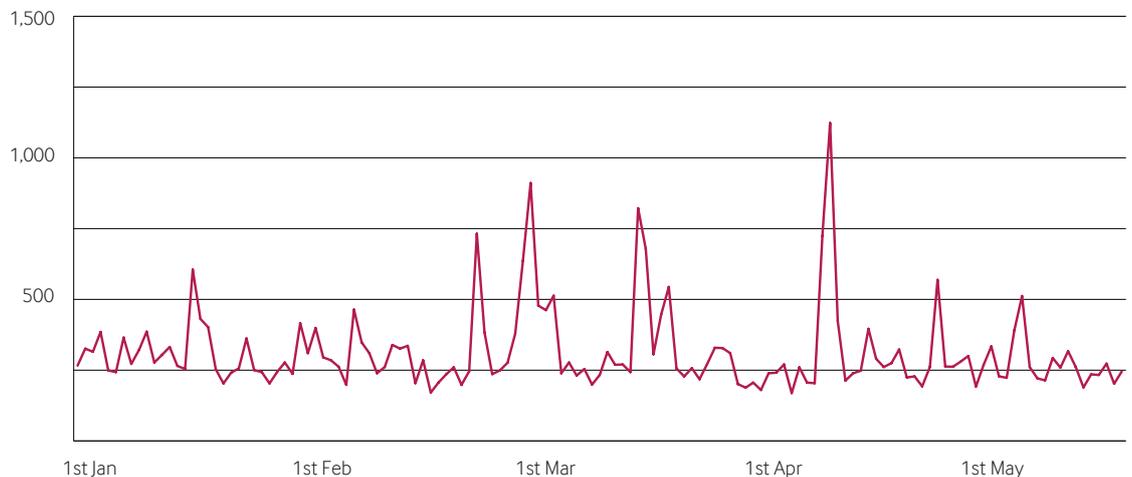


Figure 48 A tweet about slanted eyes (source: Twitter)

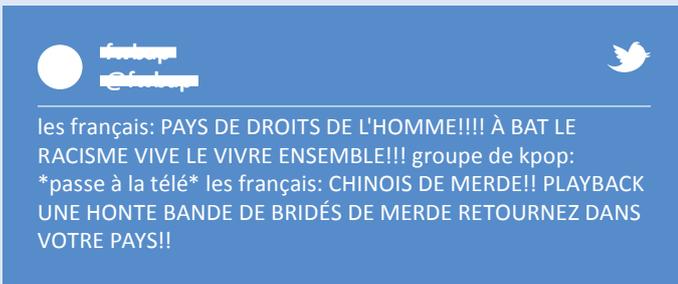
**Translation** ‘slanted eyes are super under-rated’

Figure 46 Posts containing anti-Asian keywords between 1 January 2019 and 31 May 2019



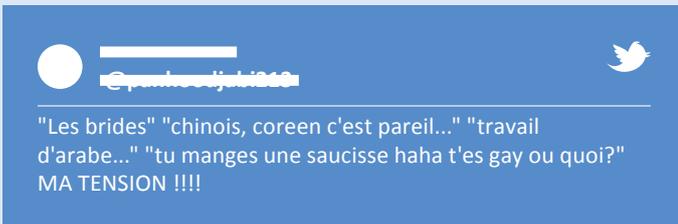
## Anti-Asian Discourse

### Examples of Counterspeech



**Figure 49** A tweet condemning anti-Asian hate (source: Crimson Hexagon)

**Translation** : 'the French: COUNTRY OF HUMAN RIGHTS!!! DOWN WITH RACISM AND LONG LIVE MULTICULTURALISM!!! Kpop group \*is on TV\* the French: FUCKING CHINESE!! PLAYBACK, SHAME, WHAT A BUNCH OF CHINKS FUCK GO BACK TO YOUR COUNTRY!!'



**Figure 50** A tweet mocking prejudices and preconceptions, including about Asian people (source: Twitter)

**Translation** 'the Chinese, Korean, all the same....' 'Arab's work' 'you're eating a sausage, haha, you're gay or what?' FOR FUCK'S SAKE!!!'

## Religion

This section includes anti-Muslim, anti-Semitic and anti-Christian discourses. The dataset gathered from anti-Muslim keywords demonstrated a high proportion of hateful speech, in some cases indicative of co-ordinated efforts. The anti-Semitic dataset contained a large subset of content related to the assault of Alain Finkelkraut at a Yellow Vest rally, as well as some critiques of Israel. Discussions related to anti-Christian keywords centred largely around public debates involving the place of Islam in French society, with little clear anti-Christian hate.

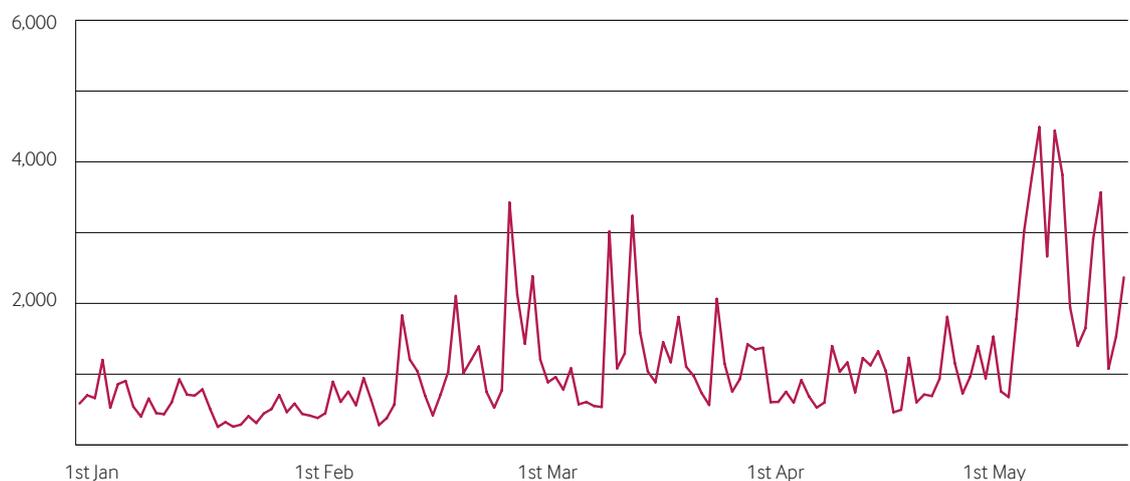
## Anti-Muslim Discourse

### Key Findings

- Anti-Muslim keywords selected for this report returned 168,324 results (Figure 51).
- **Increases in conversation using anti-Muslim keywords are closely correlated with specific events, including the beginning of Ramadan and societal controversy related to external signs of Muslim faith, such as the Etam discrimination case.**
- Qualitative analysis suggests that discourse using anti-Muslim keywords was predominantly hateful, and showed concerted anti-Muslim mobilisation around offline events.
- **Discussions about France’s alleged Islamisation and Muslim invasion dominated the sample**, with various anecdotal examples being presented by users who appeared to harbour anti-Muslim views.
- **Clearly identifiable anti-Muslim account holders were the most active in using anti-Muslim keywords, which suggests there are concerted efforts at targeting the Muslim community online.** Some accounts claimed to be based in Canada, highlighting the international nature of hateful discourse online.

“ Qualitative analysis suggests that discourse using anti-Muslim keywords was predominantly hateful ”

**Figure 51**  
Posts containing anti-Muslim keywords between 1 January 2019 and 31 May 2019



## Anti-Muslim Discourse

### Examples of online content with anti-Muslim keywords



**Figure 52** Facebook post arguing that the burkini is a sign of France's Islamisation (source: Facebook)

**Translation** 'Unbearable to watch and have to put up with this, because of politicians who are complicit with France's Islamisation'



**Figure 53** Post about Islamisation from far-right page (source: Facebook)

**Translation** 'Islamisation: a 15th-century mosque will be built in Nantes with a 9-metre minaret'

## Anti-Muslim Discourse

### Examples of Counterspeech



**Figure 54** Tweet condemning a Daily Mail article about the 'Islamisation' of the town of Saint Denis (source: Twitter)

**Translation** '6 months after its racist and Islamophobic rag about the "Islamisation of Saint Denis", the @mailonline is forced to rectify all the lies it had shared. The article has been removed and the journalist is "off"'

## Case Study

### Damocles Petition to 'End France's Islamisation' at Ramadan

During the period of study, the volume of anti-Muslim conversation peaked markedly during specific events. The largest spikes took place on the week of 5 May (over 20,000 tweets), which marked the beginning of Ramadan in France. This was largely driven by a petition calling for an end to France's Islamisation, which was the most shared piece of content in relation to the keywords chosen. The petition was launched by Damoclès, a French association whose objective is to reduce Muslim immigration to France. The petition was shared widely across Twitter and Facebook.



Figure 55 Facebook page sharing Damoclès's petition (source: Facebook)

Translation 'The latest Etam and Decathlon affairs are signs of France's rampant Islamisation. Act before it's too late!'



Figure 56 Post by Damoclès arguing that young Muslim women get bogus medical sign-offs from their doctor to avoid going to the swimming pool (source: Facebook)

Translation 'Islamisation: there is a mass epidemic of fake medical certificates which generally relate to young Muslim women, to avoid going to the swimming pool. Sign the petition against Islamisation'

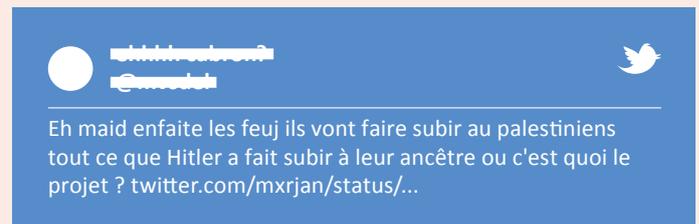
## Anti-Semitic Discourse

### Key Findings

- Anti-Semitic keywords selected for this report returned 79,289 results (Figure 57).
- **Increase in conversation using anti-Semitic keywords spiked sharply following the attack on Alain Finkelkraut**, with over 30,000 posts, during a Yellow Vest demonstration in February 2019. Demonstrators used the term 'sale juif' (dirty Jew) during this period, overwhelmingly to report the incident rather than as a direct attack (There were very few instances of hateful anti-Semitic discourse).
- France reported a 74% increase in anti-Semitic offences in 2018. The discrepancy between the low volume of hateful speech we encountered and the evidence of offline anti-Semitic hate could be the result of our choice of keywords, limited access to comprehensive data from all platforms or the fact that anti-Semitic hateful content is not limited to the use of certain keywords.
- **Critiques of Israeli policy in Palestine emerged as another key theme of this dataset**, though most content was not explicitly hateful (see Figure 58).
- The most active accounts were a combination of explicitly anti-Semitic accounts (most of which have been suspended or removed by platforms), Yellow Vest accounts (mostly related to the Finkelkraut incident), bot-like accounts (posting on diverse issues) and progressive accounts (which critiqued Israeli policies).

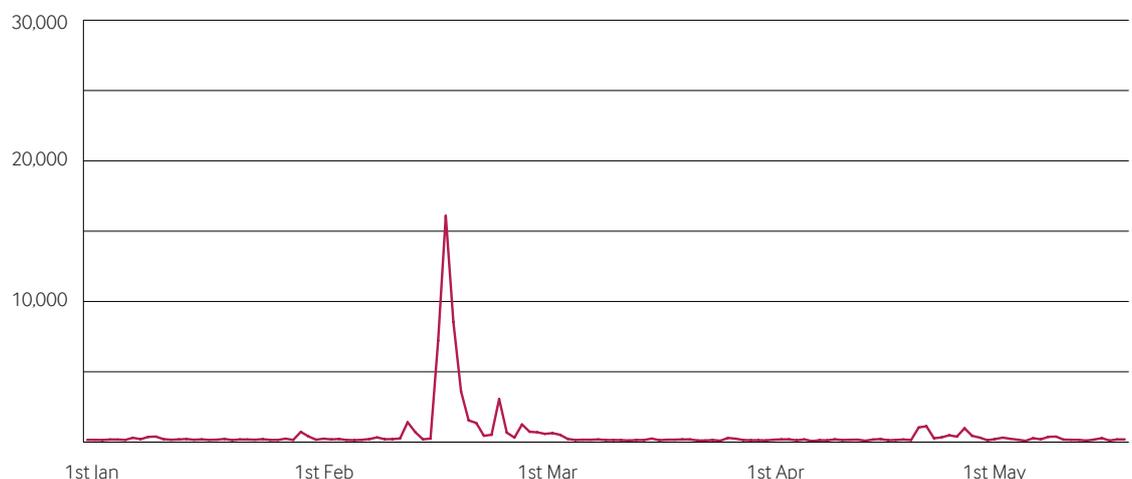
### Anti-Semitic Discourse

#### Examples of online content with anti-Semitic keywords



**Figure 58** Tweet arguing that Jews are treating Palestinians the way Nazis treated Jews during World War 2 (source: Twitter)  
**Translation** 'So basically the Jews will do to the Palestinians what Hitler did to their ancestors, or what's the plan?'

**Figure 57**  
 Posts containing anti-Semitic keywords between 1 January 2019 and 31 May 2019

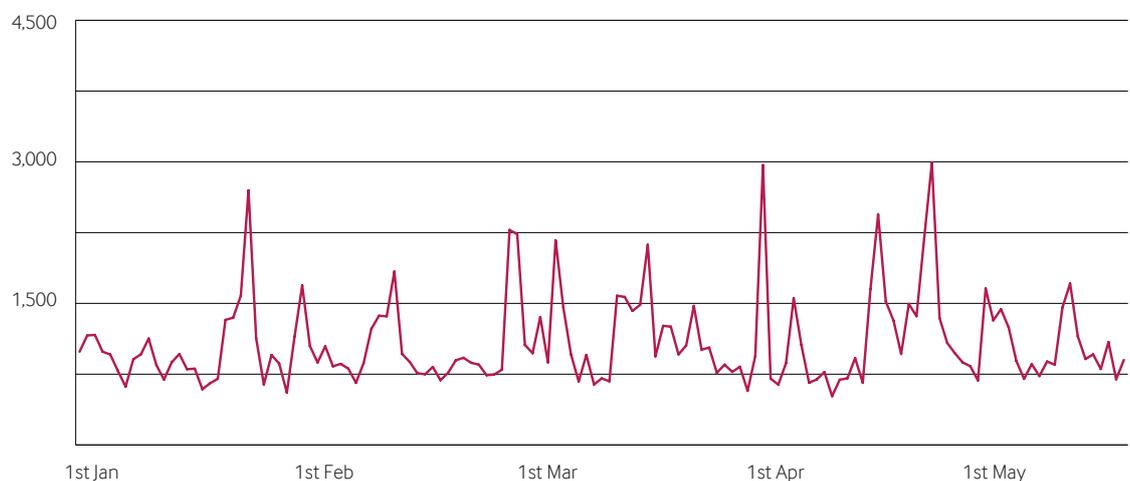


## Anti-Christian Discourse

### Key Findings

- Anti-Christian keywords selected for this report returned 156,047 results (Figure 59).
- **Use of anti-Christian keywords correlated with specific controversies related to Islam.** Discussion about a viral tweet from social media user Hugo that compared a picture of a crowd gathered for prayer in Mecca with a nightclub scene figured prominently in the dataset. The young man's religion and the support he received prompted discussions about 'pro-Muslim' bias, pitting Islam against Christianity.
- ISD's key finding is that **anti-Christian keywords were overwhelmingly used not to target Christians, but to criticise Islam and to engage in anti-Muslim conversations.** For example, many far-right account holders in the dataset claimed that less attention is paid to anti-Christian hate than to anti-Muslim hate in order to appease religious minorities.
- **The term mécréant (unbeliever) was used by politicians and influencers which condemn its use by Islamists** (see Figure 60). This led to spikes of anti-immigration content. In other instances, the term mécréant was used to quote parts of the Qu'ran. A number of far-right influencers used the word mécréant in our sample to convey anti-Muslim sentiment implicitly.
- **Terms denoting bigotry (namely bigot) appeared in 7% of posts.** They were frequently used to attack socially conservative positions on LGBTQ rights and abortion.
- Most active accounts were made up of bot accounts and accounts held by politically engaged users, most of which could be described as liberal.

**Figure 59**  
Posts containing anti-Christian keywords between 1 January 2019 and 31 May 2019



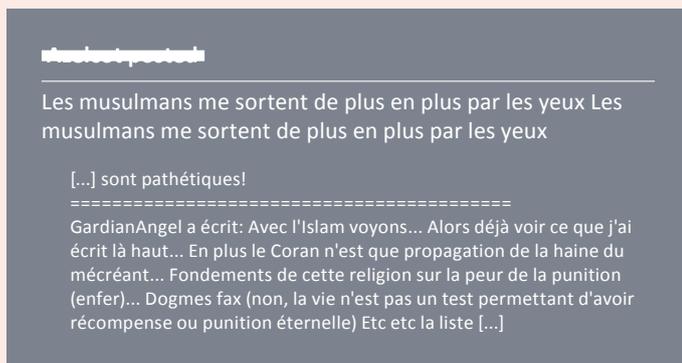
## Anti-Christian Discourse

### Examples of online content with anti-Christian keywords



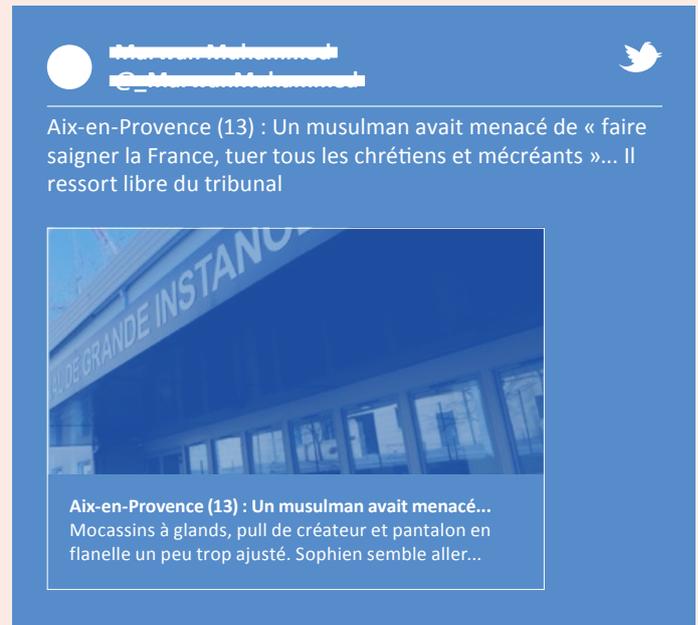
**Figure 60** Post by Marine Le Pen using the word *mécréant* arguing that Islamists are undermining French values (source: Twitter)

**Translation** ‘attacks on our values are an everyday job for Islamists, a game against the “unbelievers”, a game of chess in which they always end up taking the king #OnArrive’



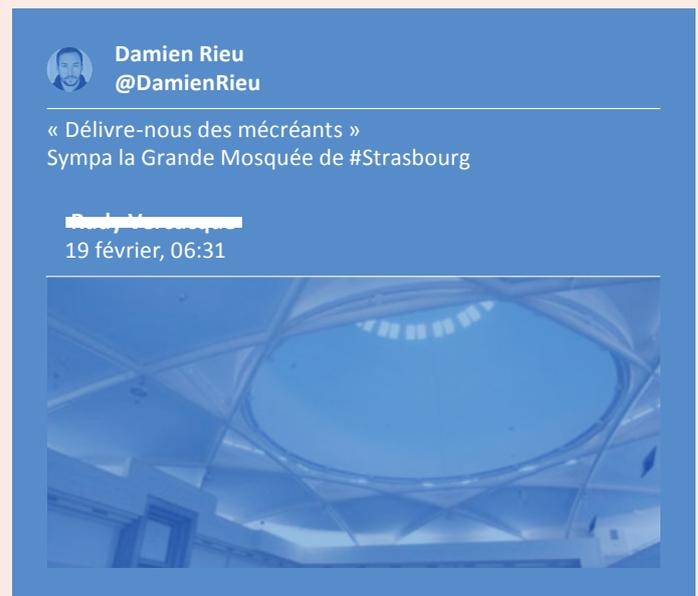
**Figure 61** Anti-Muslim post which uses the word *mécréant* (source: Doctissimo)

**Translation** ‘I can bear Muslims less and less’



**Figure 62** Tweet about the release of a man who had been tried for threatening to ‘kill all Christians’ (source: Twitter)

**Translation** ‘Aix-en-Provence (13): a Muslim man had threatened to “bleed France, kill all Christians and unbelievers...” He is leaving the courts a free man.’

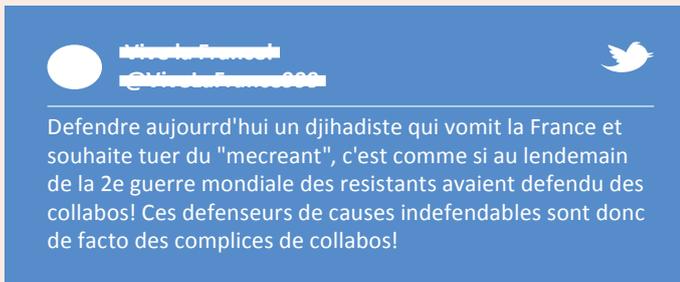


**Figure 63** Post by Identitarian influencer Damien Rieu (source: Twitter)

**Translation** ““Free us from unbelievers”; nice the #Strasbourg great mosque’

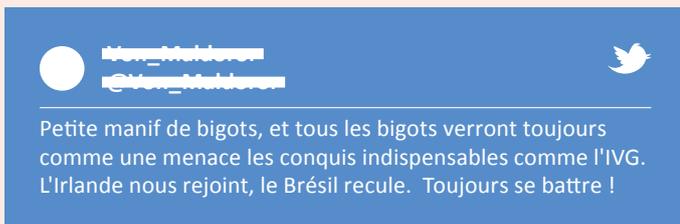
## Anti-Christian Discourse

### Examples of online content with anti-Christian keywords



**Figure 64** Post arguing that France is too lax on jihadists (source: Twitter)

**Translation** 'Defending a jihadist who despises France and wishes to kill "unbelievers", it's as if after World War 2 the resistance had defended collaborators! The defenders of indefensible causes are de facto collaborators!'



**Figure 65** Pro-abortion post using the word bigot (source: Twitter)

**Translation** 'A demonstration of bigots, and all the bigots will always look at wins such as abortion as threats. Ireland is joining us, Brazil is going backwards. Keep fighting!'

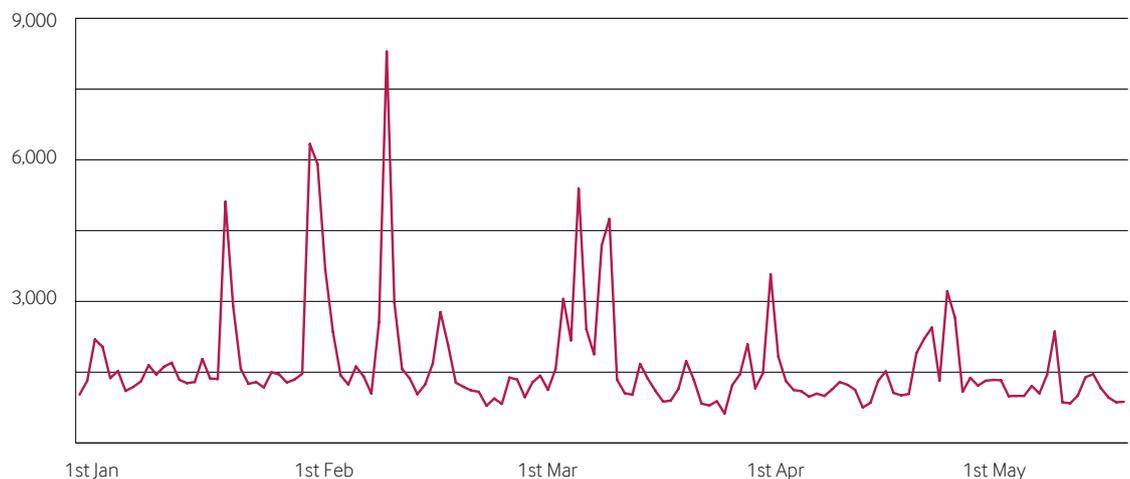
## Disability

## Ableist Discourse

### Key Findings

- Out of the 344,000 relevant posts identified, **our algorithm classified roughly 77% or 265,000 posts as hateful ableist speech during the period studied** (Figure 66).
- **Almost all of these hateful posts were insults or slurs that find their basis in ableism and are normalised in online discourse.**
- **The largest spike in hateful ableist speech was caused by posts mocking rapper Koba LaD, which characterised him as 'gogole'** (pejorative slang for someone with Down Syndrome). The tweet was retweeted nearly 8,000 times.
- **Ableist hateful discourse was largely dominated by vocabulary surrounding Down Syndrome.** This vocabulary was used in a widely indiscriminate way to insult the intelligence of fellow social media users. **'Gogole', 'mongole' and 'attardé', which are all slang terms referring to people with Down Syndrome, all featured prominently in the sample and were used in every day conversation to refer to someone's poor mental abilities, pointing to a generalisation of ableist hateful speech online.**
- The accounts most actively sharing hateful ableist content were held by individuals, some of whom demonstrated interest in gaming.
- As with many other discourses, the accounts most frequently mentioned in conjunction with hateful discourse were politicians, sports accounts and news outlets.
- World Down Syndrome Day, which took place on 21 March 2019, caused a spike in use of ableist keywords overall, as did a tweet from the mother of a girl with Down Syndrome. However, hateful posts did not spike on these occasions, suggesting that counterspeech posts may have been successful.

**Figure 66**  
Posts containing hateful ableist discourse between 1 January 2019 and 31 May 2019



## Ableist Discourse

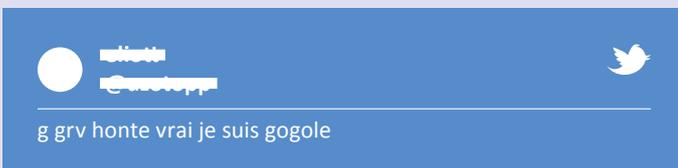
### Examples of hateful speech



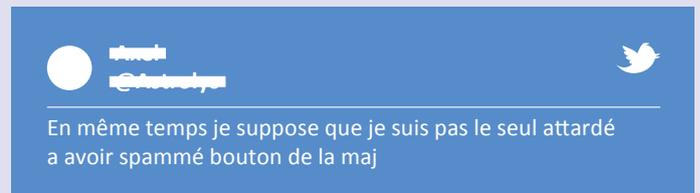
**Figure 67** Post using the word attardé (source: Facebook)  
**Translation** 'According to South Park, ¼ of the population is retarded'



**Figure 68** Tweet using ableist keywords (source: Twitter)  
**Translation** 'And let's also remember that Mongolians are not the inhabitants of Mongolia! The Mongols are!'



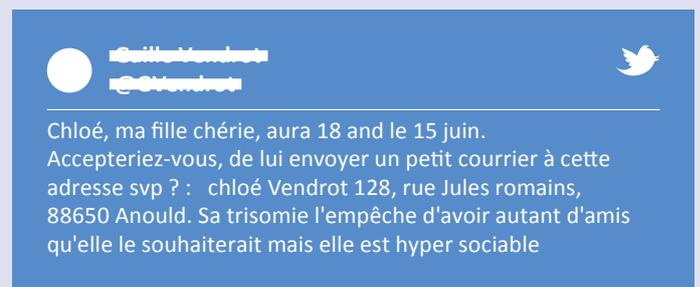
**Figure 69** Tweet in which the author uses an ableist term to refer to themselves (source: Twitter)  
**Translation** 'I'm so ashamed, I'm a retard'



**Figure 70** Tweet in which the author uses an ableist term to refer to themselves (source: Twitter)  
**Translation** 'At the same time, I don't think I'm the only retard who has spammed the caps button'

## Ableist Discourse

### Examples of Counterspeech



**Figure 71** Post by the mother of a Down Syndrome teenager (source: Twitter)  
**Translation** 'Chloé, my darling daughter, will be 18 on 15 June. Would you be so kind as to send her a card to this address please? Chloé Vendrot, 128 rue Jules Romains, 88650, Anould. Her Down Syndrome prevents her from having as many friends as she would like, but she is hyper sociable'

## Intersectionality of Hateful Speech

When analysing different types of hateful speech it is essential to keep in mind that individuals may be targeted by a number of different hateful slurs on the basis of their different identities.

Arab Muslims are frequently targeted both for their ethnic background as well as their faith; gay people of colour are targeted for both their sexuality and race; women receive multiple types of hate based on their gender and other identity categories. Women are particularly likely to be targeted with online violence. A November 2017 report from the French High Commission for Equality found 73% of women had experienced online violence.<sup>58</sup>

The way individuals or groups are targeted by various attacks or prejudices based on their multiple identities is commonly known as intersectionality,<sup>59</sup> referring to the intersections of different identities.

This concept is particularly important in understanding how individuals targeted by hateful speech online may be victims of multi-faceted attacks. It is also important to bear in mind that single users may be responsible for spreading a number of hateful or divisive narratives. A multi-level analysis is necessary to have a full understanding of the hateful speech trends which are occurring online and in order to shape solutions.

This section presents our analysis of the intersections between the hateful content targeting different groups. In order to do this, we identified the accounts that used multiple types of hateful speech as identified by our natural language processing algorithms, allowing us

“ The most significant overlaps could be found between misogynistic content, anti-Arab content and anti-LGBTQ content ”

to identify the types of hateful speech that frequently intersect.

Table 4 presents the percentage of overlap in accounts using hateful speech from each discourse that we were able to analyse using machine learning. The figures indicate the percentage of users that occur in both datasets, as a percentage of each column. Green represents little overlap and red represents significant overlap. Figures are listed as a percentage of each column.

**Table 4 Percentage of overlap between users in different groups employing hateful speech**

	Ableist	Anti-Arab	Anti-LGBTQ	Misogyny
Ableist	100%	33%	31%	22%
Anti-Arab	11%	100%	10%	7%
Anti-LGBTQ	33%	32%	100%	24%
Misogyny	60%	56%	63%	100%

Given how generalised and trivialised misogynistic speech is online, it is unsurprising to find that those who use hateful misogynistic language also make up over 50% of accounts using other types of hateful speech. Users who employ hateful anti-LGBTQ language also make up around a third of both ableist and anti-Arab hateful posters, again highlighting its widespread use.

We also examined to what extent keywords from the four hateful datasets overlapped. The keyword network maps shown in figures 73–76 display the results of this analysis. Each dot on the maps represents an account, and each account is connected to a word that the account holder used in a hateful post.

The most significant overlaps could be found between misogynistic content, anti-Arab content and anti-LGBTQ content:

- Five misogynistic keywords appeared in the anti-LGBTQ language map (Figure 73).
- There were 13 anti-LGBTQ keywords in the misogynistic map (Figure 74), including tante,

## Intersectionality of Hateful Speech

fiotte and pédé, all pejorative slang terms for gay people. Bilal Hassani emerged as a significant figure in misogynistic hateful speech as well as hateful anti-LGBTQ language, demonstrating the extent to which these types of hateful speech overlap, and the extent to which gender and sexuality are often conflated, particularly in hateful speech.

- The term pute appeared frequently in anti-LGBTQ and ableist hateful posts.
- The term pd, which was central in anti-LGBTQ hateful speech, appeared in all four hateful datasets.
- The term beurette, the feminine form of beur (a slang term used to refer to people of North African descent, or Arabs more generally), was used in roughly 9% of the posts in the hateful anti-Arab dataset (Figure 76).
- The term chienne (female dog), which was a central term in hateful misogynistic speech, was also associated with hateful anti-Arab speech.
- While hateful ableist language demonstrated more limited overlap with other types of hateful speech, words like pédé and pute emerged as significant terms within this dataset. This again demonstrates how these slurs are employed in a generalised way, much like ableist slurs such as mingo and attardé.

Our analysis also identified a general conflation of Arabs and Muslims, as well as Arabs and Islamists in hateful discourses. Hateful posts targeting Muslims were ubiquitous in analysis of the anti-Arab dataset, and many posts contained both anti-Arab and anti-Muslim rhetoric.

Figures 73–76 include keywords from our initial lists and other terms that may be used hatefully. These were included to understand to what extent they appeared within hateful speech identified by our NLP algorithms.

---

Figure 73 Network map of key terms in hateful anti-LGBTQ dataset

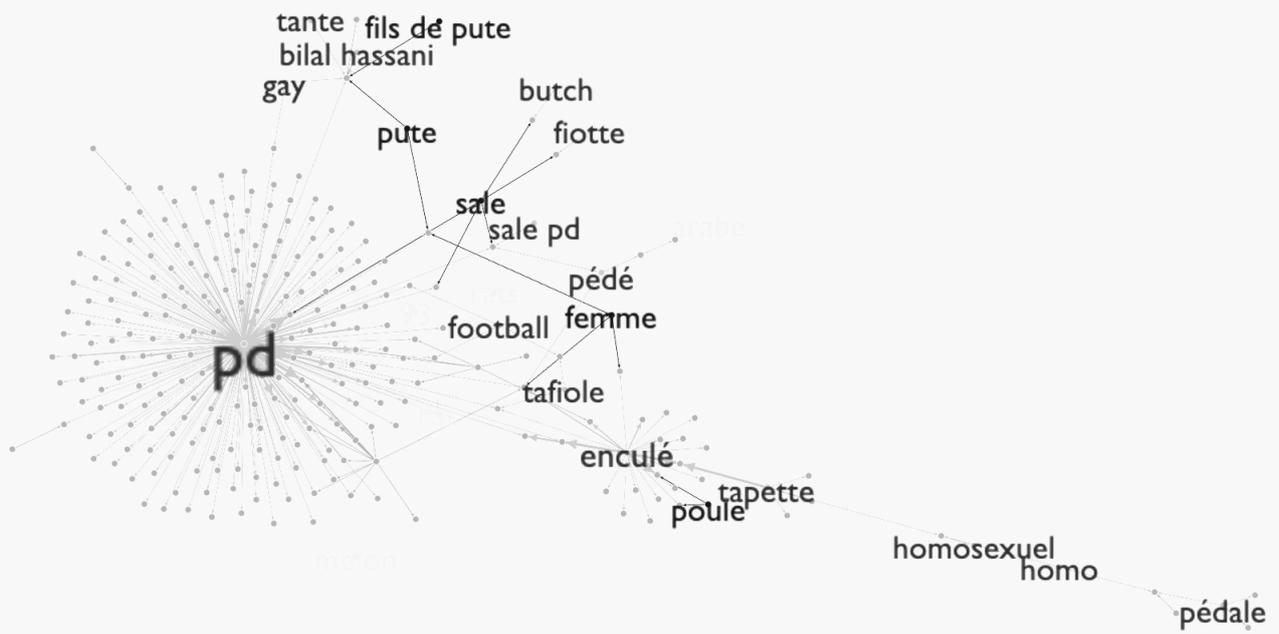




Figure 75 Network map of key terms in hateful ableist dataset

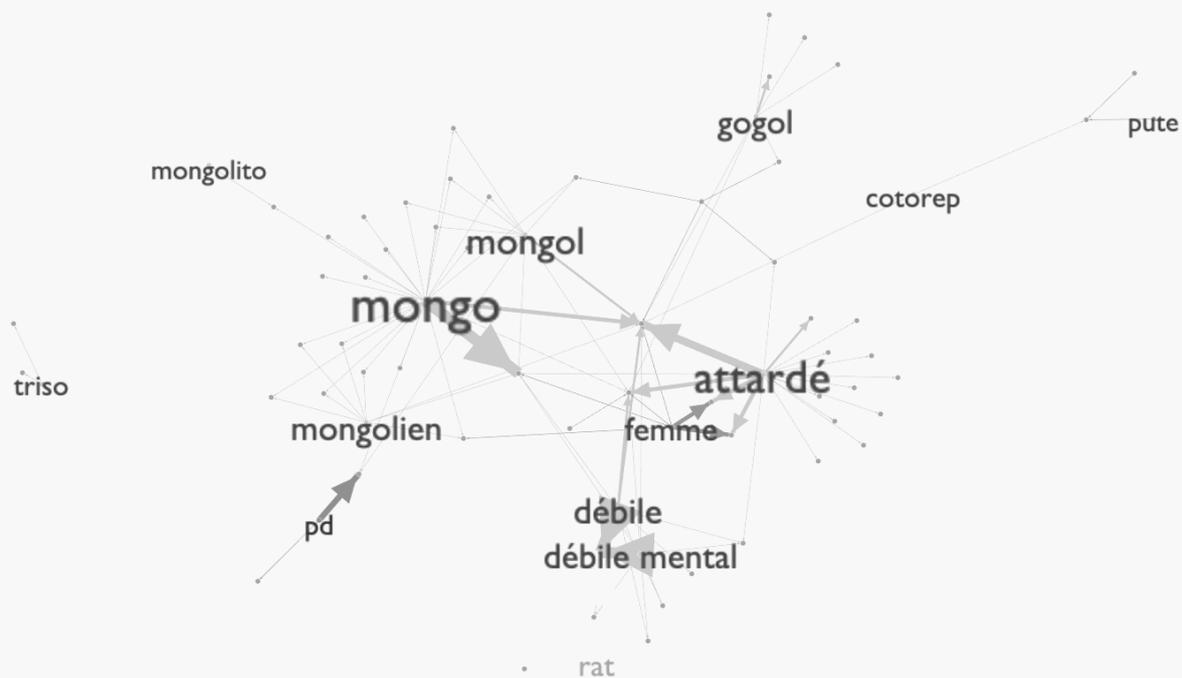
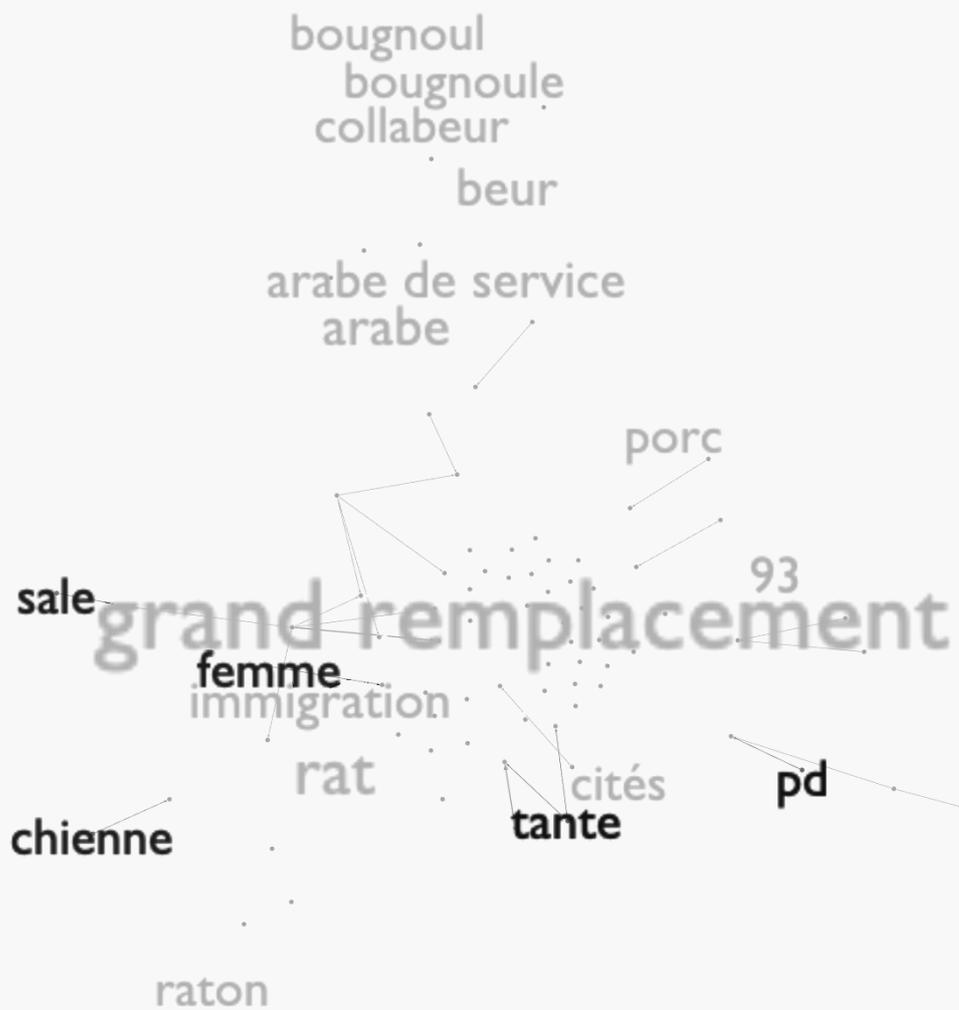


Figure 76 Network map of key terms in hateful anti-Arab dataset



## Recommendations

**In this report we have attempted to provide a comprehensive overview of online hateful discourse in France. We sought to look across all groups whose members experience or claim that they are the targets of hate in order to understand the scale, nature and drivers of those conversations. We aimed to look across social media platforms, as far as data access to those platforms was provided through their APIs and commercial software. And we sought to explore the application of natural language processing and machine learning to the challenge of identifying and analysing hateful discourse at scale. While the platforms themselves use these tools and capabilities, ours was the first attempt to do this in the public realm in France.**

The research revealed a number of insights for government, online platforms, civil society organisations and researchers working to understand the scale and nature of hateful speech online.

This research comes at an important time in the French policy context. As outlined in the introduction, the Loi Avia will soon require ‘high traffic’ online platforms to remove ‘manifestly illegal content’ within 24 hours of being notified. This includes speech that is illegal under French law, including ‘public provocation to hatred’ and ‘abuse, defamation and incitement to discrimination, hate or violence with regards to a person or group based on their origin, belonging or not to an ethnic group, a nation, a race or religion as well as their sexual orientation or disability’. The law also aims to make it easier for users to report content that appears to be illegal.

“ Our recommendations consider the positive effects that social media companies have made to date ”

The Loi Avia represents a key pillar of the French government’s approach, but it is not limited to this. The National Plan Against Racism and Anti-Semitism (2018–2020) proposes a number of measures for ‘fighting hate, racist and anti-Semitic content’, including activities at a European level. Moreover, the French’s government’s inter-ministerial mission team, which includes DILCRAH alongside seven high level experts and three permanent reports from a range of ministries,<sup>60</sup> has set out its initial thoughts on the establishment of a general framework for regulating social networks, starting with the issue of online hate speech, as captured in the interim mission report *Creating a French Framework to Make Social Media Platforms Accountable*.<sup>61</sup> The conclusions we draw from the research presented in this report, and the recommendations we make, have been considered with this policy context in mind.

Our recommendations consider the positive efforts that social media companies have made to date. For example, as cited in the inter-ministerial report, Facebook has provided more transparency on the content of ‘community standards’, increased resources for human and automated content moderation, created a ‘trusted flagger’ programme and provided transparency reports. These efforts have helped to provide researchers and civil society organisations with a better understanding of the scale of hate, as well as the challenges faced by the companies and the actions they are taking. However, being able to tackle and address the presence and impact of hateful content online will require further actions, including putting civil society at the heart of the response.

### **1. Online platforms should increase transparency on public communications and content on their platform by providing open API access to provide better understanding of the scale of hateful discourses.**

As noted in the inter-ministerial mission’s interim paper,<sup>62</sup> there exists ‘an extreme asymmetry of information’ between the social media companies and civil society, which ultimately undermines trust in the companies’ self-regulatory approach to dealing with hateful content. The current level of access limits researchers’ ability to understand the problem, making it difficult for governments and civil society to mount appropriate responses.

Online platforms need to consider how to provide greater data access to public communications through their API to help researchers and civil society better understand and challenge hateful speech online. In doing so, however, it is vital that all precautions regarding privacy are taken into consideration. For example, it would be crucial that only legitimate research institutes are granted this access and only in instances when the research meets certain criteria for public value and respecting user rights, as with greater accessibility come greater risks of this data being used by malign actors for profiling and targeting purposes with a view to sow hate and division, recruit or smear people. Recent changes to Twitter's terms for accessing their API can serve as a good model for balancing ability to enable important research while respecting users' privacy rights.

In addition to API access, increased transparency could include social media companies providing a detailed breakdown of the hateful discourse that is on their platform, sorted by groups targeted by such speech as well as the types of speech identified (e.g. whether it is Tier 1, 2 or 3 according to Facebook's policy on hateful content). As far as possible, this information should be provided at speed and scale to civil society organisations working to counter hate. If these capabilities are not possible, then the online platform operators should develop bespoke tools and algorithms that can analyse hateful discourse in an open and accessible way. One of the key ambitions for the OCCI is to help online platforms to communicate this data to

“ Increased transparency could include social media companies providing a detailed breakdown of the hateful discourse that is on their platform ”

local and regional civil society organisations in a manner that is accessible and actionable, and to advise civil society on effective response strategies.

## **2. Government regulators and online platforms must consider the limits of machine learning algorithms when identifying hateful content.**

The research revealed the limitations of natural language processing and machine learning – particularly in the form of commercial software – to deliver a confident identification of relevant hateful content, and fine-grained breakdown of the different types of hateful content. This has implications for being able to deliver this work at speed and scale.

Image and video content both pose further challenges to this approach. One of the most significant and worrying case studies from the research was the online disinformation campaign alleging that Roma communities were kidnapping children. These videos started locally, spreading via Snapchat. This type of video content is not readily analysable by keyword or language-based approaches. Further investment would need to be made to analyse such visual content algorithmically in a way that would not lead to privacy concerns. For example, mechanisms could analyse video content that is associated with captions calling for violence or using derogatory content, or slurs against protected categories, rather than analyse all uploaded video content, which would be both impractical and present privacy concerns.

As governments begin regulating online platform operators, and such platforms rely increasingly on algorithms to identify and moderate hate speech, all parties must recognise the limits this approach has, and the importance that context plays in moderating hateful content. Artificial intelligence should not be seen as a panacea and approaches to hate speech moderation should be holistic.

Efforts should focus on the testing and application of more human-centric moderation models prioritising the users' own experience regarding wellbeing and ability to self-express safely, by providing more opportunities for moderation personalisation in addition to and within the limits of existing legal frameworks.

Although online platform operators' moderation models already comprise levels of automated and

human reviews, there is scope for introducing much more robust and thorough human language learning and classification processes. Such enhanced processes should be combined with additional layers of artificial intelligence analysis drawing from a more comprehensive range of factors to observe systematically, including better consideration of users' interpretation (categorisation) and human and algorithmic observation of the nature of the interactions between the involved users. This point is particularly crucial in the context of harassment and would ensure online users can control grey area content from a legal standpoint, which can yet be highly disturbing, unwanted content from a personal point of view, with potential associated risks, ranging from their mental toll, to the possibility the viewer might self-censor, or replicate malign or extremist behaviours, and so on.

### **3. Online platforms should work closely with civil society organisations to tackle hateful content that is legal but nonetheless problematic and harmful.**

The Loi Avia will require online platforms to remove manifestly illegal hateful content. More partnerships with civil society organisations should be pursued to tackle the much larger body of hateful content that does not meet this threshold.

As part of this, social media companies should undertake research into the experience and views of groups that are frequently subjected to online hateful speech or are the basis for widespread slurs. This can help social media companies adopt a user-focused approach to understand the types of hateful content that are most harmful in order to help drive decisions around prioritisation in content moderation. This can also help companies develop and apply a range of solutions matched to the severity of content, from removal to minimising reach, targeted reminders of terms of service, education initiatives, counterspeech and victim support. Civil society organisations can help companies decide what the appropriate responses are to different types of hateful content. This can help to build trust and partnerships between social media companies and civil society groups.

Online platforms should also work with civil society organisations to trial and test a range of direct engagement techniques that go beyond simply creating and disseminating counterspeech campaigns. Research

demonstrates that attitudinal change and ideological recruitment happens less through content and more through online engagement on forums, in groups or via direct messaging. ISD's programme Counter Conversations demonstrated that direct engagement can lead to a sustained conversation with an individual publicly sharing hateful content.<sup>63</sup> The next generation of effective counterspeech and online interventions will take place in the 'iterative spaces' online where those drawn to extremist ideologies congregate: in the comment threads and forums of 'grey area' content for which there are no solid grounds for removal.

### **4. Online platforms should provide increased transparency on moderation policies and approaches, and the role of algorithms and automated accounts in spreading hateful content.**

It is difficult for those outside the companies to assess the effectiveness of user reporting of hateful content and the process for assessing and removing that content, though some effort has been made via the monitoring exercises of EU Code of Conduct on countering illegal hate speech online.<sup>64</sup> This lack of data makes it difficult to determine whether identified differences in the scale of hateful discourse targeting certain groups is due to more effective moderation policies in some instances. Lack of transparency here opens the companies up to potential criticism that their content moderation policies effectively protect some groups more than others. To respond to these issues, tech companies should pull back the curtain on their moderation practices and bring researchers and civil society organisations in to understand their approaches and some of the challenges they face.

In response to the UK's Online Harms White Paper, ISD outlined a series of recommendations and arguments for increased transparency from the social media companies across four categories: content and communications, advertising, complaints and redress, and algorithms.

Transparency on content moderation processes and greater oversight from a regulator is needed to ensure the processes are appropriate, well-resourced and accurate, as outlined in the inter-ministerial mission's interim report. This should include transparency on the scale and nature of users' complaints about hateful content and the actions taken in response.

## Recommendations

Companies should ensure they have in-house expertise on these issues and systematic dialogue with the groups frequently targeted. It is also important for tech companies to provide greater transparency on the working conditions and pastoral support provided to content moderators.

It is also vital that there is greater transparency on the role of algorithms in amplifying content that may be hateful, particularly during those events where we saw increased scale of hateful content. A regulator could also be expected to require this of online platforms. Related to this is a need for greater transparency on understanding the role of automated accounts in spreading hateful content. Our research revealed a number of bot-like accounts spreading hateful content, including many accounts that it appeared had been removed by social media platforms at a later date. Better understanding of the phenomenon and impact of online hateful content requires increased transparency on these issues.

### **5. Online platforms, government and civil society organisations need to collaborate on effective campaigns to tackle the widespread, normalised use of slurs in society.**

While Facebook and Twitter address the subject of slurs that 'negatively target' or are 'non-consensual' in their community guidelines, the widespread use of these terms as presented in this report can make it extremely difficult to identify instances at scale. This is complicated by the fact that some groups have reclaimed these terms. An effort to remove such slurs at scale would lead to a significant backlash and be overly draconian. At the same time, as our research demonstrates, it is possible to identify hateful and targeted uses of these slurs which could help to inform campaigns to address them.

Challenging hateful or targeted uses of slurs could include civil society-led campaigns to reclaim or re-appropriate terms used as slurs; educational programmes to address their use in younger communities; and communications programmes that aim to highlight the negative impact that normalised slurs can have on specific communities. Those introducing these efforts should draw from best practice understanding of behaviour change campaigns, including those that have addressed normalised slurs or casual racism

successfully. They should focus on those communities where these slurs are widespread, including football fans and gamers. Practitioners should also be mindful of the potential for backlash: a poorly designed and targeted campaign could end up retrenching some people's use of such slurs as part of subversive internet culture railing against 'political correctness'.

### **6. Online platforms, government and researchers need to pay greater attention to the intersectional nature of hateful speech.**

More research needs to be conducted on the intersectional aspect of hate speech. Very few research studies to date have considered the intersectional nature of hateful discourse. Our research provides some insight into these trends. But greater efforts, and funding, is needed to support further research.

The focus on intersectional hateful speech should extend to social media company staff and civil society campaigners, who need to be conscious of the intersectional aspects of hateful speech and its impact on different populations. Campaigns should not focus exclusively on one type of hateful speech, as the different types of hateful speech are not clearly delineated. Campaigners should consider the different types of hateful speech used to target communities. This applies not just to communications campaigns, but also to other types of intervention. Civil society campaigners should attempt to build coalitions to address the intersectional aspects of hate more effectively, working across traditional ideological lines.

“ Challenging hateful or targeted uses of slurs could include civil society-led campaigns to reclaim or re-appropriate terms used as slurs ”

Social media platforms must consider the intersectional aspect of hateful speech as they create and implement content moderation policies. If accounts are identified as being repeat offenders, they should be further investigated to understand if they are spreading other types of hate.

### **7. Media, government, local authorities, police and online platforms should try to create a co-ordinated mechanism for responding to events that tend to cause spikes in hateful speech.**

This report demonstrated how specific events can drive hateful discussions, and tech companies should be prepared for these types of spikes. Campaigners should prepare themselves for spikes in hateful speech that may occur following significant news events. They could model their response on or take inspiration from the GIFCT Content Incident Protocol to prevent the sharing of violent and terrorist content, and leverage the OCCI. For example, they could provide civil society organisations with more data about the scale and nature of hateful discourse taking place in the wake of events, and provide free advertising credits and advice to civil society organisations producing counterspeech campaigns after they occur.

### **8. Greater attention should be given to the relationship between online hateful content and offline hate crimes or incidents (such as attacks).**

This should include further research on these phenomena to determine if there is a correlation between them. ISD is currently working to develop geo-location online hate mapping capabilities that can help provide local civil society and sub-national governments with a better understanding of the scale and nature of online hateful content coming from their areas. This capability was first developed by ISD's technology partner CASM Consulting LLP in partnership with the London Mayor's Office for Policing and Crime and the Metropolitan Police. In addition to being confident in identifying and geo-locating online hate speech down to borough level in London, the research demonstrated a correlation between online speech and offline hate crime statistics. This capability could help local and regional government staff predict when and where certain communities may be at greater risk of hateful attacks. This can enable authorities to channel resources more efficiently to prevent such attacks, for example through preventative community policing,

raising awareness on reporting hateful attacks and victim support.

Official French police statistics on hate crime already provide a 'heat map' showing the distribution of hate speech across France. Efforts should be made to explore how the geographical distribution of hate crimes in France match up to online hate speech. French police should consider including a new category of online hate crime in their official statistics. In the UK, it has been mandatory to include statistics on online hate crime since 2015. If French police were required to report online hate crime, the data and statistics accumulated over time could provide a basis for understanding the relationship between online hateful content and offline hate-inspired attacks.

# Appendices

## Appendix 1: Lists of Keywords

The keyword lists created for this research report are displayed below, as they were entered into the two social listening tools discussed in the methodology section, Crimson Hexagon and CrowdTangle.

### Anti-Arab or Anti-Maghrebin Keywords

#### Crimson Hexagon

beur\* OR beurette\* OR rabza\* OR beureu\* OR bicot\* OR raton\* OR bounioul\* OR bougnoule\* OR bougnoul\* OR racaille\* OR crouille\* OR blédard\* OR métèque\* OR 'arabe de service' OR 'sale arabe' OR 'arabe voleur' OR 'arabe pas français' OR 'arabe étrangers' OR 'grand remplacement'

#### CrowdTangle

beur, beurette, rabza, beureu, bicot, raton, bounioul, bougnoule, bougnoul, racaille, crouille, blédard, métèque, arabe de service, sale arabe, arabe voleur, arabe pas français, arabe étrangers, grand remplacement

### Anti-black or Anti-African Keywords

#### Crimson Hexagon

'sale noir' OR négroïde\* OR congoïde\* OR bamboula\* OR 'face de pygmée' OR macaque\* OR métèque\* OR banania\* OR nègre\* OR négresse\* OR 'noir de service' OR 'nègre de maison' OR 'race inferieur' OR babouin\*

#### CrowdTangle

sale noir, négroïde, congoïde, bamboula, face de pygmée, macaque, métèque, banania, nègre, négresse, noir de service, nègre de maison, race inferieur, babouin

### Anti-Roma Keywords

#### Crimson Hexagon

tsigane\* OR tzigane\* OR romanichel\* OR romanichelle\* OR sinté\* OR sintée\* OR manouche\* OR gitan\* OR gitane\* OR bohémien\* OR bohémienne\* OR caraque\* OR 'noi' OR manouche\* OR manouches\* OR yéniches\* OR 'race nomade' OR 'sale roumaine' OR 'sale roumain'

#### CrowdTangle

tsigane, tzigane, romanichel, romanichelle, sinté, sintée, manouche, gitan, gitane, bohémien, bohémienne, caraque, noi, manouche, manouches, yéniches, race nomade, sale roumaine, sale roumain

### Anti-Asian Keywords

#### Crimson Hexagon

niakoué\* OR niakouée\* OR niaqué\* OR niaquée\* OR niaquoué\* OR niaquouée\* OR chinetoque\* OR chinetoc\* OR chinetok\* OR noich\* OR noichi\* OR 'face de citron' OR bridé\* OR 'sale asiat' OR 'sale chinois'

#### CrowdTangle

niakoué, niakouée, niaqué, niaquée, niaquoué, niaquouée, chinetoque, chinetoc, chinetok, noich, noichi, face de citron, bridé, sale asiat, sale chinois

### Anti-white Keywords

#### Crimson Hexagon

babtou OR toubab

#### CrowdTangle

babtou, toubab

### Anti-Muslim Keywords

#### Crimson Hexagon

muzz OR 'envahisseur musulman' OR 'immigration islamique' OR islamisation OR 'musulman terroriste' OR 'islam terrorisme' OR 'danger islam' OR 'menace islamiste' OR 'musulman délinquant'

#### CrowdTangle

muzz, envahisseur musulman, immigration islamique, islamisation, musulman terroriste, islam terrorisme, danger islam, menace islamiste, musulman délinquant

### Anti-Semitic Keywords

#### Crimson Hexagon

youpin\* OR youpine\* OR feuj\* OR 'sale juif' OR 'sale juive' OR juiverie OR 'complot juif' OR 'mafia juives' OR hoananas OR 'pornographie mémorielle' OR shoax OR 'propagande sioniste' OR 'juif voleur' OR 'juif rothschild' OR 'juif franc-macon' OR 'juif traître' OR 'sous race de juif'

#### CrowdTangle

youpin, youpine, feuj, sale juif, sale juive, juiverie, complot juif, mafia juives, hoananas, pornographie mémorielle, shoax, propagande sioniste, juif voleur, juif rothschild, juif franc-macon, juif traître, sous race de juif

## Anti-Christian Keywords

---

### Crimson Hexagon

mécréant\* OR kouffar\* OR kâffir\* OR bigot\* OR bigotte\* OR 'grenouille de bénitier' OR 'sale catho'

### CrowdTangle

mécréant, kouffar, kâffir, bigot, bigotte, grenouille de bénitier, sale catho

## Misogynistic Keywords

---

### Crimson Hexagon

'féministe hystérique' OR salope\* OR pute\* OR biatch\* OR bitch\* OR poufiasse\* OR connasse\* OR emmerdeuse\* OR pétasse\* OR salasse\* OR grognasse\* OR guenon\* OR 'mal baisée' OR blondasse\* OR 'femme hystérique' OR 'feministe de service' OR poulette\* OR tepu\* OR pouffe\* OR 'bonne a rien' OR chaudasse\* OR 'meuf vulgaire' OR 'feministe totalitaire' OR 'grosse chienne' OR gonzesse\* OR 'sale meuf' OR 'chienne féministe' OR bobonne\*

### CrowdTangle

féministe hystérique, salope, pute, biatch, bitch, poufiasse, connasse, emmerdeuse, pétasse, salasse, grognasse, guenon, mal baisée, blondasse, femme hystérique, féministe de service, poulette, tepu, pouffe, bonne a rien, chaudasse, meuf vulgaire, féministe totalitaire, grosse chienne, gonzesse, sale meuf, chienne féministe, bobonne

## Anti-LGBTQ Keywords

---

### Crimson Hexagon

pédé\* OR pd OR pédale\* OR tapette\* OR tarlouze\* OR tantouze\* OR tafiole\* OR fiotte\* OR gouine\* OR gouinasse\* OR camionneuse\* OR butch\* OR goudou\* OR follasse\* OR travelo\* OR enculé\* OR pédéraste\* OR lopette\* OR 'broutte minou' OR 'sale pd' OR 'sale gay'

### CrowdTangle

pédé, pd, pédale, tapette, tarlouze, tantouze, tafiole, fiotte, gouine, gouinasse, camionneuse, butch, goudou, follasse, travelo, enculé, pédéraste, lopette, broutte minou, sale pd, sale gay

## Ableist Keywords

---

### Crimson Hexagon

mongo\* OR mongol\* OR mongolien\* OR mongolito\* OR gogol\* OR attardé\* OR 'espèce d'handicapé' OR cotorep\* OR triso\* OR 'débile mental'

### CrowdTangle

mongo, mongol, mongolien, mongolito, gogol, attardé, espèce d'handicapé, cotorep, triso, débile mental

## Appendix 2: Method52

### Technology

The core technology used in this project is called Method52. It has been built to allow people who don't have a formal data science background to collect, analyse and visualise datasets that are very large and unstructured. This is especially the case for large, text-based datasets, such as those drawn from social media, but has also included datasets comprised of emails, forum data and internal and proprietary data held by large organisations.

The design principle of Method52 is to create a development environment through a graphical user interface. Users select, configure and connect a number of components to create a bespoke pipeline that data flows through. Each of these pipelines is designed to perform a particular task and often a number of pipelines are themselves connected together to fulfil a particular research-driven function. There are 82 components in Method52, and many of them can be configured to perform a number of different tasks.

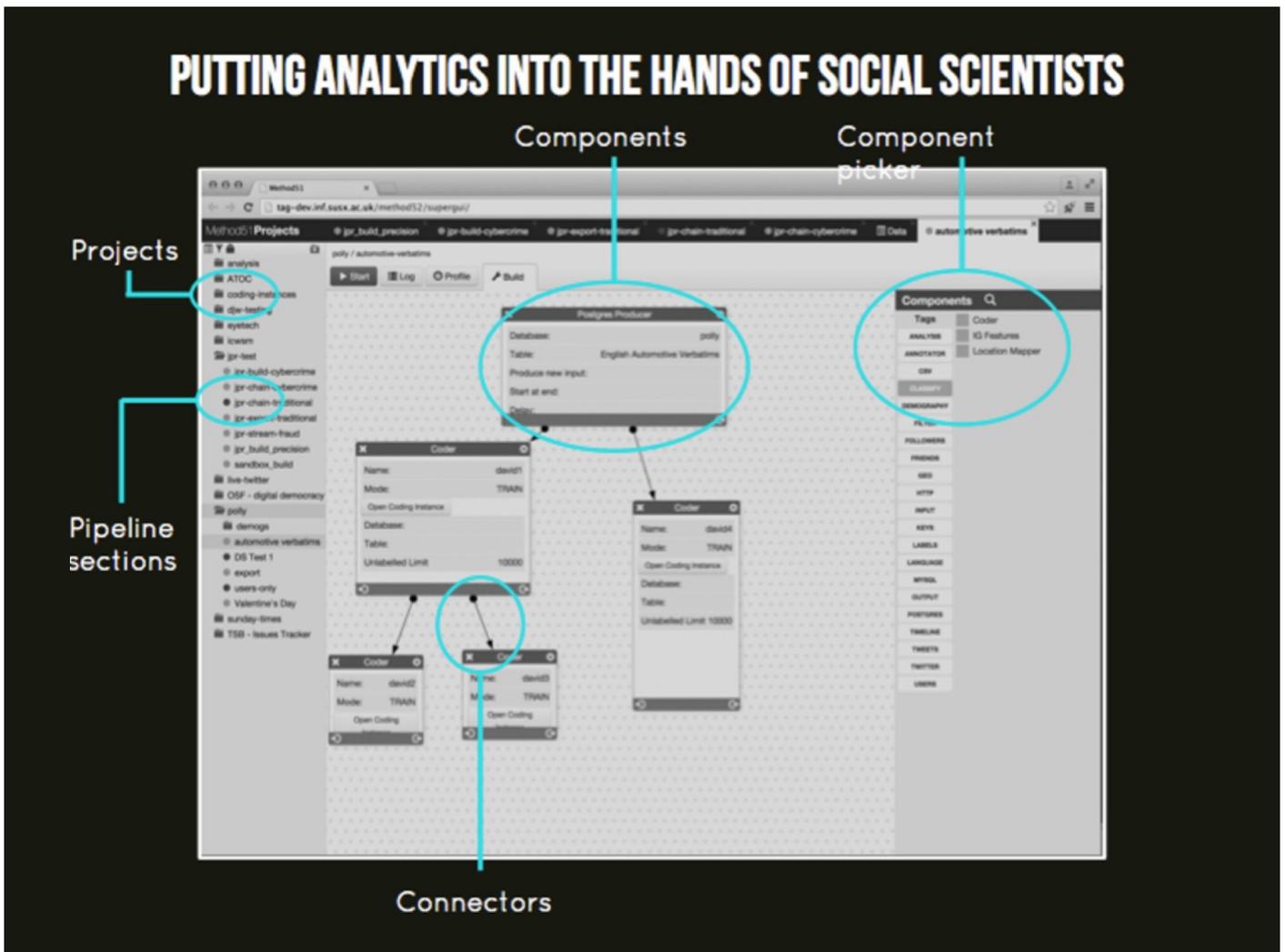


Figure 77 Method52 example pipeline

### Classification of Themes

Method52 allows researchers to train algorithms to split apart ('classify') documents into categories, according to the meaning of the document, and on the basis of the text they contain. To do this, it uses a technology called natural language processing, which is a branch of artificial intelligence research. It combines approaches developed in the fields of computer science, applied mathematics and linguistics.

Throughout this project, ISD researchers marked up posts on the basis of whether they considered them to fall inside or outside the categories each algorithm was trained to distinguish between. This 'teaches' the algorithm to spot patterns in the language use associated with each category chosen. The algorithm looks for statistical correlations between the language used and the categories assigned to determine the extent to which words and bigrams fall into the pre-defined categories.

To measure the accuracy of algorithms into the categories chosen by the analyst, we use a 'gold standard' approach. For each algorithm, around 100 documents are randomly selected from the relevant dataset to form a gold standard test set for each classifier. These are then manually coded into one of the categories defined above. The 100 documents are then removed from the main dataset and so are not used to train the classifier. As each classifier is trained, the software reports back on how accurate the classifier is at categorising the gold standard compared with the analyst's decisions.

There are a number of ways of measuring classifier accuracy. We used overall accuracy: the percentage likelihood of any randomly selected document within the dataset being placed into the appropriate category by the algorithm.

### Training Process

Building algorithms to categorise and separate posts formed an important part of the research method for this paper. This responds to a general challenge of social media research: the data that is routinely produced and collected is too large to be manually read.

Natural language processing classifiers provide an analytical window into these kinds of datasets. They are

trained by analysts on a given dataset to recognise the linguistic difference between different kinds of data. This training is conducted using Method52.

Each classifier was built by using Method52's web-based user interface to proceed through eight phases, which are described below.

#### Phase 1: Definition of Categories

The formal criteria explaining how posts should be annotated is developed. Practically, this means that a small number of categories – between two and five – are defined. These will be the categories that the classifier will try to place each (and every) post within. The exact definition of the categories develops throughout the early interaction of the data. These categories are not arrived at a priori, but rather iteratively, informed by the researcher's interaction with the data – the researcher's idea of what comprises a category is often challenged by the actual data itself, causing a redefinition of that category. This process ensures that the categories reflect the evidence, rather than the preconceptions or expectations of the analyst. This is consistent with a well-known sociological method called 'grounded theory'.

#### Phase 2: Creation of a Gold Standard Test Dataset

This phase provides a source of truth against which the classifier performance is tested. A number of posts (usually 100, but more if the dataset is very large) are randomly selected to form a gold standard test set. These are manually coded into the categories defined during Phase 1. The posts comprising this gold standard are then removed from the main dataset, and are not used to train the classifier.

#### Phase 3: Training

This phase describes the process wherein training data is introduced into the statistical model, called 'mark up'. Through a process called 'active learning', each unlabelled post in the dataset is assessed by the classifier for the level of confidence it has that the post is in the correct category. The classifier selects the post with the lowest confidence score, and it is presented to the human analyst via a user interface of Method52. The analyst reads each post, and decides which of the pre-assigned categories (see Phase 1) that it should belong to. A small group of these (usually around ten) are submitted as training data, and the natural language processing

model is recalculated. The natural language processing algorithm then looks for statistical correlations between the language used and the meaning expressed to arrive at a series of rules-based criteria, and presents the researcher with a new set of posts which it has low levels of confidence for under the recalculated model.

#### **Phase 4: Performance Review and Modification**

The updated classifier is then used to classify each post within the gold standard test set. The decisions made by the classifier are compared with the decisions made (in Phase 2) by the human analyst. On the basis of this comparison, classifier performance statistics – ‘recall’, ‘precision’ and ‘overall’ – are created and appraised by a human analyst.

#### **Phase 5: Retraining**

Phase 3 and 4 are iterated until classifier performance ceases to increase. This state is called ‘plateau’ and, when reached, is considered the practical optimum performance that a classifier can reasonably reach. Plateau typically occurs within 200–300 annotated posts, although it depends on the scenario: the more complex the task, the more training data is required.

#### **Phase 6: Processing**

When the classifier performance has plateaued, the natural language processing model is used to process all the remaining posts in the dataset into the categories defined during Phase 1, using rules inferred from data the algorithm has been trained on. Processing creates a series of new data sets – one for each category of meaning – each containing the posts considered by the model as most likely fall within that category.

#### **Phase 7: Creation of a New Classifier (Phase 1) or Post-processing Analysis (Phase 8)**

Practically, classifiers are built to work together. Each is able to perform a fairly simple task at a very large scale: to filter relevant posts from irrelevant ones, to sort posts into broad category of meanings, or to separate posts containing one kind of key message with those containing another. When classifiers work together, they are called a ‘cascade’. Cascades of classifiers were used in this study. After Phase 7 is completed, a decision is made about whether to return to Phase 1 to construct the next classifier within the cascade, or, if the cascade is complete, to move to the final phase – post-processing analysis.

#### **Phase 8: Post-processing Analysis**

After posts have been processed, the new datasets are often analysed and assessed using a variety of other techniques.

#### **Notes on Training**

##### **Foreign Language**

The scope of this research is geographically localised to France, aiming to identify hate trends in the French language, so any content which was completely in a foreign language was classified as irrelevant to this research. The most common example of this was found in the anti-LGBTQ dataset, as a major political party in Italy uses the acronym PD (Partito Democratico or Democratic Party).

Some posts combined French and another foreign language (mostly English in this case). Posts were considered irrelevant if there were fewer than four words in French.

##### **Information and News**

In certain circumstances, datasets included posts or content from news outlets. As the main focus of this report is organic online discussions, and not retweets of headlines, all posts from established news outlets were categorised as irrelevant.

---

## Examples of Classification

For the four discourses analysed by Method52, we included the following types of post in the hateful category:

- posts containing slurs which were clearly used in a way to cause offence:

 Mais où va t'on si on ne peut plus rien dire sur ces islamo bamboulo bougnoules mais où va t'on? <https://t.co/g31M3vrtE>

slurs	<b>hateful</b>	other
insults	<b>hateful</b>	other

**Translation** 'where is the world going if we can't say anything about the islamo negro ragheads where is the world going?'

 Au secours venez la famille lopez !!! Svp aidez nous contre ces pd de crs qui tabassent des personnes agees .

slurs	<b>hateful</b>	other
insults	<b>hateful</b>	other

**Translation** 'Help, come and help us, the Lopez family!!! Please come and help us fight the faggots in the police who are beating up elderly people'

- posts which were dehumanising and belittling individuals because they belonged to a protected category

 t'façon c tous d gro pd. toi t un mec bien polo écoute les pa.

slurs	<b>hateful</b>	other
insults	<b>hateful</b>	other

**Translation** 'anyway, they're all fags, you're a good guy, don't listen to them'

 Pttddd mais tu fais la frappe or que t'es le genre de grosse chienne à tourner dans les caves wsh t'es une vieille beurette à khel calme toi

slurs	<b>hateful</b>	other
insults	<b>hateful</b>	other

**Translation** 'lol, you're playing the tough woman, but you are the kind of fat bitch who is making the rounds in cellars, you're just an HIV-infected beurette, calm down'

- posts which contained threats or calls for violence (for example Marlène Schiappa, discussed in the report):

 On donne de l'argent à qui on veut grosse pute, par contre quand tu vas recevoir une balle dans la tête ça serra contraire aux valeurs de la république ou pas ?????

12:13 AM - Jan 10, 2019

**Translation** 'You're giving money to whoever you like you fat bitch, but when you get a bullet in your head, will it be against Republican values or not???'

In the non-hateful category, we included posts:

- where slurs were used to call out other users (one of the signs for this was the use of quotation marks)



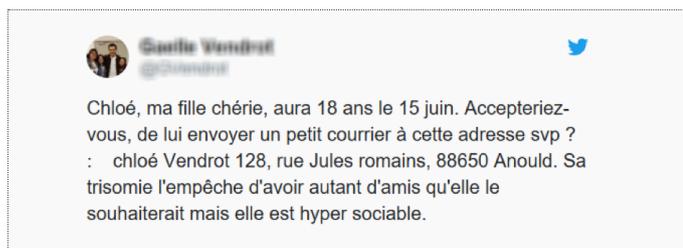
**Translation** ‘In 2019 France, people are still called dirty Arab’

- which re-appropriated hateful terms and slurs: for instance ‘gouine’ (example of gouine in the report in anti-LGBTQ section):



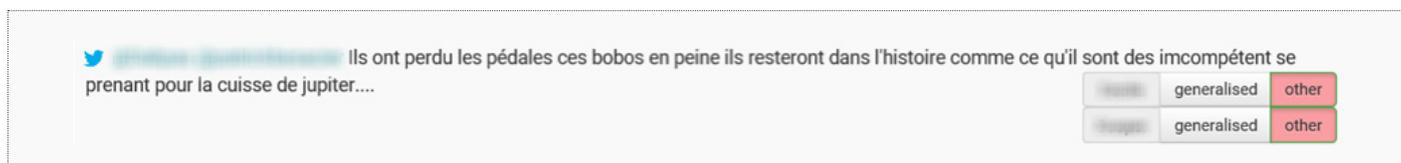
**Translation** ‘dyke is a strong feminist political statement by a group of women who are socially oppressed because of their partial or total rejection of men from their sentimental or sex life. This oppression is costing us our lives’

- that used hateful terms as part of information-sharing: triso often surfaced contents that included the term trisomique (someone with Down Syndrome) in posts about research and information on Down Syndrome, as well as to raise awareness



**Translation** ‘Chloé, my darling daughter, will be 18 on 15 June. Would you be so kind as to send her a card to this address please? Chloé Vendrot, 128 rue Jules Romains, 88650, Anould. Her Down Syndrome prevents her from having as many friends as she would like, but she is hyper sociable’

- irrelevant content that had been missed by the initial relevancy classifier, for example beurre or beur in the anti-Arab dataset and pédale in the anti-LGBTQ dataset



**Translation** ‘These champagne socialists have lost the plot they will go down in history as what they are incompetents who think they’re god’s gift to humanity’

## Generalised and Granular Classification of Misogynistic Speech

The misogynistic dataset was sufficiently large to allow for more granular classification of posts. First, a classifier was trained to identify generalised use of misogynistic terms. A second classifier was then trained to identify targeted hateful speech as well as counterspeech that appeared in the dataset.

In the generalised hateful category, we included:

- content which did not target the community or group that was targeted by the original meaning of the word (e.g., fils de pute, which has misogynistic roots, is applied to a wide range of targets)



**Translation** 'good luck to those who have behaved like sons of bitches all year, let's hope they reap what they have sowed'

- posts where a specific person cannot be identified as being directly targeted with abuse (e.g., posts that contain only the word salope as opposed to a clearly directed phrase like tu es une salope) and where the message does not contain direct threat



**Translation** 'Slut Long Live Marine Le Pen'

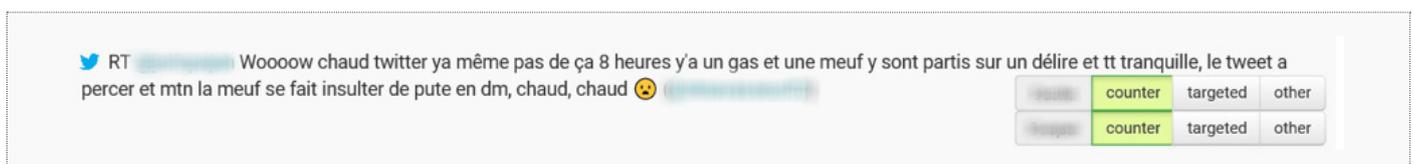
- posts where slurs were used against imaginary or fictional characters.

All other posts were included in an 'other' category.

We then took the 'other' category and separated it using a three-way classifier, which separated this content into counterspeech, targeted hateful speech and other (unclear or irrelevant).

In the counterspeech category, we included:

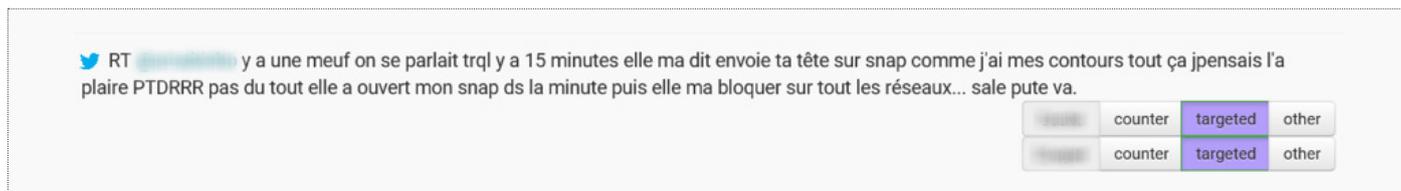
- clear condemnations of the use of slurs and hateful rhetoric



**Translation** 'wow Twitter, tough stuff, less than 8 hours ago a chick and a dude started talking shit and the tweet went viral, now the girl is called a slut, tough stuff, tough stuff'

In the targeted category, we included:

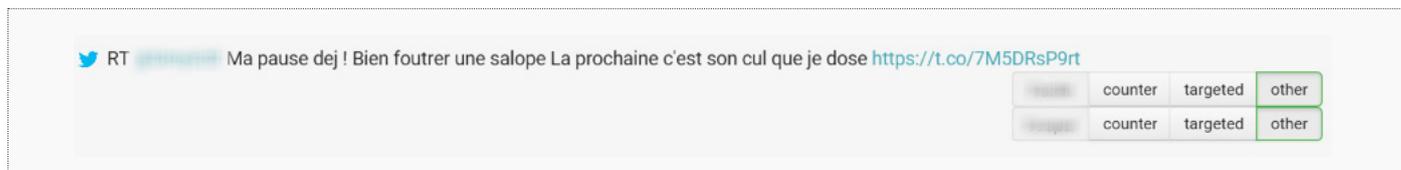
- direct threats (violence, death or rape threats)
- use of slurs and abuse directly targeting a person of the protected category (e.g., tu es une grosse pute when the user targeted can be reasonably identified as a woman)



**Translation** ‘there was a girl, we were talking 15 minutes, she asks me to send a pic on snap, I thought she liked me, she opened the snap within a minute and blocked me on social media...you dirty slut’

In the other category, we included:

- pornographic content which contained misogynistic language
- any remaining irrelevant posts



**Translation** ‘Lunch break! Shagging some slut, next time I’ll take up her the ass’

## Appendix 3: Classifier Accuracy Statistics

Overall accuracy of each classifier is displayed as the probability that a randomly selected post from the dataset belongs to the category in which it was classified.

### Misogynistic

#### Relevancy Classifier

Label	Precision	Recall	FB1	Labelled
Relevant	0.973	0.917	0.944	136
Irrelevant	0.586	0.820	0.683	104
Overall accuracy	0.905			

#### Hateful Classifier

Label	Precision	Recall	FB1	Labelled
Hateful	0.831	0.888	0.858	86
Other	0.500	0.382	0.433	52
Overall accuracy	0.773			

#### Generalised Hateful Classifier

Label	Precision	Recall	FB1	Labelled
Generalised	0.796	0.750	0.772	368
Other	0.500	0.565	0.531	257
Overall accuracy	0.693			

#### Granular Classification

Label	Precision	Recall	FB1	Labelled
Counter	0.889	0.762	0.821	29
Targeted	0.426	0.606	0.500	107
Other	0.867	0.801	0.833	182
Overall accuracy	0.765			

### Anti-Arab

#### Relevancy Classifier

Label	Precision	Recall	FB1	Labelled
Relevant	0.910	0.966	0.937	70
Irrelevant	0.889	0.741	0.808	70
Overall accuracy	0.905			

#### Hateful Classifier 1

Label	Precision	Recall	FB1	Labelled
Hateful	0.462	0.750	0.571	143
Other	0.882	0.682	0.769	175
Overall accuracy	0.700			

#### Hateful Classifier 2

Label	Precision	Recall	FB1	Labelled
Generalised	0.670	0.678	0.674	324
Other	0.750	0.743	0.747	326
Overall accuracy	0.715			

## Appendix 3: Classifier Accuracy Statistics

Overall accuracy of each classifier is displayed as the probability that a randomly selected post from the dataset belongs to the category in which it was classified.

### Anti-LGBTQ

#### Relevancy Classifier

Label	Precision	Recall	FB1	Labelled
Relevant	0.884	0.985	0.932	113
Irrelevant	0.962	0.750	0.843	129
Overall accuracy	0.905			

#### Hateful Classifier

Label	Precision	Recall	FB1	Labelled
Hateful	0.875	0.824	0.848	76
Other	0.447	0.548	0.493	67
Overall accuracy	0.767			

### Ableist

#### Relevancy Classifier

Label	Precision	Recall	FB1	Labelled
Relevant	0.968	0.772	0.859	71
Irrelevant	0.514	0.905	0.655	50
Overall accuracy	0.800			

#### Hateful Classifier

Label	Precision	Recall	FB1	Labelled
Hateful	0.928	0.772	0.847	62
Other	0.184	0.500	0.269	41
Overall accuracy	0.747			

## Endnotes

- 01 [https://europa.eu/rapid/press-release\\_IP-19-805\\_en.htm](https://europa.eu/rapid/press-release_IP-19-805_en.htm)
- 02 [https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online\\_en](https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en).
- 03 'Rapport fait au nom de la commission des lois constitutionnelles, de la législation et de l'administration générale de la république sur la proposition de loi, après engagement de la procédure général de la république sur la proposition de loi, après engagement de la procédure accélérée, visant à lutter contre la haine sur internet (no. 1785) par Mme Laetitia Avia', 19 juin 2019, <http://www.assemblee-nationale.fr/15/rapports/r2062.asp>.
- 04 We derived our methodology for detecting accounts displaying inorganic and potentially bot-like activity from guidelines set out by the Atlantic Council's Digital Forensic Research Lab (DFR Lab) and the Oxford Internet Institute: <https://medium.com/dfrlab/botspot-twelve-ways-to-spot-a-bot-aedc7d9c110c>.
- 05 See, for example: Pete Burnap and Matthew L. Williams, 'Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making', *Policy & Internet*, Vol. 7, Issue 2, 2015.
- 06 <http://www.iicom.org/images/iic/themes/news/Reports/French-social-media-framework---May-2019.pdf>
- 07 The GIFCT was established in 2017 by Facebook, Microsoft, Twitter and YouTube to help formalise industry co-operation to curtail the spread of terrorism and violent extremism online. In May 2019, New Zealand Prime Minister Jacinda Ardern and French President Emmanuel Macron announced the Christchurch Call to Action, following the terrorist attack in Christchurch in March 2019. In response, the GIFCT has created the Content Incident Protocol, which is a 'triaged system aiming to minimise the online spread of terrorist or violent extremist content resulting from a real-world attack on defenceless civilians/innocents'. Read more about the GIFCT and the Content Incident Protocol here: <https://gifct.org/transparency/>.
- 08 <https://www.un.org/disabilities/convention/pdfs/factsheet.pdf>
- 09 <https://rm.coe.int/hate-crimes-against-lgbti/168073dd37>
- 10 Definition based on article 222-33-2 of the French law (Code pénal français), <https://www.legifrance.gouv.fr/affichCode.do?idSectionTA=LEGISCTA000006165282&cidTexte=LEGITEXT000006070719>.
- 11 Debbie Ging and Eugenia Siapera, 'Introduction', *Feminist Media Studies*, Vol. 18, Issue 4, special issue on online misogyny, <https://www.tandfonline.com/doi/full/10.1080/14680777.2018.1447345>.
- 12 <https://rm.coe.int/ecri-glossary/1680934974>.
- 13 <https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680651592>.
- 14 <https://www.adl.org/education/resources/tools-and-strategies/slurs-and-biased-language>.
- 15 For more on the methodology of this index, see: [https://www.cncdh.fr/sites/default/files/cncdh\\_rapport\\_2017\\_bat\\_basse\\_definition.pdf](https://www.cncdh.fr/sites/default/files/cncdh_rapport_2017_bat_basse_definition.pdf).
- 16 Source: CNCDH, *Report on the Fight against Racism, Anti-Semitism and Xenophobia*, 2018.
- 17 Karsten Müller and Carlo Schwarz, *Fanning the Flames of Hate: Social Media and Hate Crime*, 30 November 2018, <https://warwick.ac.uk/fac/soc/economics/staff/crschwarz/fanning-flames-hate.pdf>.
- 18 In June 2019, Facebook accepted to provide French authorities with Internet Protocol addresses to identify individuals posting illegal content on their platform. This was a major shift as up until now Facebook would only share private information for content related to terrorism or child pornography. To access illegal content, French authorities had to go undergo a long and complex process through the American legal system. More information: <https://www.nouvelobs.com/societe/20190625.OBS14918/facebook-fournira-desormais-a-la-justice-francaise-les-adresses-ip-des-auteurs-de-propos-haineux.html>.
- 19 'Rapport fait au nom de la commission des lois constitutionnelles, de la législation et de l'administration générale de la république sur la proposition de loi, après engagement de la procédure général de la république sur la proposition de loi, après

engagement de la procédure accélérée, visant à lutter contre la haine sur internet (no. 1785) par Mme Laetitia Avia', 19 juin 2019, <http://www.assemblee-nationale.fr/15/rapports/r2062.asp>.

20 See here the European Commission Code of Conduct on countering illegal hate speech online, February 2019.

21 For this section the report references Commission nationale consultative des droits de l'homme, *La lutte contre le racisme, l'antisémitisme et la xénophobie (année 2017)*, May 2018, pp. 38–50, <https://www.cncdh.fr/fr/publications/rapport-2017-sur-la-lutte-contre-le-racisme-lantisemitisme-et-la-xenophobie>.

22 Anti-Defamation League, *Quantifying Hate: A Year of Anti-Semitism on Twitter*, 2018.

23 The European Commission Code of Conduct on countering illegal hate speech online, February 2019, which is mentioned by the Loi Avia context report to identify online trends of hate speech, does not mention hateful content which targets women and disabled people. The Code of Conduct relies on Framework Decision 2008/913/JHA of 28 November 2008, which includes 'all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin'.

24 Kimberlé Crenshaw, 'Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics', *University of Chicago Legal Forum*, Vol. 1989, Issue 1, pp. 139–67, <https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1052&context=uclf>.

25 'Intersectionality highlights the flaws in discrimination laws which focus on one ground at a time... focusing on single grounds at a time ignores the fact that everyone has an age, a gender, a sexual orientation, a belief system and an ethnicity; many may have or acquire a religion or a disability as well... The aim of intersectionality should be to capture and address the wrongs suffered by those who are at the confluence of all these relationships.' For more please find here the report European Commission, *Intersectional Discrimination in EU Gender Equality and Non-discrimination Law*, May 2016.

26 See European Commission, *Intersectional Discrimination in EU Gender Equality and Non-discrimination Law*, May 2016.

27 See Loi du 29 juillet 1881 sur la liberté de la presse, including articles 24, 29, 32, 33, <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006070722>.

28 Addition made in 1972: la diffamation et l'injure sont punies plus sévèrement lorsqu'elles sont commises 'à raison de leur origine ou de leur appartenance ou de leur non appartenance à une ethnie, une nation, une race ou une religion déterminée'.

29 These principles were transferred to criminal law with articles R 625-7 and R 625-8. of the Criminal Law code (Code pénal français). This is important as the 1881 freedom of press law only applies in public spaces (in the age of the internet anything published on social media which is not public would not fall under the 1881 law); with the criminal code, content shared in the private space also falls under such regulations.

30 Insults (or 'injure' in French) is understood under article 29 of 29th of July 1881 law as an outrageous expression, term of contempt or invective which does not rely on a fact (French version 'toute expression outrageante, termes de mépris ou invective qui ne renferme l'imputation d'aucun fait').

31 Defamation is defined under article 29 of 29 July 1881 freedom of press law as an allegation to facts which could compromise the honour or consideration of a person or to its body (French version: 'toute allégation ou imputation d'un fait qui porte atteinte à l'honneur ou à la considération de la personne ou du corps auquel le fait est imputé').

32 For more information please find here Nathalie Droin's article 'L'appréhension des discours de haine par les juridictions françaises: entre travail d'orfèvre et numéro d'équilibriste', 2018.

33 *Ibid.*

34 From the French 'La religion la plus con, c'est quand même l'Islam', for more information please see TGI Paris, 17ème ch., 22 octobre 2002, *Légipresse*, 2003, nos. 198-I, p. 12.

35 From the French 'pour moi les juifs, c'est une secte, une escroquerie', for more on this please find Cass. crim., 15 mars 2005, n°

04-84.463, *Bull. crim.*, no. 90; *Dr. pén.*, 2005, *Comm. no. 85*, note M. Véron.

36 This research was based on the analysis of 15,000 randomly chosen comments, from among 15 million comments extracted from 25 Facebook pages. For more information on the complete methodology, please see the methodology section (page 1) at <https://netino.fr/panorama-de-la-haine-en-ligne-2019/>.

37 The definition of insult used in this report includes more than hateful speech directed towards protected categories and includes generalised insults like 'fuck you'.

38 In this research, among the 2,964,271 tweets collected over the month of January 2017, 2,950 tweets were randomly selected and classified; for more details on this please find the methodology section here: <http://www.idpi.fr/wp-content/uploads/2018/02/2018-02-Baromètre-IDPI-Janvier-2017.pdf>.

39 See *Baromètre mensuel des manifestations de la haine en ligne – Janvier 2018*, IDPI, for more information on methodology and definitions.

40 Sylvain Mossou and Andrew Lane, *Anti-migrant Hate Speech*, Quaker Council for European Affairs, 2018, [http://www.qcea.org/wp-content/uploads/2018/06/Hate-Speech-Report\\_final.pdf](http://www.qcea.org/wp-content/uploads/2018/06/Hate-Speech-Report_final.pdf).

41 Anti-Defamation League, *The Online Hate Index*, 2018. This study was based on the analysis of 9,000 Reddit comments randomly chosen and coded among 80,000 pulled (manually coded into categories 'hate' or 'non-hateful').

42 *Encyclopedia of Political Communication*, Sage, 2008.

43 See methodology section of Jamie Bartlett, *Misogyny on Twitter*, Demos, 2014. Two studies were conducted: the first pulled out tweets geographically located in the UK using 'rape'. Among the 138,662 tweets, after relevancy classifier 108,044 were left; from those 500 were randomly selected for analysis. In the second study keywords used were 'slut' and 'whore' with 161,744 tweets found in the UK; after relevancy classifier 131,711 tweets remained, with 500 then randomly selected for the analysis.

44 Carl Miller and Josh Smith, *Anti-Islamic Content on Twitter*, Demos, 2017; study based on 143,920 tweets which were collected as found 'derogatory towards Muslims'.

45 Similar to anti-white discourse, in Western societies, the concept of anti-Christian discourse is often evoked by fringe groups to legitimise their anti-immigrant agendas, and by searching for this discourse we aimed to ascertain if this is actually a significant portion of hateful discourse online.

46 The concept of anti-white hateful speech is often leveraged by the far-right as proof of anti-white sentiment from minority groups and in some cases proof of ethnic cleansing or 'white genocide'. Part of this scoping endeavoured to understand to what extent this type of discourse exists online and to demonstrate its relative (in)significance online compared with other types of hateful speech.

47 With the exception of the anti-Arab algorithm. See section 'Anti-Arab Discourse' for more detailed discussion.

48 The key limitations are that potentially irrelevant content cannot be separated from the sample, and detailed language analysis cannot be conducted to distinguish between hateful and non-hateful uses of certain keywords.

49 [https://www.lemonde.fr/societe/article/2019/02/11/la-justice-saisie-a-cause-de-plusieurs-tags-antisemites-a-paris\\_5422154\\_3224.html](https://www.lemonde.fr/societe/article/2019/02/11/la-justice-saisie-a-cause-de-plusieurs-tags-antisemites-a-paris_5422154_3224.html).

50 With the exception of the anti-Arab algorithm. See section 'Anti-Arab Discourse' for more detailed discussion.

51 Burnap and Williams, 'Cyber Hate Speech on Twitter', 2015.

52 For further discussion on the accuracy of natural language processing algorithms and their application to social media, see Jamie Bartlett, *Vox Digitas*, Demos, 2014.

53 The Great Replacement theory is the belief that 'white European populations are being deliberately replaced at an ethnic and cultural level through migration and the growth of minority communities. This propagation often relies on demographic projections to point to population changes in the West and the possibility that ethnically white populations are becoming minority

groups'. See ISD's *Great Replacement Theory* report.

54 For further information on Renaud Camus and the Great Replacement theory see [https://www.lemonde.fr/les-decodeurs/article/2019/03/15/la-theorie-du-grand-replacement-de-l-ecrivain-renaud-camus-aux-attentats-en-nouvelle-zelande\\_5436843\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2019/03/15/la-theorie-du-grand-replacement-de-l-ecrivain-renaud-camus-aux-attentats-en-nouvelle-zelande_5436843_4355770.html).

55 Jacob Davey and Julia Ebner, 'The Great Replacement': *The Violent Consequences of Mainstreamed Extremism*, ISD, 2019, <https://www.isdglobal.org/isd-publications/the-great-replacement-the-violent-consequences-of-mainstreamed-extremism/>.

56 <https://www.bfmtv.com/societe/a-grasse-une-boulangerie-soupconnee-de-vendre-des-gateaux-racistes-867194.html>.

57 Anti-white keywords were included in this report as anti-white racism is often cited by far-right extremists as a justification for anti-immigrant prejudices. Part of the reason behind the inclusion of anti-white keywords in this research report was to understand to what extent, if any, this type of racism is expressed online.

58 *En finir avec l'impunité des violences faites aux femmes en ligne : une urgence pour les victimes*, [http://haut-conseil-egalite.gouv.fr/IMG/pdf/hce\\_rapport\\_violences\\_faites\\_aux\\_femmes\\_en\\_ligne\\_2018\\_02\\_07-3.pdf](http://haut-conseil-egalite.gouv.fr/IMG/pdf/hce_rapport_violences_faites_aux_femmes_en_ligne_2018_02_07-3.pdf)

59 See Crenshaw, 'Demarginalizing the Intersection of Race and Sex', 1989.

60 This inter-ministerial mission team comprises seven high-level experts and three permanent reporters from a range of ministries (Culture, Interior, Justice, Economy, Prime Ministerial services); DILCRAH (Inter-ministerial Delegation to Combat Racism, Antisemitism and Anti-LGBTQ Hate); DINSIC (Inter-ministerial Delegation of Digital and Information and Communication Systems); and independent administrative authorities ARCEP (electronic communications and postal authority) and CSA (audiovisual regulatory authority). This team worked with Facebook throughout January and February to explore how a new system to regulate social networks could be established to complement existing instruments and better achieve public policy objectives relating to the reconciliation of public freedoms and the safeguarding of public order on social networks.

61 <http://www.iicom.org/images/iic/themes/news/Reports/French-social-media-framework---May-2019.pdf>.

62 *Ibid.*

63 Jacob Davey, Jonathan Birdwell and Rebecca Skellett, *Counter Conversations*, ISD, 2018, <https://www.isdglobal.org/isd-publications/counter-conversations-a-model-for-direct-engagement-with-individuals-showing-signs-of-radicalisation-online/>.

64 [https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online\\_en](https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en).

---

ISD London | Washington DC | Beirut | Toronto  
Registered charity number: 1141069

© ISD, 2019. All rights reserved.

Any copying, reproduction or exploitation of the whole or any part of  
this document without prior written approval from ISD is prohibited.  
ISD is the operating name of the Trialogue Educational Trust.

[www.isdglobal.org](http://www.isdglobal.org)

