DCU

Hate Track

Tracking And Monitoring Racist Speech Online

Eugenia Siapera, Elena Moreo, Jiang Zhou

Coimisiún na hÉireann um Chearta an Duine agus Comhionannas
Irish Human Rights and Equality Commission

IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn

DCU School of Communications, Institute for Future Media and Journalism (FuJo), Anti-Bullying Centre (ABC), and Insight Centre for Computer Analytics, School of Computing.

ABSTRACT

The HateTrack Project is an experimental, exploratory research project that combines social, scientific and computational methods to understand online racist speech in the Irish context. The project used insights from civil society and experts in the field of race, racism and hate speech to build a computational tool that harvests and classifies Facebook and Twitter posts in terms of their probability to contain racially-loaded toxic contents. The tool is designed as a monitoring and diagnostic tool of the state of the Irish digital public sphere. While it is currently focused on racially-toxic contents, it can be scaled to other forms of hate and toxicity, such as misogyny and homophobia. Using HateTrack, we generated a dataset which was subsequently analysed in terms of the toxic repertoires it contained, the communities targeted, the kinds of people posting, and the events that trigger racially-toxic contents. Finally, we held workshops with students to identify their views on reporting racist hate speech online.

Contents

Acknowledgements

Acknowledgements

This project has provided us with a great opportunity to learn more about this topic which we strongly believe is of crucial social relevance. We are sincerely grateful for the funding that the Irish Research Council and the Irish Human Rights and Equality Commission made available for the project. We would also like to take this opportunity to thank the members of the Steering Committee for their time and input in this project: Ronit Lentin (TCD), Shane O'Curry (ENAR Ireland), Eoin O'Dell (TCD), James O'Higgins-Norman (DCU), and especially Walter Jayawardene from IHREC. We are also extremely grateful to Suzanne Little from the School of Computing at DCU, whose help and guidance in developing HateTrack has been invaluable. None of this would be possible without our key informants, representing anti-racist and refugee/migrant support organisations in Ireland, academics, and media professionals. We also wish to thank the students who attended our presentations at DCU, Colaiste Dhulaigh and Trinity College, for their feedback, challenging questions, and for sharing their views on reporting. We hope that we have made justice to their many points, comments, observations and concerns.

Overview and Key Findings

Overview and Key Findings

This report presents the findings of the 12-month HateTrack project, conducting research into developing and implementing a ***machine learning*** [see Glossary and Definitions] tool for the monitoring and study of online racist hate speech. The tool is conceived and designed as a diagnostic tool, seeking to identify the current state of the Irish ***digital public sphere*** [see glossary and definitions]. It is not intended as a censorship or removal tool, but as a means for tracking racist discourses and gaining a better understanding of their trajectories across the digital public sphere. The tool can be potentially extended to include different kinds of hate speech, for example misogynistic or homophobic speech. The project operated in four stages. The first stage involved a qualitative study discussing racism and hate speech with anti-racist and community-based organisations, as well as with academic and other experts. In the second stage of the research, these discussions fed into the coding of a manually collected corpus of online materials found on Facebook and Twitter. This coded corpus formed the 'ground truth' used to train the ***algorithm*** [please see Glossary and Definitions] that classifies online contents harvested from Facebook and Twitter. This constituted the second stage of the research, which was mainly undertaken by Dr Jiang Zhou. Dr Zhou developed HateTrack, a computer application that collects and classifies online contents in terms of their probability to contain 'racially-loaded toxic speech'. In the third stage of the project, we used HateTrack to collect a larger dataset, comprising over 6,000 entries, which were subsequently analysed further to pick up on the racially-loaded repertoires formed, the communities targeted, the types of people posting such contents, and the events that seem to trigger these. In the fourth and final part of the research, we sought to identify what kinds of contents people tend to report using the social media platforms' flagging systems, and how they justify these decisions. The key findings are outlined below.

Stage I: Defining racist hate speech: findings from the focus groups and interviews

- **Online racist speech is pervasive but it is not all the same**. It can be thought of in terms of a continuum, with extreme, vicious and overt racist speech occupying one end and a subtler, more masked kind of racist speech occupying the other end. Instances of extremely racist speech that dehumanise, demean and clearly mean to belittle and discriminate are easy to identity. On the other end, we encounter instances of coded racist speech, that are less clear and more difficult to decode, but which are equally problematic as they too 'racialise' and through this seek to subjugate those targeted.
- Processes of adaptation and learning mean that **racist speech is dynamic and evolving**, often using tropes such as slang, circumlocutions (speaking around something, being evasive and vague), irony, and ambiguity.
- Variants of racist discourses include '**whataboutery'** (e.g. 'what about our own'), **narratives of elsewhere** (e.g. 'look at Sweden'), **use of bogus statistics** (e.g. '80% of Africans are unemployed'), and **metonymies** (substituting a word with something closely related, here in an ironic sense, for e.g. 'religion of peace' to refer to Islam typically used with a view to associate Islam with violence).
- **Civil society is primarily concerned with the impact that racist and racialising speech** has on those targeted, such as harm, exclusion, a chilling effect, but also material losses, as certain people who could use the digital sphere to generate income, for example on YouTube or through their online writing, are now avoiding placing themselves in potentially harmful and traumatic situations.
- **Online racist hate speech cannot be understood in isolation from racist structures and institutions**, and from media and political discourses that racialise certain groups.

Stage II: Operationalising the definitions and building HateTrack

- Racially-loaded toxic speech: we develop this compound term to capture the different forms and intensities of racist speech.
- Racially-loaded toxic speech is defined as language and contents that entrench polarisation; reinforce stereotypes; spread myths and disinformation; justify the exclusion, stigmatisation, and inferiorisation of particular groups; and reinforce exclusivist notions of national belonging and identity.
- Rather than using Naïve Baynes, Method52 or other 'hand-crafted' models, HateTrack builds on deep neural network techniques, and specifically on the Long Short-Term Memory (LSTM) network.
- This method can potentially extend to cover other forms of hate contents, for example misogyny and homophobia.
- HateTrack can harvest Facebook comment threads and Twitter posts, based on account handles or keywords. The

tool classifies posts in terms of their probability to contain racially-loaded toxic speech (1=high, 0=low). Users can select, save and download contents in a spreadsheet format.

- The downloaded data is anonymised so that it does not contain any information on those who posted the information. The tool can be further refined through manually coding contents and saving the classification.

Stage IIIa: Dataset analysis: racially-loaded toxic contents in the Irish digital sphere

- Crude and coded forms of ***racially-loaded toxic contents*** [for full definitions see the glossary and definitions section] utilise different discursive strategies (including grammar, semantics, style of argumentation).
- Crude forms typically employ insults, slurs, profanity, animal comparisons, direct denigration, or appeals to well-entrenched racial stereotypes or long debunked 'race science' myths.
- Coded forms rely on supposedly race-neutral principles like culture, values, ethnicity, and tend to employ seemingly well-reasoned or common sense arguments, for example, 'taking care of our own' or distinguishing between 'deserving' and 'undeserving' groups.
- Racially-loaded toxic discourses often coalesce around notions of 'Irishness' and what it means to be Irish, which is constructed as exclusively White and Christian.
- Calling out racism in online environments typically leads to accusations of being over-sensitive or 'playing the race card', or 'being racist' against white people.
- There are clear patterns of shared language between international and Irish groups, including the adoption of racist ideologies produced in the context of the United States and the European Identitarian movement. Key terms include 'white genocide' and 'population replacement' and the localised term 'new plantation'.
- Racially-loaded toxic discourses feed on fake news and bogus statistics revolving around the alleged failures of multi-culturalism, no-go Muslim areas, and African youth gangs terrorising locals.
- Social media pages of news outlets seem to play an important role in channelling racially-loaded toxic contents through the comment threads on their posts. The way mainstream media frame and present news has an impact on the comments left.
- Expressions of racism online are punctuated with misogynist, homophobic, and transphobic attacks directly targeting women and members of the LGBT community.
- Social media affordances and tropes lend themselves to racially-loaded toxic contents, which can include memes, multimedia materials, hashtags, tagging and other forms that allow the materials to travel further.

Stage IIIb: Targeted communities

- Anti-immigrant and anti-refugee discourses revolve mainly around three inter-related tropes: access to welfare and housing; moral deservedness; and the good versus bad immigrant trope.
- Anti-Muslim discourses mobilised four tropes: terrorism; clash of civilisations; Muslim men as misogynist and sexually deviant; and a general and unspecified antipathy.
- Typically, Traveller and Roma people are targeted as undeserving, 'uncivilised', thugs and criminals; they can further be targeted using a dehumanising language.
- Jewish people are targeted as hidden figures, globalists scheming behind the scenes; as Shylock, devious merchants and userers; as 'unassimilable'; through denying the importance and magnitude of the Holocaust.
- Black people are targeted in the anti-refugee/migrant discourses, in the anti-Muslim/Islamophobic ones, as well as the attacks against second generation Irish people. But it is important to further identify the specific ways in which Black people are targeted as such. Some of the ways we identified in our dataset include the trope of criminality; the trope of being 'uncivilised', lazy, 'parasites'; and the dehumanising trope of African men as animals.
- Second-generation Irish people are targeted through the trope of population replacement or colonisation; and through making a distinction between 'real' Irishness, which is an outcome of both a 'biological' and a 'cultural' bond and Irish citizenship which is a kind of 'fake', 'paper' Irishness.
- ***Trigger events*** [see glossary and definitions]: while there is a constant undercurrent of racially-toxic contents in circulation at any given time, we identified three types of trigger events: exceptional, one-off events, for example, the case of Ibrahim Halawa or the stabbing in Dundalk that trigger a high volume of racially-toxic contents; topics that touch upon social tensions, for example, housing and welfare; and finally, topics that explicitly thematise questions of the nation, 'race' and culture, for example, of refugees and migration, Direct Provision, anything about Travellers and Roma, or Islam.
- Who is posting racially-toxic contents? The main distinction we identified is between people versed in a particular ideological and political language and discourse, and those who merely reproduce 'racial common sense'.

Stage IV: Reporting cultures

– The main finding in this part of the research is that people tend to under-report online racist speech. We identified four repertoires that may act as barriers: freedom of speech/expression as an absolute; racist speech is only uttered by people who are not worth dealing with; reporting is pointless because there is so much racist speech online; and a 'bystander' effect or disavowal of responsibility, with some respondents feeling that it was not their job to report anything.

Discussion Points for Policy

HateTrack is a small exploratory project that does not allow for concrete policy recommendations to be made. Nevertheless, we identified some relevant points for discussion that can help guide policy thinking.

1. The benefits in tackling online hate and racism are not solely felt by individuals and groups targeted but are likely to benefit the entire online community by ensuring that online spaces remain civil, safe, and democratic. The toxicity of hateful comments has a ripple effect felt across society and not only by those immediately targeted.

2. In contrast, and despite its broader toxic effects, the burden for dealing with racially-toxic speech falls disproportionately on those targeted by it. Making online racist speech an issue for the whole society to deal with will mark an important step forward. Developing understandings of digital citizenship that include codes of ethics of online behaviour and responsibility to others can be part of this.

3. Building up on this, a point that was raised repeatedly during the focus groups and interviews is that the examples set by public figures, the media, and the Garda Síochána can have a powerful effect on how victims of online abuse feel. Public commitment by a variety of key actors to counteract online racism and take racist incidents seriously can help minimise some of the toxic effects of online hatred.

4. It is noted that online racially-toxic speech cannot be countered on its own and in isolation from other forms of racism. This point emerged very clearly from discussions with anti-racist and community groups. Those representing some of the communities targeted made references to ongoing discrimination, exclusion and aggression in many face-to-face contexts, and felt that all these have to be addressed in tandem.

5. Social media platforms already rely on trusted partners – NGOs and various organisations – that promptly flag problematic or hateful content. These types of collaborations could be extended and mainstreamed.

6. To effectively neutralise racially-loaded toxic contents, counter-speech has to be tailored to the specific points made by the discourses or repertoires identified.

7. As young people increasingly use the internet as a library and Facebook and Twitter as sources of news, it is important that digital media literacy become a key part of the curriculum and that educators help young people to develop critical thinking about race and racism. Multiple literacies of digital media, social justice, and anti-racism can help minimise what Daniels (2008: 146) calls 'epistemological vulnerability', that is, the susceptibility of young people to hateful arguments.

Introduction

Evil settles into everyday life when people are unable or unwilling to recognise it. It makes its home among us when we are keen to minimise it or describe it as something else

TEJU COLE, 2016

Introduction

Social media facilitates the rapid spread of ideas online, and hate speech is no exception. Neo-Nazi, Far-right, and fascist groups have all capitalised on social media's broad reach, easy access, and anonymity to spread racist, homophobic and misogynist rhetoric through targeted online posts, videos, forum discussions. While explicit Islamophobic, xenophobic, anti-Semitic groups may be responsible for much of what would be unequivocally considered hate speech, online racism is not solely the preserve of groups with a marked ideological profile, but proliferates in a variety of more or less coded guises and through 'everyday' discourses (Essed, 1991). Researchers and media analysts have agreed that events like the 2016 US Presidential election, the 'refugee crisis' in Europe, and the Brexit referendum have resulted in a worrying escalation of racist hate speech and racist incidents. [1]

Many of our most important public and civic spaces exist online and the capabilities deriving from social media platforms to shape public attitudes are immense. The proliferation of such platforms has created an entirely new frontier in thinking about and addressing racism, bringing up challenges in terms of how online racism should be defined and whether the notion of hate speech can capture the unstable and adaptable nature of racist discourses – the 'motility' of racism, in Lentin's (2016) definition. Governments, IT companies, and civic society groups have focused their efforts on counteracting the effects of an increasingly toxic online environment on civic life and the broader public sphere, seeking to strike a balance between safeguarding the core tenets of freedom of speech and defending the rights of individuals and groups not to be subjected to vilification, the threat of violence, and abuse. However, the tools for addressing the challenges of hate speech and fake news have been at times inadequate to the task or ineffective.

Although research has shown that forms of racism such as Afrophobia (Michael, 2017), Islamophobia (Carr, 2015), anti-Traveller and anti-Roma racism (Twomey, 2017) are present in Ireland, we have less information about the online domain.

There is little research on the nature and distribution of online hate in the Irish context and the existing data is based on reported racist incidents. According to ENAR Ireland's report[2], 111 cases of racist hate speech online were lodged through the iReport mechanism between January and June 2017. Eighty-two of these incidents related to content published on Facebook (37), Twitter (35), and YouTube (10). The report found that over half of the Facebook posts reported as racist were published on the pages of named groups alongside other explicit white supremacist, racist and anti-refugee content and met the criteria under Irish law for Incitement to Hatred. Twelve other reports concerned racist speech on Facebook on personal pages. The report also found an increase in the number of organised groups reported and evidence that some of these are linked to groups already prosecuted for incitement to hatred and racist crimes in other countries. The analysis also pointed to the existence of a small number of anti-refugee groups that, while claiming to express 'concerns' about the number of asylum seekers in Ireland, stoke up hatred towards refugees and asylum seekers through falsified stories, memes, and outright racist and supremacist language. Ultimately, the report highlighted a link between racist harassment and hateful speech on Twitter, with Irish Twitter users being directly harassed, attacked, and bullied online by other accounts based in the US, UK, Australia and other locations. While ENAR has done important work through the publication of bi-annual reports, they can only present data that are based on incidents reported by victims or bystanders. The lack of a comprehensive and systematic mechanism for monitoring or collecting hate speech in Ireland means that there remains a paucity of information on its scale, features, and possible effects.

The focus of existing policies by both states and social media corporations are oriented towards improving efficiency in terms of the time it takes to take down hate contents, but there is little if any understanding of what constitutes hate speech and what may motivate users (victims and bystanders) to report some materials but ignore or simply block others. Without knowing the barriers to the reporting process and the reasoning behind reporting online hate, it is difficult to obtain an understanding of the nature of what is reported and what stays online, its severity, spread and frequency. In short, we lack a benchmarking study that will help establish the scope

1 For example, the Southern Poverty Law Centre found, between November 8th and December 8th 2016, more than 1,750 photos and memes demonising Islam and Muslims or attacking public personalities like Angela Merkel or Mayor of London Sadiq Khan. Worrying "spikes" of Islamophobic hatred were detected in the wake of the terrorist attacks in Paris, Brussels and Nice. For a few hours after the Paris attacks, #matadatodoslosmusulmanes ("kill all Muslims") became the third most used hashtag in Spain (Jubany and Roiha 2016).

2 The report is found at: http://enarireland.org/wp-content/uploads/2018/01/iReport_1516_jan-jun2017.pdf

and effectiveness of codes of practice and reporting systems compared to what circulates in social media platforms. In other words, in order to develop appropriate policy, we need to understand the various types of racist speech circulating in online environments.

An additional element concerns the overall structure and quality of the informational ecosystem, or what we refer to here as the **digital public sphere**. The past year has seen discussions of 'fake news', the use of bots, the reckless use of private data by social media companies and other issues of public concern. It is at this level that we locate the issue of racist hate speech and online racism more broadly. Democracies rely on a healthy public sphere, which is open to all, and which enables people to present, deliberate and exchange views on matters of interest thereby formulating an opinion (Habermas, 1992). While in earlier times the mass media were the main platforms for the public sphere, social media has now taken over; it is important therefore to examine the health and operations of this digital public sphere and the extent to which it is able to fulfil its functions. It is in this context that the HateTrack project was developed.

HateTrack Research and Aims

The HateTrack project sought to address these gaps with a view to contributing to opposing racism and creating a more inclusive online environment and in general improving the quality of the informational ecosystem. Specifically, the project sought to address **three related aims.** The **first aim** of the project is to develop a methodological tool for the identification, collection and tracking of racist hate materials on public Facebook pages and on Twitter. The **second aim** of the project is to generate a preliminary dataset of online hate materials from public Facebook pages and Twitter accounts collected over a period of three months. The dataset can then be used to identify the range of racist repertoires circulating in the context of Ireland. The **third aim** is to explore some of the reporting barriers and cultures that feed into decisions to report or not report online racist hate speech.

This project is not the only one to have studied racist hate speech online. European initiatives, such as the work undertaken by the UK-based Centre for Analysis of Social Media of the think-tank Demos[3], the EC REC (Rights, Equality and Citizenship) Mandola project[4], as well as the newly funded Hatemeter project[5], have developed their own computational approaches to the study of hate speech (see more details on p. 17). However, the present project departs from these in two significant dimensions: firstly, rather than relying on formal and legal definitions of hate speech, it undertakes original research with anti-racist activists and members of targeted communities in order to explore their experiences and own understandings of racist speech; secondly, it focuses specifically on the Irish context, and provides an in-depth qualitative analysis of online racist speech.

The research is limited to two social media platforms: Facebook and Twitter. Ireland is a high internet usage country, with over 89% of households having access to the internet at home[6]. Facebook is the platform of choice, with 64% of internet users in Ireland having an account. Twitter is the second most popular platform with 28% internet users having an account[7]. In terms of actual numbers, 2.2 million people in Ireland have a Facebook account, and 835,000 have a Twitter account[8]. These platforms, therefore, host a sizeable part of the population of Ireland and this is why they were selected for this project. However, it should be noted that although these platforms are extensively used by people in Ireland, they do not represent Irish society in its entirety. Additionally, these platforms are not used by all social groups. Younger, better educated, more affluent groups are more likely to use social media platforms. For example, only about 12% of the total number of users are in the 45-54 age category. Moreover, 6 out of 10 users may visit but do not post or comment on anything[9]. All this means that the present study cannot be taken as a comprehensive study of the digital public sphere, but a limited study of some of the contents found on these two platforms.

We conceptualised this research project as an exploration of the co-articulation of racism with social media. It is undertaken in an experimental manner, to examine if a bottom up definition of problematic, toxic and hateful contents can be operationalised in a manner understood by computers, and if this can be used to collect further data for analysis. Throughout this research, we were mindful of ongoing discussions on freedom of expression and censorship, a concern that was repeatedly expressed by the participants in this study. The computational tool is, therefore, conceived as a means for collecting and classifying online discourses and not as a tool for removing contents. In this manner, the HateTrack tool constitutes a monitoring and diagnostic tool that can hopefully aid civil society and academic researchers in deepening their

3 https://www.demos.co.uk/research-area/centre-for-analysis-of-social-media/

4 http://mandola-project.eu/

5 https://ict.fbk.eu/projects/detail/hatemeter/

6 Source: Central Statistics Office, http://www.cso.ie/en/releasesandpublications/er/isshh/informationsocietystatistics-households2017/

7 Source, ISPOS MRBI, https://www.ipsos.com/sites/default/files/ct/news/documents/2017-10/Social-Networking-Aug-17.pdf

8 Source: Consumer Barometer, GlobalWebIndex 2016 Q4, IPSOS MRBI, http://connector.ie/infographic/

9 Source: http://connector.ie/infographic/

understanding of current racist discourses and develop appropriates responses to these.

In combining social, scientific, qualitative methodology with computer science, the project has a clear interdisciplinary character. Its approach is uniquely innovative in that this is a project primarily driven by social scientists who undertook original research to then inform the computer scientists and play the role of the so-called 'ground truth'. A similarly innovative approach was used in the interpretation of the data generated by the tool, which relied on grounded theory and discourse analysis. It was especially challenging to understand and contextualise the problematic discourses we encounter in the Irish digital public sphere in terms of broader, global and transnational discourses associated with the rise of the so-called alt-right and identitarian movements. Similarly, the combination of racist, misogynist and anti-feminist, as well as homophobic discourses required a sound grounding in feminist theory and theories of intersectionality.

In short, the HateTrack project is the first of its kind to seek to gain an overall understanding of hate speech in Irish social media through a mixture of automated techniques and discourse analysis. To operationalise participants' definitions and collect examples to design and train the algorithm has proved challenging on many levels. The automated identification of hate speech is difficult: firstly, there are a multitude of semantic combinations and codes for channelling racist ideas, without the need to use insults, slurs and other expressions which a programmer can anticipate. Secondly, an algorithm may be unable to identify the nuances of the context within which a statement is made (Bartlett et al. 2014: 25). To address these challenges, we opted for a mixed and multi-faceted methodological approach combining qualitative methods – focus groups and interviews – algorithmic techniques for harvesting data; and content discourses analysis of a selection of the data set.

This report proceeds by explaining the theoretical approach to hate and racist speech; the research design and methodology of the project; and the findings of the analysis of the dataset generated using the HateTrack tool. The final section presents the findings of a series of workshops we conducted with students on the question of reporting problematic contents.

Theoretical Approach

Theoretical Approach

This section begins with a discussion of the literature on race and racism in Ireland. It will then outline the notion of hate speech, before embarking on a discussion of the dilemmatic construction of hate speech versus freedom of expression.

Race and Racism in the Irish Context

In discussing race and racism, the first issue that needs to be addressed concerns the very notion of race. The broader theoretical framework adopted here is influenced by Critical Race Theory, which views race and racism as a historically and geographically specific socio-political system (Goldberg, 1993). In understanding how this operates in practice, we draw upon the work of a variety of scholars, most notably Essed (1991) and van Dijk (1993). These authors understand racism not simply as an 'ethnocentric dislike and distrust of the Other' but rather as an ideology and political project: racism emerges at the point where 'differences become essentialised as hard-wired, biological attributes of particular individuals and groups and thus mobilised to justify systems of discriminatory practices, structures and institutions against them (Fredrickson, 2002: 5). While we are aware that anti-Semitism, Islamophobia, and xenophobia have distinct characteristics, in this study we subsume them under the category of racism. We further understand racism as a "scavenger ideology" (Mosse, 1985) to suggest that while racism evolves, adapting to current socio-historical contexts and capitalising upon new techno-cultural affordances, it relies on the same process involving naturalisation of supposed biological or cultural characteristics, using them to justify subjugation. Racism is reproduced through everyday, routine, banal interactions, and especially through discursive interactions – ranging from informal conversations, to elite discourses, and online discussions.

In the Irish context, the main expressions of racism include anti-black racism (Michael, 2017), Islamophobia (Carr, 2015; 2017) and anti-Traveller racism (Twomey, 2017), though there are reports of hate crime against LGBT people as well (Sheehan and Dwyer, 2017; Giambrone, 2017). Additionally, recent research has uncovered a focus on whiteness, which appears to 'racialise' migrant groups, including those coming from the EU (Joseph, 2017). Michael (2017) draws upon the ENAR iReport system which she co-designed to discuss the victimisation of persons identifying as Black in both public and intimate spaces, such as their own homes. Michael notes that the experiences of some Black people include persistent low-level harassment that creates a threatening environment and that occasionally erupts into violence. Victims report the use of racial stereotypes such as 'dirty' or 'lazy', references to disease such as Ebola, references to animals and biological inferiority and so on. Islamophobia or anti-Muslim racism proceeds by way of ***racialisation*** [see also glossary and definitions], i.e. through rendering Muslims a racial group always associated with certain characteristics and primarily through positing the essential incompatibility of Muslimness and Irishness; additionally, Islam is typically associated with terrorism, while Carr (2017) notes a gendered element in Islamophobia, in women reporting anti-Muslim hostility in greater numbers than men (44% compared to 28% respectively). Travellers, whose recognition as a distinct ethnicity occurred in Ireland only in March 2017, are equally racialised and subjected to a number of stigmatising stereotypes as criminal, undeserving, dishonest, and immoral (Joyce et al. 2017). Relying on employment statistics and semi-structured interviews, Joseph (2017) identified the operation of a racial hierarchy in the Irish labour market, with White settled Irish on top, followed by other EU white people, namely Spanish and Polish, and with Nigerians at the bottom. Joseph's work is suggestive of a particular dynamic at play whereby migrants are subjected to racialisation processes, but with 'whiteness' still prevailing. An IHREC-ESRI study (McGinnity et al., 2017) supports Joseph's findings, reporting that while women and older people tend to report more instances of workplace discrimination, Black respondents reported higher discrimination across the workplace, public service and private services. Travellers were ten times more likely to report discrimination in seeking employment and 22 times more likely to experience discrimination in private services, for example in pubs, restaurants and shops.

Looking at the attitudes of the Irish public, a second IHREC-ESRI study (McGinnity et al., 2018) confirms the studies discussed above. The racial stratification is evident in the finding that while 58% of the Irish-born respondents would accept 'many or some' immigration from the same ethnic group as most Irish (i.e. white), the figures dropped to 41% and 25% respectively for Muslim and Roma immigrants (McGinnity et al, 2018: 24). To give some context, the average figures of ten other Western European countries were at 75%, 54% and 44% respectively for white, Muslim and Roma immigrants. The survey further measured beliefs about biological and cultural superiority among Irish respondents. The findings show that 17% of the respondents believe that some races are born

less intelligent, 45% that some races are more hard-working, and 50% that some cultures are better than others. The ten country average for these questions was at 14%, 40% and 45% respectively for the intelligence, hard work and cultural superiority questions. Another relevant finding reported in this survey includes the frequent contact with outgroups, typically rated as neutral, good and extremely good. Finally, drawing on the European Social Survey findings, attitudes to immigration in the years 2000-2014 vary in accordance with the economic performance of the country, so that during times of economic growth attitudes tend to be more positive.

Overall, these studies indicate that Ireland follows a number of other Western countries in creating a racial hierarchy, with white settled natives on top followed by other ethnic groups. While most of these studies focused on reported incidents of hate crime and discrimination, there were few references of hate speech as a specific category of hate crime. Twomey (2017) referred to incidents of hate speech against Roma perpetrated through a Facebook page, and a series of threats against the Pavee Point co-director Martin Collins via Facebook's messenger. None of these resulted in any prosecutions. Carr (2017) referred to hate mail sent to various Muslim associations in 2013 and Michael (2017) found that some of the reports to ENAR iReport system concerned the use of racial slurs. It is, therefore, important to understand the specificity of hate speech as a particular category of hate crime. The next section is concerned with this.

Hate Speech as a Contested Concept

There is no universally accepted definition of hate speech and the notion itself, along with its legal and ethical implications, has been debated at great length by sociologists, political scientists, philosophers, historians, and law experts (Banks, 2011; Citron, 2009; Heinze, 2016; Herz and Molnar, 2012; Matsuda *et al*, 1993). Historically, the genealogy of the term 'hate speech' is quite recent: Walker (1994) notes, in relation to the US, that in the late 1920s and early 1930s racist and offensive expressions were referred to as 'race hate'. Beginning in the 1940s, speech attacking or defaming a particular group on the basis of its race, ethnicity, gender orientation, religion, or other such characteristic, was known as 'group libel'. The term 'hate speech' only became popularised in the 1980s.

There is no universally accepted definition of hate speech and the notion itself, along with its legal and ethical implications, has been debated at great length by sociologists, political scientists, philosophers, historians, and law experts.

In its current usage, hate speech is generally taken to refer to a rather heterogeneous set of manifestations, ranging from unlawful criminal acts to speech which is offensive and disturbing, but not necessarily unlawful (Gagliardone *et al.*, 2014). Some definitions of hate speech include any expression of contempt and animosity towards groups and individuals and utterances that are stereotyping and demeaning (Coliver, 1992; Wentraub-Reiter, 1998). Raphael Cohen Almagor's (2011: 1-2) definition is typical of this 'broad' conceptualisation of hate speech:

> 'Hate speech is defined as bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics. It expresses discriminatory, intimidating, disapproving, antagonistic, and/or prejudicial attitudes toward those characteristics, which include gender, race, religion, ethnicity, colour, national origin, disability, or sexual orientation. Hate speech is intended to injure, dehumanise, harass, intimidate, debase, degrade, and victimise the targeted groups, and to foment insensitivity and brutality against them. A hate site is defined as a site that carries a hateful message in any form of textual, visual, or audio-based rhetoric.'

The Council of Europe's Committee of Ministers' Recommendation 97(20) defines hate speech as encompassing *'all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin'*[10]. While this definition is quite wide-ranging, the Council of Europe distinguishes between expressions which, although offensive, shocking, and insulting, are fully protected by the right to freedom of expression, and expressions that do not enjoy that protection.

Generally, definitions of what constitutes unlawful hate speech emphasise the element of 'intent', or 'incitement', assessing the unlawfulness of speech acts on the basis of their direct and immediate harmful potential to individuals or public order. Individual legislations diverge greatly in terms of the types of speech that are prohibited, with national statutes often mirroring the political and constitutional traditions of that country and local cultures of speech. In the US, the First Amendment means that prohibited hateful speech only applies to 'fighting words', 'those that by their very utterance inflict injury or tend to incite an immediate breach of the peace'[11]; offensive and demeaning remarks or racist slurs do not qualify as "fighting words", unless they are personally abusive. On

10 https://rm.coe.int/168071e53e

11 https://www.law.cornell.edu/wex/fighting_words

the contrary, legislatures in most European countries prohibit certain forms of speech based on the content *itself* even in the absence of direct and explicit threats to violence. However, there remains considerable national variations: in some European countries, the denial of the Holocaust, the apology of fascism, or blasphemy are considered hate speech, whereas in others these conducts are not proscribed (Banks, 2011) notwithstanding the EC Framework Decision of 2008, which is discussed below. The Council of Europe has developed a series of measures in order to lead to some kind of international harmonisation, most notably through the introduction of the Additional Protocol to the Convention on Cybercrime, which criminalises the publication of 'racist and xenophobic material' that promotes or incites hatred, discrimination or violence, against any individual or group of individuals, based on race, colour, descent or national or ethnic origin.[12] The protocol also extends the scope of extradition provisions to include those sought for Internet hate speech crimes.

Social media corporations broadly follow international and EU legal guidelines when it comes to policy rules regulating hate speech on their platforms. Both Facebook and Twitter contain references to the same vulnerable groups identified in the Universal Declaration of Human Rights and enshrined in the International Convention for Civic and Political Rights (ICCPR, see below). Facebook's community standards are influenced by all these and prohibit "*content that directly attacks people based on their: race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender or gender identity, or serious disabilities or diseases*".[13] An analysis of Facebook's leaked training materials[14] shows that this understanding of hate speech is operationalised in terms of removing contents that incite to violence or are dehumanising against the above 'protected categories' but not contents that are, for example, 'degrading generalisations' or that address non-protected categories or 'quasi-protected categories', such as for example 'migrants' (Siapera, Viejo-Otero and Moreo, 2017).

In the Irish context, the principal legal instrument to tackle hate speech is the 1989 Prohibition of Incitement to Hatred Act. The Act does not contain a precise definition of hate speech but makes it an offence *'to publish, display or distribute written or visual materials – as well as saying words or engaging in behaviour – which are threatening, abusive or insulting and are intended or, having regard to all the circumstances, are likely to stir up hatred'*[15]. The Act prohibits incitement to hatred *'against a group of persons in the State or elsewhere on account of their race, colour, nationality, religion, ethnic or national origins, membership of the Travelling community or sexual orientation'*. While the Act is technically broad enough to include online offences, there are difficulties in adapting current legislation to online platforms; to date, public shaming and media outrage, rather than legal sanctions, have been the typical response to racist online utterances in Ireland as elsewhere (O'Mahony, 2011; Twomey, 2017).[16] Anti-discrimination campaigners and international monitoring bodies have also argued that the current legislation is outdated and that much broader hate-crime legislation is required at a time when racist incidents and cyberhate are on the rise (ENAR, 2018).

In 2016, the Law Reform Commission (LRC) published a report on 'harmful communications', focusing in particular on the need to develop a new legal framework for dealing with new forms of harassment, such as victim-blaming and sharing intimate pictures without consent, practices often referred to as 'revenge porn' and 'upskirting' (LRC, 2016). The report further examined the intersections between the online harmful communications and hate speech but considered that online hate speech would be better addressed through reforming hate speech regulation.

At the same time, the LRC report noted that Ireland is expected to ratify the Council of Europe Convention on Cybercrime, which may also include the ratification of the Additional Protocol to the Convention concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems. Ireland is also required to implement the 2008 EU Framework Decision on combating racism and xenophobia, discussed more extensively below. In the most recent report on the implementation of the Framework Decision in 2014, it was noted that Ireland does not have any legislation criminalising acts such as public condoning, denial or gross trivialisation of genocide, crimes against humanity and war crimes, as well as denial or gross trivialisation of the crimes defined in the Charter of the International Military Tribunal (i.e. the crimes of the European Axis countries during WWII). Moreover, Ireland has included an exception to criminalising incitement to hatred, by making dependent on it being threatening, abusive or insulting (EC, 2014).

Recent initiatives taken by the Irish Government to target abuse and harassment online may signal an increase in enforcement as well as a shift towards recognising the specificity of the online domain. As the then Minister for Communications

12 http://conventions.coe.int/Treaty/Commun/ChercheSig.asp?NT=189&CM=8&DF=17/02/2006&CL=ENG

13 https://www.Facebook.com/communitystandards#hate-speech

14 https://www.propublica.org/article/Facebook-hate-speech-censorship-internal-documents-algorithms

15 http://www.irishstatutebook.ie/eli/1989/act/19/enacted/en/html

16 https://www.irishtimes.com/news/crime-and-law/courts-service-reveals-five-convictions-for-hate-crime-since-1989-1.3124352. The act has been remarkably underused: according to documents released by the Courts Service under the Freedom of Information Act, there have been 44 prosecutions under the Act since 2000, of which only five resulted in convictions. Of the 44 cases, 22 were struck out or dismissed by the court and seven were withdrawn by the Director of Public Prosecutions.

Denis Naughten announced in February 2017, the intention is to appoint a statutory Digital Safety Commissioner with the authority to compel social media platforms to remove harmful content promptly from their services. It is proposed that the new office will be tasked with drafting a statutory code of practice on digital safety: the focus seems primarily to be on child protection issues, like online bullying and harassment, so it is unclear whether it will cover broader issues of hate speech (Linehan, 2017)[17] .

To summarise, the main issue here concerns the specificity of online hate speech and online racism. Some of the questions emerging include the extent to which existing measures address online hate speech, and the role of digital platforms and their technical features in enabling such forms of speech. The Law Reform Commission operates under the principle of technology neutrality, according to which regulation should address the behaviours or actions and not the means used. This points to a general understanding that unless the behaviours or actions are themselves different, existing legislation should cover them wherever they take place and via whatever means. The next section considers the legal instruments and the debates they are part of.

Some of the questions emerging include the extent to which existing measures address online hate speech, and the role of digital platforms and their technical features in enabling such forms of speech.

Hate Speech or Freedom of Speech?

Public debates around hate speech and freedom of speech tend to be highly polarised in the Irish context, as elsewhere[18]. When Nicholas Pell penned a controversial article in the *Irish Times* about the alt-right movement – which included a glossary of sexist, racist and hateful terms used – many journalists and readers protested against the newspaper's decision to publish the article, claiming that it normalised racism and hate rhetoric (Mullally, 2017).[19] The *Irish Times* stood by its decision, claiming that the publication of the article did not amount to condoning the views of the Alt-Right but rather sought to inform and challenge readers to form their opinions on the matter. In a poll on *Claire Byrne Live show*, conducted soon after the article controversy, 65% of respondents expressed the view that no limits should be placed on freedom of speech to protect people from being offended (Leonard, 2017)[20]. However, when Katie Hopkins was invited to appear on *The Late Late Show*, RTÉ received over 1600 complaints from people arguing that Hopkins' hateful rhetoric should not be given exposure on TV or through other media (Griffin and McMahon, 2016)[21]. When the Irish online news outlet Journal.ie decided to introduce a series of changes to their comments section, with the aim of improving readers' experiences and standards of 'decency and civility', many readers protested what they considered was an attempt to stifle debate and their freedom of speech.[22]

While such debates take place in print and broadcast media, they are echoed in the social media sphere, where increasingly, as noted by Titley (2017a), freedom of speech is appropriated and mobilised by the far right. Titley discusses the post-Charlie Hebdo context, where ostensibly in the name of 'freedom of speech' far right groups and political parties staged events such as 'Everybody Draw Mohammed Day', showing that the notion of freedom of speech is in danger of changing from a political right to a racialised strategy. The tensions in how hate speech and freedom of speech are conceived and mobilised become more apparent in the digital domain, firstly because of the '***informational libertarianism***' [see glossary and definitions] that is an integral part of cyberculture (Jordan, 2001; Barbrook and Cameron, 1996); and secondly, because of the '***spreadability***' [see glossary and definitions] of ideas and information in the digital sphere (Jenkins, Ford and Green, 2013). Both of these add to the already existing complexity of these questions. In order to understand the current regulatory and legal framework, the sections below offer a discussion of the legal instruments that pertain to questions of racial discrimination and freedom of expression, and their conceptual antecedents and tensions.

17 https://www.irishtimes.com/news/social-affairs/digital-safety-watchdog-could-prove-a-milestone-online-1.2967204, accessed 11/10/2017

18 A similar debate erupted during the recent controversy surrounding broadcaster George Hook's 'rape comments', as evidenced for example, in Fintan O'Toole and Kitty Holland's articles in the Irish Times, available at: https://www.irishtimes.com/opinion/george-hook-s-right-to-free-speech-ends-where-it-does-gratuitous-harm-1.3222409?mode=sample&auth-failed=1&pw-origin=https%3A%2F%2Fwww.irishtimes.com%2Fopinion%2Fgeorge-hook-s-right-to-free-speech-ends-where-it-does-gratuitous-harm-1.3222409 and https://www.irishtimes.com/opinion/george-hook-should-be-challenged-not-silenced-1.3219952 [Accessed June 15, 2018]

19 https://www.irishtimes.com/opinion/una-mullally-why-the-irish-times-should-not-have-published-nicholas-pell-1.2926726 [Accessed March 12, 2018]

20 http://www.sin.ie/2017/01/25/free-speech-vs-hate-speech-wheres-the-line/ [Accessed March 12, 2018]

21 https://www.irishtimes.com/news/ireland/irish-news/rt%C3%A9-receives-1-300-complaints-over-katie-hopkins-interview-1.2864436 [Accessed March 12, 2018]

22 http://www.thejournal.ie/journal-comments-section-changes-3286318-Mar2017/ [Accessed March 12, 2018].

Legal Instruments: A Brief Exposition

Historically, the current understandings of hate speech and its relationship to freedom of speech/expression can be traced to the 1948 Universal Declaration of Human Rights. Three articles are relevant here: Articles 2, 7 and 19. Article 2 is concerned with establishing the right of non-discrimination: "Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status."

Article 7, which refers to a universal right to equal protection: "All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination." This is almost identical to Article 2 but with the crucial addition of the notion of incitement to discrimination.

Thirdly, Article 19 refers to the issue of freedom of expression: "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."

These articles are contradictory to the extent that protections offered under Article 19 can clearly compromise the right to non-discrimination of Articles 2 and 7. The main way in which the tension between the right to be protected from discrimination and the right to freedom of expression was resolved in most legal systems is through making punishable incitement to hatred and violence against groups of people.

The Universal Declaration of Human Rights constitutes part of the tri-partite International Bill of Human Rights; the other two parts are the International Covenant of Civil and Political Rights (ICCPR) and the International Covenant of Economic, Social and Cultural Rights (ICESCR). The ICCPR constitutes a revision and reiteration of the main articles of the Declaration. While the UDHR takes the form of recommendations, the ICCPR is legally binding. In terms of racism and hate speech, the relevant Articles do not depart substantially from those in the UDHR . Article 19 on freedom of expression remains substantially the same, complemented by Article 20[23] on prohibiting incitement to hatred, while Article 26 provides the necessary protection against discrimination. It should be noted here that the well-known Article 19 on freedom of expression is not absolute but comes with 'special duties and responsibilities'.

While these instruments are addressing human rights in general, the United Nations felt that it was necessary to address the specific issue of racism and discrimination. The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) which was adopted in 1965, defines racial discrimination and develops a set of articles seeking to eradicate this across the signatory states, currently 179 out of the 193 state-members of the UN. Crucially, Article 4 obliges states to prohibit incitement to racial hatred and violence.[24]

23 *Article 19.* 1. Everyone shall have the right to hold opinions without interference. 2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:(a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order, or of public health or morals.
Article 20. 1. Any propaganda for war shall be prohibited by law. 2. Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.
Article 26. All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.
" [ICCPR, 1976, pages 178-179, full text available at: https://treaties.un.org/doc/publication/unts/volume%20999/volume-999-i-14668-english.pdf, accessed June 14, 2018]

24 *Article 4*: States parties condemn all propaganda and all organisations which are based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination and, to this end, with due regard to the principles embodied in the Universal Declaration of Human Rights and the rights expressly set forth in Article 5 of this Convention, inter alia:
(a) Shall declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin, and also the provision of any assistance to racist activities, including the financing thereof;
(b) Shall declare illegal and prohibit organisations, and also organised and all other propaganda activities, which promote and incite racial discrimination, and shall recognise participation in such organisations or activities as an offence punishable by law;
(c) Shall not permit public authorities or public institutions, national or local, to promote or incite racial discrimination. [full text available here: https://www.ohchr.org/EN/ProfessionalInterest/Pages/CERD.aspx, accessed June 18, 2018]

Regionally, in the Council of Europe system, Article 10 of the European Convention on Human Rights[25] recognises the right to freedom of expression, subject to 'formalities, conditions, restrictions or penalties as are prescribed by law and necessary in a democratic society'.[26] The case law[27] of the European Court of Human Rights (ECtHR) has demonstrated that any such restrictions to freedom of expression must be 'proportionate to the legitimate aim pursued',[28] and has found against governments which have failed to meet this standard. The ECtHR has also found certain forms of speech, including negationism and Holocaust denial, to be excluded from the protections of the Convention where they negate its fundamental values.[29]

The EC Framework Decision 2008/913/JHA of 28 November 2008[30] requires member states to take measures to tackle xenophobia and hate speech that includes incitement "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"; the "public dissemination or distribution of tracts, pictures or other material" that incites to violence or hatred; the trivialisation of genocide, crimes against humanity, war crimes, and crimes as defined by the Tribunal of Nuremberg (EC, 2008: 328/56). In the European Union, national governments are expected to make the necessary amendments to their national legislation to bring it in line with this decision. According to the 2014 report on the implementation of this framework decision, Ireland has no criminal law provision for public condoning, denial or gross trivialisation of genocide, crimes against humanity and war crimes or for the crimes crimes defined in the Charter of the International Military Tribunal, while it has also added an exception to the inclusion of racist motivation as an aggravating factor in crime, arguing that it can always be considered by the courts (EC, 2014)[31].

Conceptual Tensions

In the legal instruments, therefore, the tension is resolved through limiting the category of hate speech in the ways seen above, and in offering explicit protection to freedom of expression. However, as the digital domain is expanding, opportunities for expression, tensions re-emerge.

> *As the digital domain is expanding, opportunities for expression, tensions re-emerge*

Similar tensions are encountered in the academic and scholarly debates on hate speech which are equally shaped by the dilemma of freedom of speech/expression versus controlling hate speech (Garton Ash, 2016; Heinze,2015; Matsuda et al., 1993; Butler, 1997; Waldron, 2012). Typically, scholars and legislators focus on the role of freedom of speech in a democracy, on the one hand, and the harm and injury caused by hate speech on the other. While it is beyond the scope of this report to offer more than an overview of these debates, or to make a case for or against hate speech bans, it is useful to summarise the different and often contradictory social imaginaries which animate them. The absolute freedom of speech/expression position is most clearly articulated in Eric Heinze's work. Heinze (2015) argues that freedom of speech must be seen not merely as an individual right but as a fundamental attribute of democratic citizenship itself: the constitutions upon which democracies rely, and their amendments, and by extension all laws and procedures that make up democracies, are based on public discourse. If this is in any way limited, curbed or compromised, all these processes are themselves compromised. Hence, there can be no limits based on the contents of public discourse, and this is what hate speech regulations are seeking to do. Heinze is not oblivious to the potentially pernicious effects of hate speech; he argues that such effects are largely dependent on specific socio-historical contexts and that, in liberal and well-established Western democracies, the harms of hate speech can be dealt with in non-legal ways without placing limits on

25 The European Convention of Human Rights entered into force in 1953 of as part of a regional system put in place as a response to the Second World War on the one hand and on the rise of the power of the USSR and what was perceived at the time as a potential threat to liberal democratic values (Rainey, Wicks and Ovey, 2014). The Strasbourg-based European Court of Human Rights (ECtHR) hears cases alleging breaches of the Convention.

26 ECHR Article 10(2). The Article continues 'in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.'

27 For an overview of ECtHR Case Law in the area of hate speech and freedom of expression, see the Court's regularly updated fact sheet on Hate Speech. The most recent version, updated March 2018, is available at https://www.echr.coe.int/Documents/FS_Hate_speech_ENG.pdf [last accessed 26 June 2018].

28 *Handyside v. the United Kingdom*, judgment of 7 December 1976, application no. 5493/72, § 49.

29 For example, the Court's negative admissibility decision in *M'Bala M'Bala v. France*, application no. 25239/13, concerning a public comedy performance which included demonstrations of hatred, anti-Semitism, negationism and Holocaust denial. Article 17 of the ECHR states that 'Nothing in this Convention may be interpreted as implying for any State, group or person any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms set forth herein.'

30 Full text available here: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:328:0055:0058:EN:PDF [accessed March, 12, 2018]

31 Full text on the report of the implementation is found here: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52014DC0027 [accessed on June 19, 2018].

public discourse. The social imaginary that underpins Heinze's account is one that prioritises liberal principles and pluralism and underplays the role of hate speech in legitimising systemic forms of inequality or oppression. For Heinze, the existence of stable and mature democratic institutions are in themselves a sufficient guarantee that hateful speech does not translate into actual violence or discrimination. The question for Heinze's view that emerges here concerns the extent to which the State and its institutions are neutral (Lentin and Lentin, 2006).

In contrast, Matsuda, Lawrence, Delgado and Crenshaw (1993) begin from a different standpoint, that of critical race theory. Critical race theory focuses and values the historical experience of oppressed communities and prioritises analysis and tools that aim to eradicate race-related inequalities and racial injustices. It, therefore, focuses on the harm caused by hate speech and its role in reinforcing and perpetuating the social structures that enact discrimination. In Matsuda *et al.'s* (1993) view, what is at stake in hate speech regulation is a vision of society where ideals of freedom and equality are redefined. Hate speech regulation is, therefore, necessary to protect the rights of those who have been and continue to be denied access to the full benefits of citizenship and to repair the effects of historical wrongs.

Heinze critiques Matsuda *et al*'s position based on the argument that in liberal democratic contexts, anti-hate speech measures are unnecessary, ineffective, and counterproductive. A different critique is put forward by philosopher Judith Butler (1991) who argues that hate speech bans can be used to take the politics out of the struggle for racial justice, limiting political engagement to calls for regulating or suppressing harmful language, with the attendant risk of reducing issues of racism to individual grievances rather than viewing them as manifestations of systemic oppression.

Like Matsuda *et al*'s, Jeremy Waldron's analysis focuses on the harms of hate speech but seen through the lens of classical republicanism. Hate speech is harmful because it damages societal inclusiveness, which Waldron conceives as a 'public good', something that societies should sponsor and be committed to. Waldron (2012: 4) explains that people and groups should be accepting of the fact that society is "not just for them; but it is for them too, along with all of others. And each person, each member of each group, should be able to go about his or her business, with the assurance that there will be no need to face hostility, violence, discrimination, or exclusion by others". For Waldron, hate speech functions as a slow-acting toxic agent, incrementally and effectively poisoning the fabric of public life and the very civic spaces we inhabit. From this point of view, the 'health' of Western democracies that Heinze seems to take for granted, may be allowed to erode because of this toxicity.

It is evident that societies are called to develop a balanced set of policies that protects the rights of minorities while safeguarding the fundamental democratic right of freedom of expression. The civil rights of minority members, especially the right to enjoy all aspects of social life, including participation in the digital domain, need to be protected. At the same time, democratic societies cannot exist without the freedom to express diverse opinions.

The civil rights of minority members, especially the right to enjoy all aspects of social life, including participation in the digital domain, need to be protected. At the same time, democratic societies cannot exist without the freedom to express diverse opinions.

In May 2016, the European Commission signed a voluntary Code of Conduct with the main four social media corporations (Facebook, YouTube (Google), Microsoft, and Twitter), concerning 'illegal hate speech'[32]. In January 2018, two more platforms signed up (Instagram and Google+) followed by Snapchat in May. The main commitments of the social media companies include the following:

1. to develop and implement processes by which to review notifications of hate speech and to provide clear guidelines to their users, prohibiting incitement to violence and hateful conduct;
2. to quickly and efficiently review notifications based on their own rules and the EC Framework Decision 2008/913/JHA or national legislation that supplements it;
3. to undertake the majority of reviews and removals of contents as appropriate in less than 24 hours;
4. to work with civil society actors in order to identify hateful conduct more readily and secondly in order to enable them to develop effective counter-narratives.

In the latest review in January 2018, the Commission reported significant progress, with 70% of illegal hate speech reported by NGOs and other organisations participating in the evaluation and in more than 81% of the cases this happens within 24 hours[33].

32 European Commission – Press Release, European Commission and IT Companies announce Code of Conduct on illegal online hate speech, Brussels, 31 May, 2016, full text available at: http://europa.eu/rapid/press-release_IP-16-1937_en.htm. Accessed March 12, 2018.

33 EC, Countering illegal hate speech online – Commission initiative shows continued improvement, further platforms join, Press release, January 19, 2018, full text found here: http://europa.eu/rapid/press-release_IP-18-261_en.htm - accessed June 20, 2018.

Adding to the Code of Conduct, in March 2018, the Commission released a set of recommendations regarding illegal content, which are primarily operational rules for the effective removal of such contents. These recommendations are:

> Clearer 'notice and action' procedures: Companies should set out easy and transparent rules for notifying illegal content, including fast-track procedures for 'trusted flaggers'. To avoid the unintended removal of content which is not illegal, content providers should be informed about such decisions and have the opportunity to contest them.
>
> More efficient tools and proactive technologies: Companies should set out clear notification systems for users. They should have proactive tools to detect and remove illegal content, in particular for terrorism content and for content which does not need contextualisation to be deemed illegal, such as child sexual abuse material or counterfeited goods.
>
> Stronger safeguards to ensure fundamental rights: To ensure that decisions to remove content are accurate and well-founded, especially when automated tools are used, companies should put in place effective and appropriate safeguards, including human oversight and verification, in full respect of fundamental rights, freedom of expression and data protection rules.
>
> Special attention to small companies: The industry should, through voluntary arrangements, cooperate and share experiences, best practices and technological solutions, including tools allowing for automatic detection. This shared responsibility should particularly benefit smaller platforms with more limited resources and expertise.
>
> Closer cooperation with authorities: If there is evidence of a serious criminal offence or a suspicion that illegal content is posing a threat to life or safety, companies should promptly inform law enforcement authorities. Member States are encouraged to establish the appropriate legal obligations.
>
> (EC, 2018, A Europe that protects: Commission reinforces EU response to illegal content online)[34]

While the EC commitment to the implementation of the EC Framework Decision of 2008 and its corresponding national legislations is important, societies need to remain vigilant in terms of new developments in the sphere of social media. Further, in order to be able to counter more effectively hate speech and online hateful conduct, an in-depth knowledge of this conduct is necessary. The first step for this requires the development of tools to monitor the digital domain. The current project represents one such monitoring effort in the context of Ireland. The section below offers more details of previous relevant research.

Monitoring Hate in Social Media: Previous Research

Monitoring hate speech on social media can take a variety of forms depending on the purpose and the methodology used. Some projects, like the Umati in Kenya, focus on real time monitoring, often operating as 'an early warning system' during times of political volatility or tension (e.g. political elections, referendum campaigns). Other studies undertake after-the-fact analysis, looking at archives of messages and posts which are analysed through automated techniques or by researchers. Monitoring projects may prioritise content and discourse analysis techniques with the aim of understanding the features, ideologies, and effects of such messages within specific social and political contexts. While they may track trends in frequency or location, their main goal is to understand how hate messages are constructed; how they influence recipients; and to identify recurrent themes and patterns of speech as well as networks of hate sites (MRAP 2009; British Institute of Human Rights 2012). Online users can also play an important role in monitoring cyberhate by using hate speech hotlines or apps to alert relevant authorities to incidents or sites that warrant law intervention or notifying ISPs of material breaching their code of conduct (Lentin and Humphry, 2017). For instance, iStreetWatch[35] is an app for reporting and tracking racist and xenophobic harassment in UK public spaces. Similarly, Islamophobia Watch[36] is used for reporting incidents of Islamophobic abuse. Kick It Out[37] has a number of reporting methods available to anybody who has seen, heard or been on the receiving end of discriminatory abuse in a football environment. However, reporting tools are designed for reporting only and no automatic data analysis is performed.

34 Full text of the press release is found here: http://europa.eu/rapid/press-release_IP-18-1169_en.htm - accessed June 20, 2018.

35 https://www.istreetwatch.co.uk/

36 http://islamophobiawatch.com.au/

37 http://www.kickitout.org/get-involved/report-it/

Human monitoring enables researchers to assess the subtleties of content and context but is also labour and time intensive, and thus usually applied on relatively small sets of data. Approaches to monitoring hate speech increasingly use or experiment with automated techniques and machine learning tools capable of generating and analysing large datasets, often accessed in real time (Gagliardone et al. 2014; Prentice et al. 2012; Warner and Hirshberg, 2012). Large-scale monitoring projects often combine corpus linguistic techniques for the automated processing of messages with a qualitative analysis of smaller datasets to highlight the nuances of context and the specific features of online discursive interactions (see Brindle, 2009)[38].

The Centre for the Analysis of Social Media based at the British think tank Demos has been at the forefront of this type of research using data mining through complex algorithms and qualitative content analysis of selected content[39]. Demos has published studies on the prevalence and patterns of use of racial and ethnic slurs on Twitter (Bartlett et al., 2014); on Islamophobia spikes on Twitter in the immediate aftermath of news events (Miller et al., 2016); on the volume of derogatory and/or hateful anti-Muslim tweets; as well as the impact of the Brexit Referendum on xenophobia and racism.

Another large monitoring project that uses IT and big data is the Mandola project co-funded by the Rights, Equality and Citizenship (REC) Programme of the European Commission. The Mandola project aim is 'to monitor the spread and penetration of online hate-related speech in Europe and in Member States using big-data approaches, while investigating the possibility to distinguish, amongst monitored contents, between potentially illegal hate-related speeches and potentially non illegal hate-related speeches'. Another aim of the project is to provide actionable information to inform policy, to identify and share best practices across Europe, and set a reporting mechanism through which ordinary citizens can report illegal hate speech.

It is in this context, and recognising the nuances and difficulties involved in dealing with hateful speech, we designed this study as a way into empirically apprehending online racist hate speech and its range, as well as the ways in which people deal with this kind of hate when they encounter it online. As mentioned in the introduction, HateTrack differs from previous projects in two main ways: (i) it moves beyond the notion of illegal hate speech, deriving a definition of racism and racially-loaded contents from civil society actors; (ii) it focuses on the specific national context of Ireland. We explain below the main research questions and research design of the project.

38 Andrew Brindle (2009) used the computer programme WordSmith5 to analyse messages posted on the white supremacist web-forum Stormfront and identify words and phrases that appeared unusually often and/or together. He also carried out a critical discourse analysis of a small sample of the messages to understand supremacists' ideologies. Prentice et al. (2012) combined content analysis and semantic analysis in a study on Islamic extremists. The content analysis component involved researchers reading texts to identify occurrences of "persuasive devices" - such as persuasion, direct pressure, inspirational appeals, etc. - used by extremists to influence audiences. The semantic analysis relied on the computer programme WMatrix to identify concepts that appeared in the studied texts significantly more often than in 'normal' texts. It also identified how concepts occurred together in texts and revealed trends in the appearance of these concepts over time.

39 https://www.demos.co.uk/project/counter-speech-on-Facebook-phase-2/

Research Design and Methodology

Research Design and Methodology

The project has posed the following three research questions:

RQ1: What are the defining characteristics, the range and severity of online racist hate speech?

RQ2: How can these materials be tracked on public pages on Facebook and on Twitter?

RQ3: What kinds of online racist incidents tend to get reported in Ireland and how do they compare to the broader racist hate materials circulating? What are the perceived barriers to reporting and what kinds of experiences do victims and bystanders of online racist hate have to report?

To address these questions, the project is divided into four stages.

In **Stage I**, we sought to obtain a more nuanced understanding of what constitutes (online) racist hate speech by looking at how civil society actors define and experience it.

In **Stage II**, we used these definitions to develop the HateTrack tool.

In **Stage III**, we harvested materials to generate a dataset comprised of about 6,000 online posts classified as containing racially-loaded toxic speech. The posts were subsequently analysed in terms of the repertoires they contained, the groups they targeted, the events that seemed to 'trigger' them, and the kinds of people writing and disseminating them.

Finally, in **Stage IV**, we held semi-public discussions with groups of students, discussing what gets reported and why or why not. The next section offers more details in terms of the methodology used for each stage.

The project adopted a mixed methods approach, using different methodologies for the different parts of the research. This section offers specific details on these methods and sampling decisions as these were applied to each of the different stages. Overall, in developing the conceptual framework for the analysis, we relied on two distinct scientific fields: in researching hate speech, we made use of legal and political theory; and in researching race and racism, we relied on the sociology of race, and especially critical race theory, which underpins current approaches to legal theory (especially Matsuda et al., 1993). Additionally, we reviewed several recent publications concerning online hate speech, including Hate Spin (George, 2016), Hate Crimes in Cyberspace (Citron, 2014), Countering Online Hate Speech (Gagliardone et al., 2015) and After Charlie Hebdo (Titley et al., 2017). In terms of the actual analysis, given the exploratory nature of this study, we employed a combination of grounded theory and discourse analysis (Glaser and Strauss, 2017 [1999]; Fairclough, 2013) , that allows insights to emerge from the bottom up, looking for patterns and regularities as they occur in the text of interviews or the dataset through the use repetition of certain phrases, expressions, figures of speech and so on.

Stage I: Focus Groups and In-Depth Interviews

We conducted five focus group discussions with anti-racism and migrant support NGOs and voluntary organisations, and Roma- and Traveller-led community-based organisations; as well as twelve one-to-one, in-depth interviews with: journalists and media professionals; activists; the communication officers of migrant support NGOs (who typically manage online platforms); ethnic minority broadcasters; and academics with expertise in the area of race, ethnicity, refugee and asylum, anti-Semitism, and Islamophobia. The sample is not comprehensive, nor do we consider it representative of the range of civil society groups active in this area. Rather, we consider it as an entry point into the theme of racist hate speech. Indeed, as discussed below, there was a high degree of consistency in what our informants imparted and what they viewed as important.

Before holding the focus groups, the research team organised an informal pilot focus group with PhD students and post-doc researchers based in DCU's School of Communications to get some input on methodology, facilitation skills, and tease out some ethical issues around discussing sensitive topics. The pilot focus group also provided an opportunity to discuss the complexities inherent to defining what constitutes racist hate speech and how best harvest examples to inform the design of the computation tool. The notions of 'racially-loaded content' and 'racist discourses' were suggested as a more nuanced and sensitive lens through which to analyse online content and the dynamics of cyberhate.

Focus group participants and interviewees were selected on the basis that they are trained on issues of racial discrimination, work with clients who have directly experienced such problems, and have researched and written about such issues in relation to the Irish context. We treated participants' accounts as 'systems of knowledge' in their own right (Essed, 1991: 109). Consistency between accounts given by research participants, independent of each other, and between accounts and scholarly sources on online racism and hate speech is evidence of the reliability and validity of our approach.

Overall, participants' inputs provided: 1) the evaluative framework upon which specific online content was classified to inform the design of the algorithm; 2) insider knowledge of the specificities of the Irish socio-historical context and local cultures of speech; 3) examples of hateful, racially-loaded content as well as names of public organised groups and accounts operating from within Ireland.

While the topics discussed during the focus groups and interviews were raised with the requirement of training the algorithm, they were not limited to it. We asked participants to describe what type of racially-loaded concepts/ideas/images they had encountered online and to provide examples, if they felt comfortable; which social media platforms are more conducive to the circulation of hateful racist material/messages; which are the groups/communities most targeted; what topics/issues tend to trigger racist or racially-loaded language; what was their experiences when reporting racist hate speech to Facebook and Twitter. Participants also elaborated upon their own personal experience of online harassment and shared the names of organised groups on Twitter and Facebook they had come across in their work or had blocked/muted because of abusive and racist posts.

Stage II: Building the Tool

Research participants' definitions and understandings of racist hate speech reflect their concern with the impact and the effects it has on those attacked. Our informants were firmly oriented towards understanding how racism operates and becomes diffused in society. A similar commitment underpins the design and aims of the HateTrack. The main idea behind the HateTrack tool was to test the methodological viability of using definitions derived from civic society actors to train an algorithm to identify and classify instances of racist hate speech across social media platforms (specifically Twitter and Facebook). We used the definitions and problematic social media accounts that our informants provided, alongside with others linked to them, to gather examples of racist discourses and to help identify particular tropes/narratives (in total 113 Twitter public accounts were examined). More examples were collected through looking at comments posted under news articles published by the Irish Times, Irish Independent and Journal.ie on their eFACEBOOK pages. The examples collected were used to compile two corpora, each comprising nearly eight hundred instances of 'racially-loaded toxic content' and neutral content. The classification followed along the lines suggested by our informants, and the corpora were used for the initial training of the algorithm.

The section below outlines the more technical methodology and procedure by which HateTrack was built.

HateTrack: Technical Information

The HateTrack system is concerned with identifying various forms of online racist hate speech ranging from racist slurs to more banal or 'everyday' racist discourses. It relies on a computational tool that can track hateful contents on public Facebook pages and Twitter accounts.

Front-end

The front-end of the HateTrack system consists of two modules. The web-scraping module and the racially-loaded toxic content ranking module. The page for web scraping is shown as figure 1. The "Operation" drop-down list allows user to choose scraping posts or comments from Facebook and tweets from Twitter. Selecting an operation, more criteria such as the keywords, user name or post ID can be input to narrow the scraping. A jobs table records the user scraping history and displays the records chronologically. By ticking one or more records and clicking the view button, the system will navigate to the toxic content ranking page.

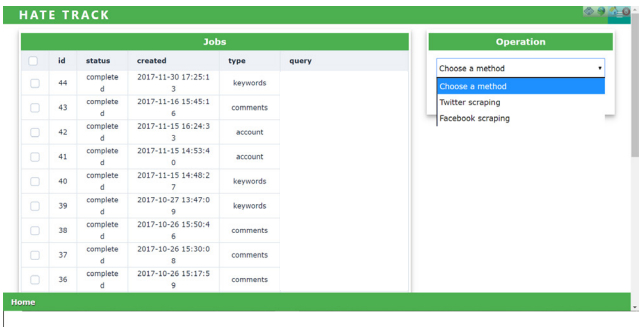

Figure 1: Web-scraping page

As shown in figure 2, the racially-loaded toxic contents are categorised into two tabs, "high probability" and "low probability", depending on their toxicity.

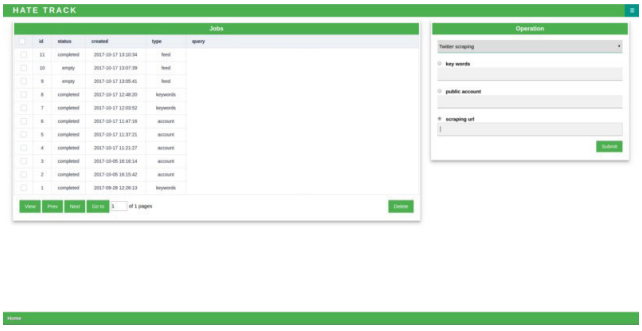

Figure 2: Toxic content ranking

In each tab, contents are ordered according to their scores by default, which represent how likely those contents are to be in that category. This score with consecutive values is a key attribute of the HateTrack system as it allows users to explore the scraped content much quicker than inspecting unordered text line by line. For example, users may only focus on the extremely toxic content and the most ambiguous text when they analyse thousands of scraped tweets. Moreover, users can also order contents according to the number of 'likes/ favourites' or 'shares' they received, which may indicate their popularity and influence. An operation panel is also provided to refine the tab content by setting selected text as "high probability" or "low probability" manually. The score of hand-labelled text will be 1.0 marked with green background. This gives a quick and easy way for researchers to feedback and train the algorithm further.

Back-end

The data collected by HateTrack are organised in a database as shown in figure 3. Users' scraping actions in table "ht_scraping" can be detailed in table "ht_Twitter" and "ht_Facebook" while table "ht_method", "ht_category" and "ht_result" are used for recording the classification algorithm results of scraped text.

Method52 and its precursor Method51 are social media analysis platforms coupled with Naive Bayes models that help social scientists harness information from large amounts of unlabelled data. The platforms provide user interfaces to enable researchers to customise their data processing pipeline and employ supervised machine learning approaches for tailored automatic data analysis. Using Method52, a software that allows a collection of Tweets that contain specific keywords from Twitter's Stream and Search 'Application Programming Interface' (or APIs), Demos[40] conducted a study on tweets considered to be derogatory anti-Islamic. The classifiers detected 143 hatred tweets out of 200 with hand-crafted features such as words "Paki" and "terrorist" etc. The Mandola project (Dikaiakos et al., 2016) also deployed the Naive Bayes model for its hate speech data analysis with manual feature engineering. However, hand-crafting features are generally hard and expensive. Compared to Method52, HateTrack is a lightweight but more precise system designed specifically for online hate speech screening. HateTrack is flexible and extensible. Plugging in other classification models in the back-end, HateTrack can be easily deployed as a versatile platform. In other words, in the future it can be extended to cover, for example, misogynistic hate speech.

Deep learning has brought a new era in machine learning. Rather than following the old school of machine learning techniques such as Naive Bayes and SVM etc., which have reached a plateau in performance, HateTrack uses deep ***neural network*** [see glossary and definitions] techniques that have shown significantly better results in many applications (Krizhevsky et al., 2012; Mikolov et al., 2013). More specifically, the HateTrack system developed a method based on the Long Short-Term Memory (LSTM) network that does not require expensive hand-crafted features. A dataset containing 290 high probability racially-loaded toxic text and 239 low probability racially-loaded toxic text was prepared using human labelling by our social scientists. Of these, we used 353 instances as balanced training data from this dataset and applied the remaining 176 instances for testing. Our method proved to have a 75.9% classification accuracy which is very promising.

40 https://www.demos.co.uk/wp-content/uploads/2017/04/Results-Methods-Paper-MOPAC-SUMMITDemos.pdf

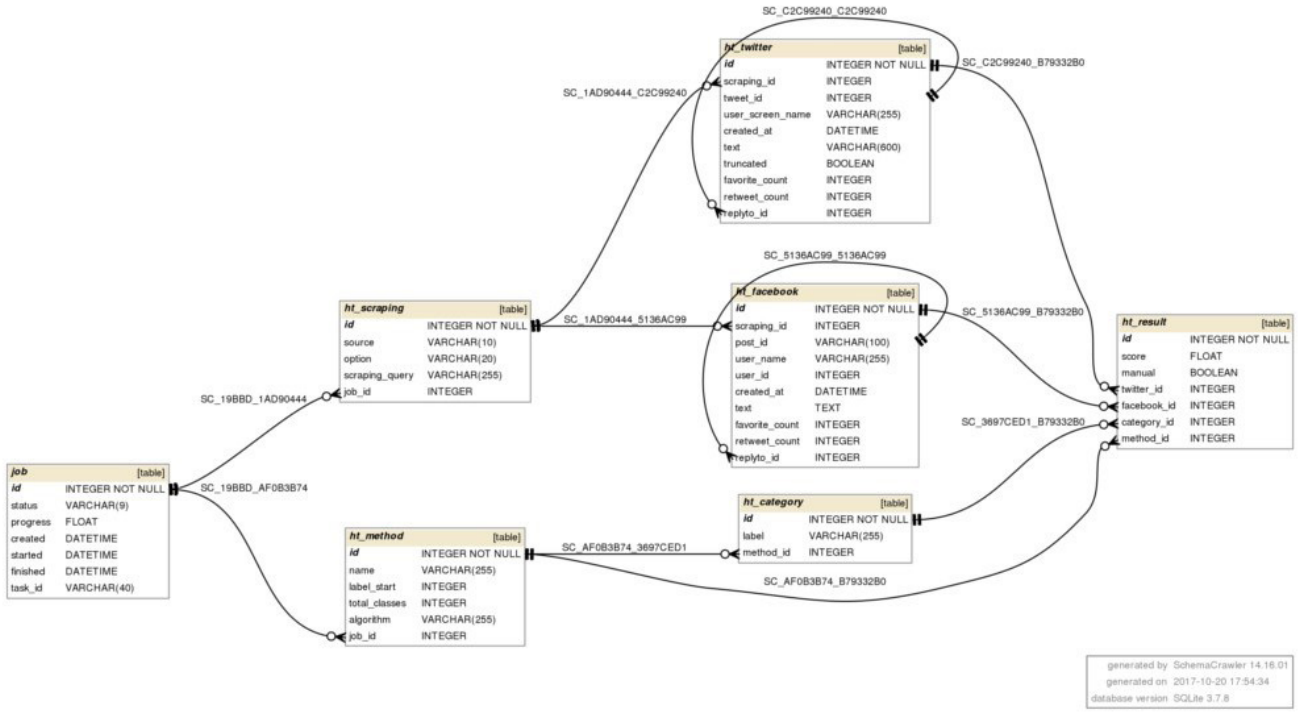

Figure 3: ER diagram of HateTrack database

Stage III: Discourse Analysis of the Dataset

The dataset was generated in two parts: firstly, we manually collected and saved contents, using the account names and other information provided by our informants. This resulted in a sample of 5,725 comments and 113 Twitter accounts. Secondly, we conducted 92 searches (46 valid, 46 void) through HateTrack, using specific account handles, keywords, hashtags or news articles posted on Facebook. Tables 1 and 2 offer more detail. Valid searches were considered; those that returned results while void searches did not return any contents.

News Outlet	Number of Articles	Number of Comments
Journal.ie	23	2,146
Independent	12	1,065
Irish Times	20	2,514
Total	55	5,725

Table 1. Facebook comment threads analysed

Number of searches	Number of Comments/Tweets
92 (46 valid)	10,728

Table 2. HateTrack searches

It should be noted that not all of these contents were problematic. While most of the searches were specific to the Irish context, focusing on accounts known to be Irish or Irish keywords, for example Lisdoonvarna or Direct Provision, some were much broader. The broader keyword searches generated the most contents; for example, the keyword that generated the highest number of results was 'white genocide' (1180[41]).

To analyse these contents, we employed a combination of grounded theory (Glaser and Strauss, 2017 [1999]) and discourse analysis (Fairclough, 2013 [1995]). Discourse analysis looks for patterns and regularities in texts, and then interprets these in terms of the relationships they engender and the socio-cultural practices they are associated with (Fairclough, 2103: 132). Rather than formulating hypotheses and then examining the data in order to see whether they support the hypotheses or not, in grounded theory researchers generate insights from the data in a bottom-up manner. Researchers read and re-read the contents in order to understand the patterns that emerge and then sought to connect these to

41 It should be noted here that searches on Twitter are limited to the past seven days from the day the search takes place. This is due to the platform's data policy.

socio-cultural practices of exclusion and racialisation. We developed several initial taxonomies which we then refined through examining the contents iteratively. In this, we were guided by two main questions: what kinds of discursive repertoires emerge around racialised others; and who are those racialised by these repertoires? The findings of this stage of the research are presented in terms of these two questions.

In parallel, we conducted research on a case study revolving around a fatal stabbing incident in Dundalk, in January 2018. The suspect was from Egypt, but before the facts of the case were made clear, there was considerable speculation regarding a possible terrorist incident. This incident made apparent the synergistic ways in which the mainstream media operate in tandem with some anti-immigration right wing accounts. Although the media may not be doing this on purpose, their speculative and sensationalistic headlines and tweets lend themselves to further exploitation by right wing accounts for their own purposes. The case study relied on an analysis of the first 1000 (out of 6302) tweets harvested through HateTrack using the hashtag #Dundalk and #Dundalkattacks.

Stage IV: Reporting Cultures

In the last part of the research, we were concerned with identifying the circumstances and the kinds of contents that people tend to report. As with the other parts of the project, this research was conducted as an exploration of the 'reporting cultures' among young people rather than as a definitive study of reporting. The main aim here was to gain a sufficient first insight to allow for the development of this part in future research.

We ran three workshops with young people in higher education. These took place at DCU and Colaiste Dhulaigh (Coolock), attended by 3rd year journalism students and at Trinity College attended by MPhil Race and Ethnicity students. These seminars provided an opportunity to present the findings of the research; discuss different definitions of hate speech; examine the moderation polices of Facebook, Twitter, and newspapers like the Guardian and New York Times; and collect data on attitudes towards reporting. Students were asked to discuss real Facebook posts and tweets collected through HateTrack; whether they considered them problematic or racially-loaded; and whether they would have reported them or not and why. Students were divided into groups and were asked to consider the materials for about ten minutes. We then reconvened and discussed the examples together. The sessions were recorded, and the recordings are securely held. None of the students was identified by name.

Research Ethics

The project followed the standard DCU procedure for a 'low risk' project, that is a project that does not involve vulnerable people or children. We sought and obtained informed consent from all participants, using plain language to describe the study and its objectives. We sought, and achieved gender balance in our research participants, with 7 out of 10 individual interviewees and 15 out of the 29 focus group participants being women. All recorded interviews, transcripts and data are stored in a password-protected digital folder. Full anonymity was promised to individual participants, although we noted that we may refer to the organisations they work for.

At the same time, the project required careful consideration of some ethical implications of this research. In Stage I, the main concern was the extent to which asking people who were themselves targeted was risking re-traumatising them. In asking participants to repeat racist or hateful materials, we may unintentionally cause them harm. This emerged out of our pilot focus group as well as in the first focus group we conducted. We remedied this by prefacing the request with 'we are asking you to share only things you are comfortable with'. Further, we asked participants to send us materials they come across with email.

In Stage II, in designing the algorithm, the main ethical issue concerned the names of social media users generating the problematic contents. This was resolved by designing the tool in ways that anonymise user accounts and return only the text. On the other hand, the issue of the names of those targeted still remained. How ethical is it to remove the names of those generating or sharing problematic contents and then keep the names of those targeted? This led to the decision in Stage III not to share openly the dataset created[42] and to remove the names of specific individuals targeted in reporting our findings. When, however, the target was a public figure, for example a politician, we have kept the name to show the extent to which such figures are publicly attacked with race-related hate speech. We have further decided to remove all pictures and memes. The reason for this removal is to avoid sensationalism and to avoid the repetition of hateful contents that often use the photographs of real people. In Stage IV, we obtained oral consent for the recording and use of the discussions in this research project. All students remained anonymous.

42 However, we will share the dataset with researchers upon request.

Findings and Discussion

Findings and Discussion

Stage I: Defining Racist Hate Speech

Given the complexity of the debates on hate speech, we do not pretend to have a ready-made solution. However, we note that, notwithstanding Matsuda *et al*'s invaluable contributions, the debate has not heard from those at the 'forefront' of dealing with racist hate speech and its fallout. Precise definitions of hate speech (i.e. definitions that prioritise the element of imminent threat and direct violence) may be necessary to define the remit of judicial intervention and to focus monitoring efforts on speech that is 'dangerous', especially in contexts of socio-political instability (Benesh, 2011; 2012; Pohjonen and Udupa, 2017). Similarly, a set of clear rules may be necessary for social media and their users in order to manage the circulation of contents in these platforms. On the other hand, more dynamic and nuanced definitions of racist speech have the advantage of shedding light on the cultures of hate that proliferate online, assisting in gathering information on worrying trends with the aim of shaping policy interventions and educational initiatives.

Dynamic and nuanced definitions of racist speech have the advantage of shedding light on the cultures of hate that proliferate online, assisting in gathering information on worrying trends with the aim of shaping policy interventions and educational initiatives.

This is why we turned to civil society, and specifically to groups and individuals with an involvement in experiencing and/or addressing racism in all its forms. One of our main findings here is that the various expressions of racist hate speech, from illegal hate speech to thoughtless 'banal racist' comments, tend to exist on a continuum of discursive online toxicity and reinforce one another. This is directly derived from the focus group discussions and interviews we had with the project's informants.

The civil society view and definitions of online racist speech

During the focus groups' discussions and interviews, the challenges of defining what constitutes 'racist hate speech' and 'online racism' were debated at length. The general consensus was that 'hate speech', as defined in legislation and the community rules of social media platforms, fails to capture a broad range of 'everyday' racist discourses and race-talk that circulate in online environments. Research participants noted that extreme forms of racist hate speech – crude epithets, racist slurs, grossly offensive and dehumanising utterances – tend to be 'easy' to identify. This type of racism has a kind of compelling visibility and is generally frowned upon by the majority of online users as expressive of bigoted beliefs and ignorance. Dehumanising language is especially important to note as it is linked to genocide. Stanton (2004), who researched the Rwandan genocide, identified eight stages of genocide, placing dehumanisation at number three[43].

However, online racism manifests itself in a variety of more or less 'coded' discourses, which often do not make explicit reference to 'race', narrowly intended as a descriptor of skin colour or other phenotypic or observable features, such as hair or eye colour and so on. This racism without race (Bonilla-Silva, 2003)[44] or 'cultural racism' is often normalised through ordinary speech rather than extremist speech.

43 According to Stanton (2004), the stages of genocide are: classification (society is divided into different categories of people, 'them' and 'us'); symbolisation (these groups are given different symbolic attributes, for example through identity cards or in Nazi Germany, the golden star); dehumanisation (the targeted groups are likened to vermin, animals, or disease, and in general their humanity is removed); organisation (where the hate groups become more organised); polarisation (where any middle ground between 'them' and 'us' is effectively removed); preparation (where concrete plans are made); extermination (the actual killing of people en masse); denial (the refusal to accept and recognise what has happened). We are not claiming here that anything like this is occurring or even likely to occur. But we also need to be aware of these stages and the links between dehumanisation and genocide.

44 Racism without race is not a new phenomenon as, for example, the racialisation of Jews in Europe and elsewhere, Travellers in Ireland, and Irish, Italian, Greek immigrants in the US shows. A focus group participant, who is a Traveller, said : *'I think in Ireland with Travellers, you don't need a different look, a different skin colour... because straight away the Irish people can recognise you as Traveller, I don't know how that happens... they can just give a look at you... so you are faced immediately with discrimination'*

[…] Some of the terminology, especially on Twitter… sometimes you go into a conversation… and you follow a conversation and it may take a long time before you figure out what's the origin of it, what are they actually saying, who is saying what… because it is unclear… and it may be through a complete search down of something that you may find out that someone is actually perpetrating hate speech because it is not immediately obvious… so it may be difficult to pick in terms of specific terms… (focus group)

I think as well the homeless crisis that we have had here in Ireland… I think a lot of people justify their comments by saying 'what about our own'… 'our own homeless people', 'why are we not looking after them first?'… 'surely we should look after them first and then'… I think that argument has been used, particularly in the general public as a way of justifying hate speech… (focus group)

There is a vast body of literature on race and racism, which firmly supports the epistemic validity of participants' definitions. Critical scholars in the areas of race and ethnicity (Barker 1981; Back and Solomos, 2000; Balibar and Wallerstein, 1991; Bonilla-Silva, 2003; Essed, 1991; Fredrickson, 2002; Gilroy, 2000; Goldberg, 2008, 2015; Hesse, 2004; Lentin A., 2004; Lentin and Lentin, 2006; Miles, 1982, 1989; Omi and Winant; 1986; van Dijk, 1993; for the Irish context see in particular Carr, 2015; Garner, 2004; Lentin and McVeigh, 2006; Michael, 2017) have all argued that, in the past few decades, racist discourses have tended to mobilise notions of culture, ethnicity, religion, that while non-racially specific, are seen as having quasi-biological properties. **Participants pointed out that both crude and coded references co-exist, while there is also a sense of escalation and learning or adaptation.** These instances often do not breech social media platforms' guidelines:

[…] [a] few years ago, they were more openly… going like 'oh they should all be put in concentration camps', 'they are scum… they are this and that'… but now they are… they may have a picture of Hitler as profile picture… but you go through the page and you cannot really report under the guidelines of the platform… there's nothing you can say, like 'ok, they are attacking Muslims… they are attacking Travellers or black people'… it is obviously racist but… (focus group)

Participants pointed out that social media users and trolls in particular have become more and more skilled at evading possible accusations of racism as well as circumventing hate speech community rules by **using slang, circumlocutions, irony, and ambiguity.**

what she was doing was, every time she was writing really racist comments, like 'all Travellers should be deported to Ireland and put in concentration camps'.. question mark at the end… and all her posts that were really racist had this question mark at the end… so she can say 'well, I was asking a question…' so trolls kind of become more 'educated'… they have become more and more sophisticated in the way they use hate speech online… (focus group)

At the same time, **crude racism seems to be making a come-back**, this time supported by pseudo-scientific references to genetics. Another informant, a Traveller activist, mentioned the case of an anti-Traveller Facebook page:

'this page was just putting up everything negative on Travellers, just like racist debates 'are Travellers even human?', 'are Travellers Neanderthals?'… all this kind of stuff… and debates about DNA stuff and genetics… like our brains are not able to absorb information and you know… all this kind of stuff and you get a message from FB saying 'it doesn't breach our community standards'… (focus group)

Participants also noted that **white supremacist ideologies and their vocabulary have become more widespread**, not only among Irish groups linked to them, but also in comment threads.

I think what I have been seeing for the last few years… it is racist concepts that have originated in the pages you were listing… Celtic Warriors types of pages and Youtube videos that maybe four years ago had fourteen views… I now see them replicated in comments by members of the public… one recently was this population replacement concept that was kind of 'niche' at the beginning and now comes up constantly (focus group)

Research participants also suggested other variants of racist discourses, such as whataboutery ('what about our own'), narratives of elsewhere (e.g. 'look at Sweden'), use of bogus statistics ('80% of Africans are unemployed'), and metonymies (e.g. 'religion of peace' to refer to Islam ironically). Age-old anti-Semitic and Islamophobic stereotypes – e.g. the Jews as Christ-killers, and Christian crusades against Muslim invaders – seem to have been given **a new lease of life by the online right**, especially its Christian fundamentalist fringe:

'This notion [that Jews are responsible for multiculturalism] goes back to the old Judeo-Bolshevik conspiracy theory… that you know 'the Jews are trying to keep themselves apart from the white races and they are trying to destroy the white races by bringing in blacks, by bringing in communism', etcetera, etcetera… this is the classic right-wing conspiracy theory… which I thought was gone… but it has totally come back in… you can see it in the attacks of George Soros, because you know what George Soros stands for… it stands for international cosmopolitan Jews… so the idea that the revolution will be Soros-funded as a way of the Jews trying to destabilise white Christian culture… is really a long held trope… that is acquiring more salience right now' (focus group)

These discussions point to multiplicity and the ever-changing nature of the various categories of racist hate speech. In a discussion of the notion of racism in a post-racial context, Lentin (2016) distinguishes between *frozen* and *motile* varieties[45], which operate in tandem and further complicate and obscure the workings of race. Lentin points to the 'acceptable' cases of calling out the crude racism encountered, for example, in instances of public racism on buses, at work and so on as particularly obscuring how race pervades the very structure of society and therefore cannot be dealt with merely by focusing of performances of ***frozen racism*** [please see glossary and definitions]. Similarly, pointing to frozen racism overlooks the way in which 'motile' racism operates: as the taken for granted backdrop of societies that are structured on the basis of racial divisions. We are, therefore, actively seeking to avoid the practice of pointing the finger to 'racist' discourses and accounts, turning instead towards identifying the variety of discourses that seek to naturalise and subjugate racialised others and ultimately to uphold and justify the current racial order. For this reason, and based on our informants' understandings, we shifted the focus from 'racist hate speech' to *racially-loaded toxic speech*[46], in the hope that it will lead to a better diagnostic of the operations of race and racism and their pervasiveness in the online context**.** The following section explains how this was operationalised.

We shifted the focus from 'racist hate speech' to racially-loaded toxic speech, in the hope that it will lead to a better diagnostic of the operations of race and racism and their pervasiveness in the online context

Stage II: Building HateTrack

In designing the HateTrack tool, we deliberately adopted the term 'racially-loaded' content characterised by varying degrees of toxicity. The terminology 'racially-loaded' reflects the wide array of discourses we were hoping to capture through the algorithm including expressions of everyday, mundane race-talk online, an area of study which has been thus far neglected (Sharma and Brooker, 2016) but which is nevertheless important in maintaining and reinforcing a 'racial' common sense.

We define racially-loaded language as 'toxic' when it conveys messages that entrench polarisation; reinforce stereotypes; spread myths and disinformation; justify the exclusion, stigmatisation, and inferiorisation of particular groups; and reinforce exclusivist notions of national belonging and identity. Racially-loaded toxic language typically uses expressions and arguments that make certain words/concepts/images and the negative emotions they evoke – fear, disgust, or distrust – 'stick' to particular bodies (Ahmed, 2004). Expressions like 'rapefugees' or 'religion of peace' or 'bogus asylum seekers' are used routinely online and they serve to evoke a whole set of racialising assumptions about specific groups.

We define racially-loaded language as 'toxic' when it conveys messages that entrench polarisation; reinforce stereotypes; spread myths and disinformation; justify the exclusion, stigmatisation, and inferiorisation of particular groups; and reinforce exclusivist notions of national belonging and identity.

Our understanding builds upon the notion of toxicity as found in biology, where it is defined in terms of the degree of harm it causes. Definitions of what constitutes online toxicity vary greatly: some definitions refer to toxic language as language that is uncivil, aggressive or rude, while others focus on the demeaning or stereotyping content of a message irrespective of the language used (York, 2017). From our point of view, toxicity does not describe the words and style used to express an argument but refers to the specific content of online expressions and the ideologies shaping them. Indeed, our examination of the materials collected shows that 'counter speech' – speech that seeks to respond to racism – often contains strong language or ad-hominem attacks; it is language saturated in emotions including anger, indignation, and hostility. In some ways, anger may be the most civil response in the face of hateful or racist rhetoric. Focusing on toxicity as lack of

45 See glossary and definitions

46 See glossary and definitions

civility, as for example, the machine learning tool Perspective[47] is doing, would miss these responses, as well as those contents couched in civil and formal terms, but which are nevertheless toxic in causing harm and justifying exclusion.

The HateTrack tool was not designed as a means for policing or censoring specific racist content or for identifying potentially illegal material. Neither was it intended as a tool for labelling specific individuals or specific statements as racist or not-racist[48]. Rather the aim is to use the tool as a monitoring tool and a diagnostic of the current state of the Irish digital public sphere with respect to racism. We used it to harvest a dataset which may help shed light on the type, severity and recursive character of racially-loaded toxic content online and to contextualise such content through a broader analysis of online hate.

In short, in parallel to existing monitoring projects, the HateTrack tool aims to identify racially-loaded toxic contents in the Irish context, and to seek to understand the scope, spread, and forms that this takes.

This tool 'scrapes' Twitter and Facebook and allows users to track keywords, pages/accounts, or specific posts.

It then automatically classifies contents in terms of the probability to contain racially-loaded toxic contents (1=high, 0=low).

The tool further allows users to manually enter their own classification of the material as high or low probability, which can be used to feed back and improve the accuracy of the algorithm.

Finally, users can save, download and export the results in a spreadsheet format, allowing further processing.

At present, the tool is only available on DCU servers and is password protected. In the future, we plan to continue hosting it at DCU but offer the possibility for external users to use it through a registration process.

Stage III: Analysis of the Dataset

The complexity of the theoretical framework and methodology underpinning our study, alongside the large amount of data collected, makes the task of interpreting and summarising our research findings rather challenging. The analysis is based upon the dataset including the content manually classified with the purpose of training the algorithm; and the contents harvested through the HateTrack tool. We hope that this study may contribute to understanding the range and pervasiveness of online racism – both the crude 'frozen' type and the subtler varieties that tend to pass unnoticed. While our dataset allows for some modest generalisations, it should be noted that further research is necessary to support, validate and extend these findings, or to modify and refine them.

Overall, and as noted by our informants, we can observe that hate speech is pervasive on Facebook and Twitter and runs the spectrum from crude biological racism and white supremacist views, mostly encountered in specific Twitter accounts to coded or common sense racism most typically found in the comment sections of the Facebook pages of the Irish Times and Journal.ie.

1. **Crude and coded forms of racism utilise different discursive strategies**, including grammar, semantics, style of argumentation. The cruder forms typically employ insults, slurs, profanity, animal comparisons, direct denigration, or appeals to well-entrenched racial stereotypes or 'race science' myths.

Too many uninvited and unwanted bogus, smelly immigrants and fake asylum seekers

Ireland 2 bcome crime-ridden darky shthole 70,000 eyeballing liar muslims here already &40 fakeKid rapefugees on way

that's why Marxism in all its depraved manifestations are best left in the century gone by. If you feel adding a few 70 IQ sub-saharan Africans into your families genepool is evolutionary progress then off you pop

47 Perspective can be accessed here: https://www.perspectiveapi.com/#/

48 Writing about the Charlie Hebdo shooting, Gavan Titley (2015) makes a powerful case for the need to 'to critique racism without the reductive certainty of categorising racists and anti-racists'. Available at http://www.irishleftreview.org/author/gavan-titley/

Coded racism relies on supposedly race-neutral principles like culture, values, ethnicity; and employs seemingly well-reasoned or common sense arguments (e.g. the need to distinguish between genuine refugees and economic migrants; taking care of 'our own' first).

We shouldn't be housing Africa's surplus population, let's house our own people first

Its not racism thats going on in Ireland its survival of the fittest. Providing housing, benefits and education for foreign nationals over our own causes people to lash out

Read our proclamation, every man woman and child be treated equally. The floods of immigrants being allowed into Ireland is ridiculous. What happened to taking care of our own first

2. **Racially-loaded toxic discourses often coalesce around notions of 'Irishness' and what it means to be Irish**, with specific individuals and groups being targeted directly, often through ad hominem attacks, or indirectly. Ethnic minority Irish people, especially if they have a public profile, have been and remain at the receiving end of racist hate speech, as evidenced in the case of the #WeareIrish campaign[49] and the abuse directed at Ibrahim Halawa and footballer Cyrus Christie. The #WeAreIrish campaign for instance, while receiving widespread support, also attracted a number of racist tweets from users (many based in the US) trying to hijack the hashtag. Some of these read:

#WeAreIrish Is like me saying im indian because i dont have a toilet lol

The patriots of Ireland martyred themselves for the Irish people – oriental deracinated transplants are NOT Irish! #weareirish #paperIrish

There is now an active campaign to flood Ireland with biotrash from all over the globe to atone for being too safe and white.

Narrow articulations of 'Irishness' can make explicit reference to race, ethnicity, and/or religion - positing Irish identity as exclusively white and Christian (and specifically Catholic)- or are expressed in more generic formulations such as 'we have to preserve our culture'.

[being Irish] does not mean just being born here. It means both parental genetics is W. European & at least 1 parent being descended from Irish

you look African, you're not Irish. Irish people are white

As an #Irishman you are the oldest and Whitest of the Aryan peoples, the Irish are the furthest there is from black, brown or yellow peoples

3. **Calling out racism in online environments typically leads to accusations of being 'over-sensitive' or 'playing the race card', or 'being racist' against white people.** Discursive retorts such as 'you're being too sensitive' or 'why are you bringing race into this?' function in two ways: firstly, they silence or undermine the grievances of minority and ethnic communities treating racism as a problem of the past, all the while 'recycling' old racist tropes via a more civilised vocabulary. Secondly, the casting of Blacks, Muslims, or Travellers as profiting from cultures of victimhood erase these communities' long histories of political, cultural, and grassroots mobilisation and their hard-fought battles against institutional and State discrimination.

Foreign nationals always, always, always play the race card even when they are downright rude and belligerent themselves, it's just too easy to play that racism card so that they can get what they want.

agree 100% why are immigrants given first choice? Because Irish society doesn't want to be called racist. So much huff about racism that we are forgetting to house our own it's so sad

stop blaming everything on whites. the fact is it was white people who ended slavery for all and whites definitely have a lot to be proud of

49 The #WeareIrish tag and campaign began as a means for celebrating the diversity of Irishness. The campaigners were directly targeted by racist hate speech and the campaign generated a multitude of racist tweets, comments and memes.

4. **There are clear patterns of shared language between international hard right and alt-right groups and parts of the Irish digital public sphere**, including the adoption of racist ideologies such as 'white supremacy' produced in the context of the United States and the Identitarian movement originating in France[50]. Particular expressions like 'white genocide' and 'population replacement' and references to a 'globalist conspiracy' with a clear anti-Semitic streak have spread to the Irish context.

Just in case anyone thinks Jews aren't involved in the replacement of the white race

Deport them all. Ireland is finished if they keep letting in invaders #KalergiPlan White replacement

Screams of racism are efforts to stop an ethnic group asserting their right to protect themselves and their rights of sovereignty over their ancestral ethnic territory and thus are complicit in the genocide of that ethnic group

'Identitarian' ideologies use the seemingly neutral vocabulary of ethnicity, ancestry, and genetic difference to advance both white supremacist arguments and 'ethnic tribalism'. This draws upon primordial ideas of the ethnic or the nation which posit that ethnicities are socio-biological entities that offer important evolutionary advantages hence people's attachment to them (van der Berghe, 1978). In the popularised version of this idea, people are seen as naturally belonging to different ethnic/racial groups and should live separately to preserve their cultural and biological uniqueness and specificity.

The Irish are a people that share a common heritage that's unique to them. The same is said for the Polish. Just ask the Polish and how they are not Greek. When a people share a common heritage a common ethnicity and generally look like one another. That's what makes them, them. Just go to Pavee point and they'll agree with that. Now a nice guy like you wouldn't deny a minority their heritage now would you? … Human nature boils down to this. Identity and the want for territory. If a people are unwilling or unable to defend their land people will just come and take it

Identitarian discourses tend to naturalise and normalise hatred by presenting it as the inevitable result of illegitimate attempts to mix and amalgamate primordially incompatible or distinct groups.

Flood a country with open-door mass immigration and tensions are bound to pick up… Working class communities have been ravaged by the scourge that is ''multiculturalism''. Our very culture, history, heritage and identity is at stake, and there will eventually be a large pushback'

Syrians, refugees, Islam etc not at fault for #Dundalk attack, no more than you can blame a pitbull for savaging a child, when it merely acts on its nature. Ultimate responsibility lies w/the owner, in this case, the political class #RefugeesWelcome brigade. Blood on your hands

Identitarianism and ***primordial nationalism*** [see glossary and definitions] are sometimes bolstered by nationalist and sectarian sentiments, a mix which can be considered idiosyncratic to the Irish context. For instance, migrants and refugees are seen as the new settlers - thus rendered equivalent to British colonialists in the 17th century – intent on establishing a new 'plantation'.

The new Plantation, Zionists invade Ireland, Anti-White, Anti-'Christian', the new Federal Europe of ONE identity

Stop stealing our money to finance the Invasion and Plantation of Ireland by a migrant horde. What you are doing will cause more violence and war in Ireland than the Plantation of Ulster did

Look around you, the natives are only 4 out of 5 of the population here and shrinking. It's a plantation.

The eclectic nature of these discursive constructions testify to the fact that some aspects of Irish online racism partake in a global political movement while infusing it with local, historically specific inflections.

50 The Identitarian Movement was founded in 2012, after a split from the Bloc Identitaire, when the youth part of Génération Identitaire decided to go its own way. Soon, other similar groups emerged across Europe: the German Identitäre Bewegung (Identitarian Movement), the Austrian Identitäre Bewegung Österreichs, the Italian Generazione Identitaria, Generation Identity United Kingdom and Ireland, but also the US-based Identity Europa, all with the same anti-immigration, anti-multiculturalism nationalist agenda. Their symbol is the Greek letter Lamda, found on the Spartan shields (standing for Lacedaemonia, another name for Sparta) and alluding to the battle of Thermopylae and King Leonidas (Virchow, 2015). The Identitarians are against the EU and against what they perceive as the old and corrupt ruling elites of Europe that push a multicultural agenda. The references to the 'Kalergi plan' refer to Richard von Coudenhove-Kalergi and his ideas of a pan-European Union and a mixed-race future.

5. **Racially-loaded toxic discourses feed on fake news, bogus statistics, research published by institutes with dubious credential and 'recited truths'** (Lentin and Titley, 2011) coalescing around the alleged failures of multiculturalism, no-go Muslim areas, and African youth gangs terrorising locals.

If immigration is a good thing, why did a 100-strong African gang rampage through Lusk recently causing it to go into lock-down? What forced 700 people come out and protest African gangs in Balbriggan? Why was an African gang leader's home fire-bombed in Tyrrelstown?

Blackbriggan needs Right Wing Death Squads #BeyondThePale[51]

Look at sweden. ....over 50 No Go areas....and all started by anti social behaviour perpetrated by people who they had welcomed into their country.....now they cannot even walk the streets in some areas of their own country

Some accounts in particular seem to be part of a densely linked network of right-wing, alt-right, accounts – some of these in the US and UK – all circulating and amplifying stories about migrants, Muslims, and refugees – through linking in powerful news sites (e.g. Breitbart, Infowars) and sharing and re-tweeting unverified or fabricated facts. We examined a small number of accounts whose sole *raison d'être* appears to be rumour mongering and posting negative stories involving Muslims or migrants found in mainstream media (BBC, CNN, Irish Independent, Irish Times, etc.) that are then shared in a deliberately misleading context. This is not false information as such but is framed in a mendacious way, with the sheer accumulation of 'incriminating evidence' serving a very clear racist agenda (Titley, 2017b). Some accounts also circulate images that are often manipulated, unrelated to the context, or again misleading (typically involving large groups of black or Muslim men). Unverified facts or misleading facts often end up integrated into comment threads, thus 'contaminating' the broader public sphere.

It should be noted here that Titley (2017b) cautions against the risk of seeing fake news as merely a problem caused by social media platforms and their dynamics. He contends that one of the reasons fake news works so well in racist terms is because the work of pointing out certain populations as a problem has already been done politically and not just by the right wing. For Titley, this has already been done by the political mainstream, and therefore it is not simply that there is an informational infrastructure through which this kind of material circulates successfully, but there is also a political and ideological infrastructure which means that people are predisposed to share it, to feed it as legitimate, and are predisposed to believe that it is legitimate to act on it because it has already been sanctioned in many ways by the mainstream.

6. **Facebook pages of news outlets and their comment threads seem to play an important role in channelling racially-loaded toxic contents**. There are a number of news topics that tend to trigger racist responses and commentaries, with Ibrahim Halawa, refugees, terrorist attacks, Direct Provision, Islam, and crime involving non-Irish nationals topping the chart. We define these as ***trigger events***[52]. News articles about Muslims, Roma, and Travellers appear to elicit dehumanising racism, irrespective of the article's context. **The way mainstream media frame and present news is likely to have an impact on the type of comments that are likely to appear**, with sensationalist headlines attracting a large volume of hateful comments[53].

The way mainstream media frame and present news is likely to have an impact on the type of comments that are likely to appear, with sensationalist headlines attracting a large volume of hateful comments

An illustration of this was found in the *Irish Times*, which in August 2017 published an article detailing the findings of a report on children living in Direct Provision under the questionable headline 'Children in direct provision complain about food, overcrowding'. Some of the comments under the piece, posted on the IT Facebook page, read:

'Ungrateful shower of freeloaders, send them back if they don't like it'

'We can't afford this bullshit anymore .. Ship them out and finally house the nearly 8000 Irish homeless'

'Round them up and repatriate them to their own country where they can enjoy their own food, ungrateful fcuks

51 'Beyond the pale' is used to denote the boundaries between civilised and non-civilised world, between what is acceptable and what isn't. In contemporary language, it can also be used to critically refer to the Government allocating all its resources to Dublin, leaving other areas unprotected.

52 See glossary and definitions

53 This may be because people tend to read headlines but not necessarily the whole article. In a 2016 study, Gabielkov et al. found that 59% of shared news items the link hadn't been clicked, meaning that users never read anything beyond the headline.

7. **Expressions of racism online are punctuated with misogynist, homophobic, and transphobic attacks directly targeting specific women and members of the LGBT community in general**.

Varadkar being half-Indian erodes the national identity of Ireland. And being a homosexual means he is immoral & mentally ill.

[Trump is] Just what the West needs. A real man. Feminisation of the west means, the west will become far more emotional, far less intellectual, far more submissive, less likely to rebel against oppressors. Look around Europe and it becomes apparent. We are losing our identities and culture to alien cults

shitskins and middleaged women that no sane european man will fuck them so they wait for the rapefugees to take care of them

Debates around abortion and reproductive rights constitute another arena for racially-loaded toxic speech targeting pro-choice campaigners and anti-racism activists – as conspiring to bring about the demographic destruction of the Irish race.

I don't see much potential with the Irish women these days, who are hurriedly having the next generation flushed down into sewers as clinical waste in English abortion clinics, so maybe traditional Polish women do have a place ;)

Thousands of #SoyBoy and #refugeeswelcome cucks bravely battle strong evidence of migrant terror today. Normal service resumes tomorrow with familial bleating about "muh bodily autonomy". #Dundalk

8. **Social media affordances and tropes lend themselves well to racially-loaded toxic contents, which can include memes, multimedia materials, hashtags, tagging and other forms that allow the materials to travel further.**

The expressive possibilities afforded by social media – especially share and re-tweet buttons, the use of memes and hashtags, and the ability to upload pictures and videos – means that racist discourses can be expressed in a variety of non-textual formats.

The term 'meme' comes from Richard Dawkins' (1976) book The Selfish Gene, and is defined as small cultural units of transmission, which, much likes genes, are 'replicators' that spread from person to person by copying or imitation. In digital culture, memes are defined as instances of digital content that share common characteristics of content, form and stance, which spread quickly and become a shared cultural experience (Shifman, 2014). Image macros, where an image is superimposed with text, are the most common forms of memes. Because they are easy to recognise, relatable and often funny and cleverly done, these kinds of contents are more likely to spread. In this research, we came across various racist memes, often with distorted or unflattering pictures of people of colour, and accompanied by ironic hashtags. In general, visual elementss tend to be recalled faster than audio or text and retention for images is better and more accurate compared to verbal and textual information (Stranding et al., 1970). This is important to note here because it implies that images of hate may be more pernicious than words alone.

The practice of hashtags on Twitter can serve to re-signify or re-contextualise either ironically or metonymically seemingly neutral content in ways that activate racist inferences. For instance, links to news articles reporting on criminal cases involving ethnic minority individuals can be accompanied by the hashtag #refugeeswelcome or #culturalenrichment. These hashtags, typically used to suggest 'humour', 'irony' or express a 'factual observation'– are a key strategy of denying racist expression and propagating the ambiguities of race talk (Sharma and Brooker, 2016)

Zappone must be checking to see if all her Syrian children are at home reading up on transgender theory. #Dundalk

Arab #Muslim terrorism has arrived in #Dundalk #Ireland #RefugeesWElcome1 dead 2 injured

The new irish integrating in local events what wonderful doctors and engineers we have invited to #Jobstown #Tallagh #LidlLooting #Lidl @AMDWaters

Targets of racially-loaded toxic discourses

While the above analysis presented the main tropes and forms of racially-loaded toxic discourses, the section below discusses the dataset in terms of the communities targeted. These include the following groups in no apparent order: the refugee and migrant communities; the Muslim community; the Traveller and Roma communities; the Jewish community; the Black community; and second generation Irish people. In this, our findings add to the existing body of research on race and racism in Ireland that has so far identified anti-Black, anti-Muslim, and anti-Traveller/Roma racism – in parallel to anti-LGBT and disablist hate (see Haynes, Schweppe and Taylor, 2017; also above pp. 11-13). At the same time, these findings support the recent findings by IHREC-ESRI and the focus on 'whiteness'.

Anti-refugee and anti-immigrant discourses

The majority of anti-refugee and anti-immigration discourses mobilise three inter-related tropes:

1. access to welfare and housing;
2. moral deservedness; and
3. the good versus bad immigrant.

Access to welfare and housing

The first trope shows that socio-economic anxieties tend to be conflated with notions of national identity and race. Online debates on refugee quotas are often punctuated with comments that cast migrants as unworthy recipients of public funds and to be blamed for the current housing crisis.[54]

Ireland needs to close its borders and start vetting ppl... Send all refugees home.... Not only would it make the country safer by knowing who is in it but it will also resolve some of the housing problem the hospitals been under pressure and it will lower the weekly welfare needed .. Time people got their PC heads out of their asses and looked after OUR country and OUR problem… THEN we can think about extending our charity to others

The welfare tourists that abandon their own countries and want a ready made answer to their economic problems on the backs of decades of Irish nationals who worked hard to progress

Some of these ethnic people are robbing us left right and centre. Social scams, rent allowance and some bringing 3 and 4 wives into our Country and getting dole and every benefit going

Moral deservedness

Discourses about refugees are often couched in terms of a polarisation between 'people fleeing wars' who may deserve protection and economic migrants or bogus asylum seekers taking advantage of the system. Generic expressions like 'send them back', 'don't let them in', 'deport them all', are particularly widespread. The casting of asylum seekers as 'African immigrants' and 'welfare shoppers' shows that articulations of illegality co-exist with forms of racialisation premised on bodily features.

I wouldnt even bother with 'fit in or fuck off', they should just fuck off – on a leaky banana boat preferably for all the trouble they made!

Too many bogus "asylum seekers" coming here, illegally, through other safe countries. Time to call their bluff, get tough and kick them out. They are unwanted, unnecessary, unwelcome and, in the case of most mohommedans, parasites who are unwilling to work or to integrate

The Irish military should not be being forced to act as a ferry service for smugglers and illegal economic migrants in the Mediterranean

Good versus bad immigrant

A trend consistent with discursive racist repertoires found in other countries is the good migrant-bad migrant dichotomy: if migrants are perceived as being, for instance, Roma, Muslims, or from Africa, they are stereotyped as inherently lazy, breeding too much, sexually rapacious, bringing diseases, etc. The moral deservedness of refugees/immigrants is for example discussed by Holmes and Castaneda (2016) in the context of Germany, where they found that mainstream media typically make distinctions between refugees from Syria, who are seen as 'deserving' and 'economic migrants' from other countries in the Middle East and Africa, who are seen as undeserving despite the fact that they too may flee war or conflict. This stereotyping language is racialised in as much as it attacks specific moral traits to specific bodies but is also gendered as it targets and pathologises men and women differently.

Hey you guys, in the hazmat suits...Do you fear contamination? What about us? These illegal African men, bring us numerous diseases

It is now illegal to buy sex in #Ireland http://jrnl.ie/3309170 That won't stop immigrant prostitutes from coming. (pun intended)

THESE APES ARE NO REFUGEES, Fully Trained Muslim Troops, Ready to get their Raisins in Heaven. The Enemy is RELAXING, WAITING FOR ATTACK!

In terms of content, we notice a difference between 'everyday' anti-immigrant rhetoric and the types of discourses put forward by certain political groups, with the latter often coalescing around the notions that multiculturalism has failed, and that 'diversity' is inherently bad or socially corrosive; such views are also infused with a nostalgia for an assumed ethnically homogenous past. These notions betray a broader ideological agenda combining anti-Muslim racism, sexism, nationalism, and racialised definitions of Irishness.

The reason why multiculturalism exists is to pretend that inferior cultures aren't inferior and that superior cultures aren't superior. It's a way to tell nice lies about rotten cultures and rotten lies about great cultures

54 It should be added though that the housing crisis has also become a key arena for the expression of anti-racist ideas and grassroots solidarity projects.

the whole of Europe needs to unite and rid the continent of the cancer that is #islam and # multiculturalism

#dublintogether – is this your future? hiv aids pakis raping your women? #enrichment

Anti-Muslim racism and Islamophobia

During the period this study was conducted, terrorist attacks in Europe, the Ibrahim Halawa case, and the Dundalk stabbing were the key news events shaping online discursive interactions around Muslims and Islam. Anti-Muslim discourses occasioned by these online debates can be classified into four broad categories: terrorism; clash of civilisations; Muslim men as misogynist and sexually deviant; and a general and unspecified antipathy.

Firstly, **discourses around terrorism identify all Muslims as terrorists** and cast any crime committed by a Muslim as a terrorist act.

In addition, his theory about Muslims biting the hand of those that feed them is sadly very true. Ireland will not wake up until there is a terror attack here

Mass deportation, get all these fucking parasites off welfare and deport them they FUCKING HATE THE WEST it's obvious when another savage kills innocent civilians....

#Dundalk looks like we finally got it, some fucking sand monkey finally attacked us with intent on killing as many as possible

Clash of civilisation narratives typically construct Muslims and Islam as a threat to European values of democracy, civility, and enlightenment.

Islam actively promotes this. There is no "white supremacy" religion that promotes murder of innocents. A serious conversation needs to be had about the compatability of Islam and western civilisation

these rats who follow the pedophile Mohammad certainly have no place in Ireland or any other country outside there own shitty hell holes... Let them rape and behead until there satanic hearts content in there own sewer infested CUNT REES.

Islam is an existential and real threat to the European way of life. I commend you for speaking out as you are entitled to.

Pseudo-feminist and sexualising discourses represent Muslim men as barbaric and abusive towards women, as having abnormal sexual proclivities, and a repressed sexuality.

Give it a few years & we will have cases in courts of Muslim sex grooming gangs. People need to educate on what we face

The dirty perverted Muslim men involved WITHOUT DOUBT..recorded the evil gang-rapes of these very young vulnerable girls

whilst your average moohamiden has 3 or 4 wives and at least a dozen mini moohamiden's

#Lisdoonvarna. This is simply a kick start to the old matching tradition there. " Abdul, you look like a man in need of a fourth wife "

Finally, what appears as a ***sui generis* anti-Muslim racism** is expressed in antipathy, dislike, disgust, and aversion.

People in Europe do not believe all Muslims are terrorists. the reason they dont respect them is the fact that Muslims living in France England Holland, and Belgium are usually on Welfare, have too many children, get involved in petty and/or violent crime

Muslims as a whole are nauseating people

Bungee jumping.1000ft drop. £35 per person. Muslims go free. No strings attached

Anti-Roma and anti-Traveller racism

Discourses stereotyping, dehumanising, and denigrating Roma and Travellers are pervasive. Keeping in mind the structural inequalities that Roma and Travellers suffer, the damaging impact and possible corollaries of hate speech on these groups' ability to feel safe, be part of society, and enjoy equal status cannot be under-estimated. Twomey's (2017) discussion of the connections between Facebook pages against Travellers/ Roma and street violence is an illustration of this dynamic. On the other hand, an important dimension to note here concerns the difference between direct attacks against individuals of Traveller or Roma ethnicity and generalised anti-Traveller racism. One of our Traveller informants made the point that 'online no-one knows I am Traveller' highlighting the face-to-face racism that they experience. At the same time, events such as the recognition of Traveller ethnicity or the Carrickmines fire trigger generalised attacks. Typically, Traveller and Roma people are **targeted as undeserving, 'uncivilised', thugs and criminals; they can further be targeted using a dehumanising language.**

Travellers trying to make a quick quid no surprise there. They are thugs like the are betrayed and the women usually dress like hookers from a young age

Reminder that knackers are a foreign people, ethnically distinct from us Irish, and that this is recognised by the Oireachtas

I always considered knackers to be like Muslims and Nigerians they get almost everything handed to them

Anti-Roma discourses in particular can contain bestial metaphors and other dehumanising language. One post in a Facebook business page we were directed to by one of our informants attracted hundreds of comments of this tenor:

ther rats from sewers

Voted the The most hated animal by the whole of the human race .. Yes that is correct the Romanian gypsies

Dirty foreign smelly cunts should be burned out

Anti-Semitism

Anti-Semitism remains pervasive. In our dataset, it takes mainly three forms: it is often woven through anti-immigration discourses that depict the cosmopolitan and rootless Jew as **the agent of globalisation and the 'hidden hand' orchestrating international finance and 'mainstream' media** (Linehan, 2012); it reproduces the figure of **the Jew as Shylock**; it **constructs Jewish people as 'unassimilable'**, or what Gellner (2008 [1983]) referred to as 'inhibitors of social entropy'; and it **denies the importance and magnitude of the Holocaust**.

time magazine a kike owned shit rag

Oven dodging kikes run Twitter censor that ya fucking faggots

big people are behind it like #rothschilds #goldmansachs all zio jews of course that worship the devil...they want to tear us apart.....

You mean "shush" don't destroy the revenue source funding Jewish world hegemony, while they continue with their campaign to destroy Whites

Age-old racist stereotypes of Jews as scheming merchants, greedy, nit-picky, and Christ-killers are not uncommon:

So called Alt-Right people are shilling for Jews. These kikes must never be trusted

American jews pretending to be socialists. Jews can never be real socialists. Judaism is a mercantile religion

One day we will shut their dirty lying jewish mouths! Our patience has its limits!!

A final anti-Semitic current seeks to **minimise or deny the Holocaust** or to caricature it.

Every Jew I've ever met or read about or had to endure listening to, had a relative in Auschwitz. How big was that party.

Ah sure are we not allowed to call them all big noses anymore or what? :p Big bunch of PC pussies wanting to lock people up and destroy them over a few words ffs

There is No scientific evidence to say anyone died in a gas chamber. It's about time that a full scientific investigation to prove exactly what happened and how,the dead are worth that more than ridiculous false claims made by people with hidden agendas surely.

Anti-black racism

Anti-Black racism cuts across some of the categories above, in particular the anti-refugee/anti-immigration and Islamophobia, as well as the attacks against second generation Irish people, as Black people can be targeted as refugees/migrants, as Muslims and more broadly as not belonging. But it is important to further identify the specific ways in which Black people are targeted. **The ways we identified in our dataset include the trope of criminality; the trope of being uncivilised, lazy, parasitic; and the dehumanising trope of African men as animals.**

In terms of criminality, the main references are to 'African gangs' as for example below:

If immigration is a good thing, why did a 100-strong African gang rampage through Lusk recently causing it to go into lock-down? What forced 700 people come out and protest African gangs in Balbriggan? Why was an African gang leader's home fire-bombed in Tyrrelstown? https://t.co/M67aS3r8Vt

1000 residents of Balbriggan, Dublin attended a protest over the weekend against the African gangs plaguing their community. Well done!

African men are particularly targeted as being lazy or parasites:

The African immigrants no women no children just young lusty males in their droves ready to do anything for Islam except work of course

Traumatised, starving African children, stay in Africa. Instead, thousands of African parasites & predators like these arrive in Europe daily

Big nose, big lips, big appetite for social welfare

Africans are further constructed as uncivilised and uneducated:

Was Macron right about Africans? That they have a "civilisational" problem...

Why should I feel pity for ARMIES of uncivilised anonymous, fit, young African & Arab Muslim men invading our peaceful, civilised continent?

You won't get an answer either, even for sub-saharan standards he's thick. In any case there's no point debating Irish affairs with African blow-ins who married their way into the country.

Dehumanising comments and especially comparisons to animals are used to establish the 'inferiority' of Black people and the 'need' to raise concerns regarding the purity of the white race. There is an evident misogynistic element here, as partners of Black people are directly targeted:

it seems a lot more common that women are willing to sleep with monkeys then men. You're much more likely to see a black man with a white woman in Europe than the other way around.

The parents of Ireland should be very concerned about the kind of porch monkeys we're letting into this country

Based on all this, violence or calls to violence are made justifiable:

I want to go outside and start punching random black people while wearing the flag of Ireland as a cape

Second-generation or mixed-background Irish

Second generation Irish people are specifically targeted in terms of their lack of any biological or ethnic connection to Irish-ness. Their claims to belong are dismissed and Irishness is constructed in exclusively White terms. The two main ways in which this group is targeted is firstly through the trope of **population replacement or colonisation** often using this community to make political points; and secondly, through making a distinction between **'real' Irishness, which is based on a 'biological' and 'cultural' bond and Irish citizenship which is a kind of 'fake' Irishness**. What is striking here is the use of rhetoric associated with identity politics and anti-colonial politics to attack any claims of this community to belong to Ireland. This identity politics from the right is directly linked to the Identitarian movement and rhetoric.

The trope of **colonisation and replacement** is explicitly political and tries to score against so-called liberals and multiculturalists:

Liberals: Colonialism was wrong!
Also liberals.. Africans can take over Ireland! It's fine! Pick one, idiots! #WeAreIrish

#WeAreIrish is about making indigenous white Irish people a minority in their own homeland, despite never colonising or enslaving anybody

Irish people are Irish. They are an indigenous people native to the island. Stop appropriating their culture #WeAreIrish

The claims to Irishness are ridiculed and denied outright – second-generation Irish people are **only Irish in name or paper**:

Who are these extra million people? The Irish birth rate is below sustaining, our population should be decreasing. Or is this the beginning of a new plantation? Brits out, everyone else?! #paperirish

They're not new irish. They're africans with irish passports. #paperirish

We all know these people aren't Irish, in fact their only form of "being one of us" is their passport. The #PaperIrish will pull any shit they can to claim this is their Homeland.

Who posts racially-loaded toxic comments?

We developed a preliminary typology of the posters of such contents. We argue that a crucial distinction needs to be made between organised political groups and ordinary online users in the way they produce and reproduce racially-toxic discourses. Unlike the latter, people or groups behind explicitly Islamophobic or anti-immigrant pages and accounts invest real labour, time, and resources in promoting the everyday circulating of racist discourses. They do so by carefully curating the content on their online platforms: spreading ill-founded stories (e.g. gangs of African youths terrorising locals in Balbriggan); misinformation; and attacking/harassing other online users. Strong similarities in language and the memes or links found in some accounts and pages may also indicate that these are managed by the same individuals, which again testify to their investment in spreading hateful messages. On the other hand, racist expressions found in comment threads tend to be more 'reactive' – occasioned by certain news content – and generally do not appear connected with any organised ideological project.

It could be argued that one of the functions of organised political groups is precisely to provide a set of ritualised scripts and 'merchandise' (links, memes, sources) through which racist hate can be channelled and expressed. However, this may be over-estimating the power of such groups in terms of shaping and circulating racist discourses: banal, everyday racist utterances may emerge from a 'psychic reservoir' (Davis, 2008) that is nourished by a broader range of more powerful actors – borrowing from the media, judiciary, academic and political institutions. This preliminary typology points to some general

categories of posters, making some initial distinctions, but more work needs to be undertaken in refining this further.

4. **Shitposters/gamers and trolls.** The mobilisation of racially-toxic memes and tropes here is mostly undertaken 'for the lulz[55]' (Coleman, 2014) or in order to 'bait' or annoy others – there is a clear performative dimension, where these posters are trying to cause as much outrage as they can.
5. **Cultural racists or nativists**. They are against diversity, Islam, migrants, all seemingly responsible for the cultural and ethnic genocide of 'Irish natives'. They use terms such as 'rapefugees' and welfare migrants and refer to the purported incompatibility between Islam and the West. They feel they have been wronged by 'globalist' governments embracing diversity and bringing in migrants to drive down wages or dilute whiteness. They are often staunch defenders of Christianity, and specifically Catholicism, seen as the foundation of Western civilisation and fiercely against pro-choice arguments and Irish women having full reproductive rights.
6. **Identitarians:** Separated from the above category as they are part of organised racist groups, forging links with right-wing groups in the UK and elsewhere and pursuing a certain political agenda.
7. **Contrarian/libertarian posters:** They are against migrants, multiculturalism, diversity, Islam, women but use a more sophisticated language (especially dark irony), alpha-male rhetoric, and have a clear and ugly misogynist streak. They appear to be media savvy and able to fully exploit the potentialities of new online platforms. Their shared hatred of feminism, the welfare state, political correctness, 'cultural Marxism', mainstream media, and 'normies'[56], and their peculiar aesthetic sensibilities have emerged from online environments and are influenced by alt-right social media celebrities.
8. **Everyday, casual, or banal racism**. This broad category encapsulates a wide range of online utterances that rarely make reference to 'race' or use demeaning and offensive language but still routinise racialising meaning. Everyday racism of this kind is expressed through comments such as *'we need to look after the homeless before we accept more refugees'*; *'Most Muslims are ok, but Islam poses a threat to democratic values'*; *'Migrants should adapt to our way of life'*; *'I have no problems with hard-working migrants'*, etc. This discursive variety of racism employs a stock of familial and 'common sense' arguments that may not be explicitly hateful but remain central to the collective reproduction of racist ideologies and their rationalisation.

Trigger events

Our analysis identified the central role of so-called **trigger events** in shaping the configuration (in terms of frequency and content) of online racist utterances, a finding consistent with research conducted elsewhere (Legewie 2013, King and Sutton 2014, Hanes and Machin 2014; Williams and Burnap, 2015). 'Triggers' can be events that have a transnational resonance (e.g. terrorist attacks, the Brexit referendum, the US Presidential elections) or national relevance (e.g. the opening of a Direct Provision centre in Lisdoonvarna). Legewie (2013) argued that events that construct an out-group as threatening or events that direct attention towards potential sources of intergroup conflict may trigger negative attitudes in response, at least in the short term. Online, such events, and their framing by the media can function to 'validate' and channel prejudicial sentiments, opening up a space for the escalation of cyberhate, the circulation of rumours and calls for retribution. On the one hand, trigger events seem to unleash and legitimise public expressions of hatred amongst 'ordinary' online users; on the other hand, organised groups capitalise upon the expediency of such events, intentionally circulating ill-founded scare stories, misinformation, and other narratives demonising migrants and Muslims. Trigger events can also be used to justify forms of more explicit, crude racism. An example of this is the stabbing incident in Dundalk, which was immediately construed as a terrorist attack because of the ethnicity of the person accused – see the case study below.

In more theoretical terms, Sharma's (2017) model of web racism is particularly useful to describe and understand the role of trigger events, showing the rhythms of racist talk online. Sharma argues that web racism has a power law distribution, and can be divided into (i) spectacular, highly visible racism, that follows highly publicised events, such as for example, terrorist attacks; (ii) explicit racism, typically triggered by milder events, for example, statements by politicians or public figures; (iii) ambient or 'long tail' racism, which forms a constant backdrop is not necessarily triggered by any events as such. What we can add to this model based on the current research is that the mid-range 'explicit racism' described by Sharma can also be triggered by media reports; specifically reports on topics such as immigration and refugees, housing and welfare or anything that has to do with Travellers or prominent second generation Irish people.

55 An online expression meaning doing something for a laugh, with no purpose behind it. Coleman (2014) has identified this seemingly purposeless behaviour as integral to the early hacking culture.

56 This refers to 'normal', conventional or mainstream people, and it is used in pejorative sense.

Stage IV: Attitudes Towards Reporting

In monitoring and regulating hate speech, social media platforms rely almost exclusively on users. It is, therefore, crucial to understand the circumstances under which users are inclined to report hateful content. To this end, in order to gain an entry point in the emerging cultures of reporting among ordinary users, we presented our findings to students and asked them whether they would be reporting any of the contents we showed them. This is the least developed part of the research, as we only had the time to conduct three of these group discussions. Nevertheless, the results are suggestive of a dynamic of under-reporting and we believe that they merit consideration. It is also hoped that these initial findings will lead to further research in this area, which evokes the responsibilities and duties of all of us towards hateful contents and our duty of care towards those targeted by these contents.

The main finding of these discussions is that people will not report racially-toxic contents even if they do recognise them as problematic. In fact, only three people admitted they had reported contents. Two of those reported materials that targeted them or their community (one self-identified as non-heterosexual and one as Muslim); the third reported a death threat, admitting that the bar was set 'ridiculously high'. In general, students acknowledged the moral or ethical implications of hateful words, but did not see these as connected to broader, systemic issues of inequality. They view racist outbursts as a form of social ignorance. Efforts to reduce misogyny or racism online, even when these efforts simply revolve around pointing at the pervasiveness of such contents, are viewed as censorship, a way of policing social media and stifling freedom of speech. Based on these discussions, we identified four 'cultures' that contribute to or justify the reluctance to reporting racially-toxic contents.

9. **Freedom of speech trumps everything else**; while respondents recognised the contents as problematic, they insisted that those posting these were entitled to their freedom of opinion.
10. **Racist utterances online are the preserve of idiots and bigots** who expose themselves as such and can be dealt with by others through taking them on and arguing against such views. In these terms, respondents felt that those posting racist contents were collectively judged as idiots and there was no need for any further action.
11. **Reporting hateful or racist comments is pointless because online racism is too pervasive and intractable.** This was a common response by our respondents, who felt that the whole process is 'disheartening' to quote one of the terms used.
12. There was a form of **disavowal of responsibility or a 'bystander effect'** whereby some felt someone else would deal with it; it was not their job to do anything about it because it did not concern them; or felt it was not their job to do this.

Considering our discussions with students as an entry point to 'lay' understandings of hateful and racist contents, a preliminary analysis seems to suggest that:

- Unlike focus group discussions and interviews with anti-racist activists and experts, where the emphasis was very much on the effects and roots of online hate, students **limited their analysis of racism to the discursive field** (i.e. the content itself), without reflecting on the connection between racist discourses and structural racism, and the silencing, intimidating effects that online racism can have on 'minority' writers/bloggers/social media users.
- Students tended to **view online hate speech as a product of the specific features of online interactions (anonymity, unfettered access to global audiences)**. They see online racism as 'a glitch' of the internet (Nakamura, 2013), caused by single individuals' utterances. Little attention is paid to the discursive, everyday and collective reproduction of racism and racial meanings and the real consequences this has for those targeted.
- Students tend **to view race-based offensive remarks through a post-racial lens lubricated by 'a culture of racial equivalence'** (Song, 2013): on the one hand, racism is a relic from the past; on the other hand, every sort of racially inflected abuse is considered equally damaging. Some students, rather vehemently, took issue with the notion that there are privileges that often accrue from being white; instead they argued that being a 'young, white, working class man' nowadays makes them an easy target for all sorts of abuse.
- Yet, students were also alert to the nuances of language and were able to identify when words or phrases were deployed indexically to create racial meaning or pathologise groups without direct reference to race. Students made distinctions between utterances that they found offensive and yet to be tolerated and utterances that go beyond the threshold of what is acceptable. The type of content that they found most abhorrent and unacceptable was broadly the same that was identified as intolerable by focus group participants. This is typically content that contains slurs/epithets or dehumanising comparisons and describes targeted groups as biologically inferior, sexual threats, or carriers of diseases.

Conclusions

Conclusions

The HateTrack project posed three research questions. Below, we summarise how these were addressed.

RQ1: What are the defining characteristics, the range and severity of online racist hate speech?

This question was addressed in Stage I and in Stage III of the research. Stage I relied on a series of focus groups and in-depth interviews with expert informants, while Stage III relied on a discourse analysis of online contents. Our analysis generated rich insights in how anti-racist activists and other people in the 'front line' understand and define online hate speech. Their understandings are more fluid and broader than those codified into law as 'illegal hate speech'. We have operationalised these understandings as racially-loaded toxic contents, as they contain discourses that cause harm through racialising and othering groups of people. Our results point to the variety of toxic contents, groups targeted, categories of posters, and trigger events. We found both crude 'frozen' instances and more nuanced, subtle and 'moving' forms, that may not be immediately apprehended as racially-toxic, but which nevertheless racialise and discriminate against targeted groups. We identified both organised and semi-organised groups, operating through specific pages and accounts, mobilising specific kinds of exclusionary rhetoric, as well as random, generic, 'casual', 'common sense' racially-toxic material, emanating from people who may not understand themselves as involved in any form of racism. We highlighted the specificities of the Irish varieties of racially-loaded toxic discourses, but also continuities and commonalities with US, UK and European racism. We traced the increased volumes of racially-toxic contents that are triggered by highly publicised events – for example, the Ibrahim Halawa case or the supposed 'terrorist' attacks in Dundalk – alongside the mid-range volume triggered by occasionally sensational media headlines covering Direct Provision and other refugee-related topics, migration, Traveller-related topics, or the pressures on social welfare, hospitals and housing. These were occurring on the backdrop of always-present racially-loaded toxic contents understood by Sharma (2017) as ***ambient racism*** [please see glossary and definitions].

RQ2: How can these materials be tracked on public pages on Facebook and on Twitter?

This was addressed in Stage II of the research. We employed the information obtained through Stage I to collect and classify a corpus of online contents which we then used to train the HateTrack algorithm. The tool now harvests contents from Twitter and Facebook, classifies them in terms of their probability to contain racially-loaded toxic contents and allows users to export as spreadsheet files. The tool can be very useful to both civil society actors who can use it for monitoring purposes, and to academics studying racism. It can also be extended to include more functionalities and to cover a broader range of toxic discourses, for example, against women, non-heterosexual people, or disabled people.

RQ3: What kinds of online racist incidents tend to get reported in Ireland and how do they compare to the broader racist hate materials circulating?

What are the perceived barriers to reporting and what kinds of experiences do victims and bystanders of online racist hate have to report?

This question was addressed in the fourth and final stage of the research. It should be noted once more that this was the least developed part of the research, drawing on three workshops with students. The main finding here is that there is a great reluctance to report contents. In the three workshops we ran, only three people had ever reported contents, pointing to a reluctance or reticence in dealing with racially-loaded toxic contents. Significantly, the burden is placed on those directly affected to report or deal with such contents. This echoes the work of Nakamura (2015) who thematised the extra burden faced by those targeted by online racism and misogyny, discussing their work in calling this out as a form of unpaid digital labour. The barriers to reporting that we identified here include a kind of 'first amendment absolutism', which suggests a poorly understood notion of what constitutes freedom of speech/expression in Europe; a position that such contents are better dealt with by the broader community, who will identify and appropriately shut down the 'idiots'; a view of the reporting process as pointless in the face of extremely large volumes of online racism; and a 'bystander' effect, in which responsibility is diffused because there are many others exposed to the same contents.

Overall, the findings of this research suggest continuity with what has been found by researchers studying racism in other contexts within Ireland, notably in employment (Joseph, 2017), in print media (O' Regan and Riordan, 2018), and in social interactions (Carr, 2017; Michael, 2017) and it reflects the ambivalent attitudes of Irish citizens towards race and cultural diversity (McGinnity et al, 2018). This research complements these studies by pointing to the concurrent existence of three further inter-related discourses found in social media:

i. the emergence of cross cutting-categories and specifically in the emergence of Irish people of either ethnically mixed backgrounds or descendants of migrant parents as specific targets for racism and whose Irishness is explicitly and constantly questioned and denied;
ii. the intersection of categories of race/ethnicity, gender and sexuality, with women and men positioned differently, and with sexuality featuring prominently in some racist discourses; and
iii. the cross-fertilisation of Irish-based discourses with international discourses revolving around 'white supremacy' and discourses associated with the Identitarian movement. This suggests the operation of organised or quasi-organised groups in this sphere.

Our research further noted a synergy – which we assume is unwitting – between coverage of certain events or themes by the media and a surge in online racially-toxic contents. It is important, finally, to point to the specificity of online racism which is found both in the kinds of discourses mobilised but also in the accumulations of racially-toxic discourses across platforms and over time. What we found here in terms of the substance of these discourses constitutes for the most part a repetition of racially problematic discourses circulating in society at large. In theoretical terms, we see both frozen and motile versions of racism, that suggest both a regression to old-fashioned forms of racism as biological or cultural inferiority, and more sophisticated discourses of preserving cultural integrity and uniqueness through cultural separation, but which end up victimising those deemed as culturally 'alien' or unassimilable. What is specific to social media, however, is the constant repetition and 'pile up' of all these discourses in places that are understood as public, such as the Facebook pages of online news outlets, and Twitter, which may end up silencing those targeted and forcing them to remove themselves from these spaces. Returning to Habermas' ideas of an inclusive public sphere as a requirement for a properly functioning democracy, it is clear that the circulation of racially-toxic speech may compromise this inclusivity affecting the overall state of our society.

Limitations

The definitions and categories we use throughout the report –'hateful content', 'racist discourses', 'racially-loaded toxic contents' – are sociological constructs. They are not intended to match or approximate any legal definition or threshold outlined in criminal law. There is no suggestion that any of the content analysed is unlawful or that it should be removed.

It should also be noted that Facebook and Twitter are not a representative window into Irish society so our analysis of online racist discourses can only shed limited light on the broader dynamics of racism. Social media are not used evenly by different groups; it is likely that socio-economically disadvantaged groups are the least represented on these platforms.

Overall, this research was undertaken with the aim of testing the methodological viability and reliability of using automated techniques and qualitative methods to identify and classify hateful content. As such, it is intended to be an indicative, first-take analysis of the type of racist discourses that proliferate on Facebook and Twitter and that pertain to the Irish context. The findings presented here are neither exhaustive or definitive but shared with the hope of stimulating further research on these very important topics.

In terms of the focus group discussions, we interviewed only those designated as experts in the field. Clearly, there is scope to broaden this to include a wider range of informants. Indeed, we tried to identify some 'lay' understandings of online racist speech in the discussions with students, conducted in the final part of the research. But more research needs to be conducted to examine the range of understandings of what constitutes racially-loaded toxic speech across the Irish society.

It should be noted that the HateTrack tool is not perfect: the technology used is inherently probabilistic, and the algorithm needs further refinement. Additionally, it could be made to be more user-friendly, especially working with Facebook posts, but also in terms of saving and downloading selections.

A further issue is that although we identified a set of racist discourses, targeted groups, posters, and trigger events, we were not able to perform any statistical analysis of the frequencies or the distribution metrics of these. Similarly, our analysis focused on a particular period of time and it is uncertain whether these discourses are valid over a longer period of time.

Finally, our analysis of the reporting cultures was limited to groups of students and its general validity needs to be tested in broader segments of the population. The emphasis on freedom of speech narrowly understood may be the outcome of talking to journalism students.

Future research directions

This research was an inter-disciplinary collaboration between social scientists and computer scientists. It has proven fruitful in generating new insights. Future research may extend these and look for other ways of combining the two disciplines. As a first step towards this, we made available the HateTrack tool and dataset to two MSc Digital Analytics students who will use it for their dissertations. One project will look at geotagging and locations and try to associate these with specific racially-toxic discourses; this seeks to replicate the work of the Geography of Hate project[57]. The second project will perform a network analysis of the accounts offering more information on the role of organised groups. The dataset we generated, and which can be extended through further searches using HateTrack, affords different kinds of future study. One particularly interesting one would be to use topic modeling, which refers to applying a statistical model for discovering the clustering of terms, thereby allowing the emergence of other semantic patterns from the data, that our manual analysis could not identify.

In terms of the HateTrack tool itself, this is built in a way that allows it to be extended. Three such extensions can be as follows: (i) the tool could be recalibrated to identify other kinds of hate speech, for example, misogynist, homophobic or ableist contents; (ii) it could extend to search for contents in YouTube and other social media platforms and it may also be adapted for the comments sites of news and other websites; (iii) it could be extended to operate in other languages. Future research may, therefore, help develop this tool further.

While the current project focused on defining and tracking racially-toxic contents, it would be useful to know which of these discourses tends to be more frequent or which groups are the most targeted. This would require a quantitative methodology that would look to track the rhythms of racially-toxic contents across time. In doing so, it will help clarify the operation and role of trigger events. Additionally, future research can identify the responses and counter-arguments used by others when they encounter racially-toxic speech. These 'lay' counter-narratives can be studied both for their range and their effectiveness in dealing with the various problematic discourses identified. Further, while this study derived its definitions of racial toxicity from anti-racist and community-base groups, future research may examine the extent to which these find broader resonance within the Irish society.

Examining the role of organised or quasi-organised groups was beyond the scope of this research but it has emerged as especially significant. Future research can develop relevant research questions, such as looking at the diffusion of specific vocabularies or terms, the extent to which these have infiltrated the mainstream, and their links with organised groups located elsewhere in Europe or the US.

Finally, future research can examine the 'reporting cultures' by looking more closely at the kinds of contents people tend to take issue with and report, and what they may perceive as the main barriers. Further research here can firstly expand the sample of people to include different ages and different backgrounds; and secondly, examine the diffusion of the various justifications concerning reporting across different demographic categories.

57 See here for more details: http://users.humboldt.edu/mstephens/hate/hate_map.html

Glossary and Definitions

Algorithm: a step-by-step process for solving a problem.

Ambient racism: Sharma (2017) defines ambient racism as the kind of always-present, low level, banal racism forming the backdrop of social media.

Digital public sphere: Habermas (1991) defined the public sphere as a sphere between civil society and the State, which allowed for a critical public discussion of matters of general interest to the public. Habermas considered the mass media and, in particular journalism, as the principle institutions of the public sphere and rational-critical discussion as the principle communicative form. The digital public sphere can be considered as a communicative space that is comprised of digital and social media, where participation is open to all, who can then discuss matters of common concern, using a diverse range of communicative forms, from rational discourse to memes and emojis (Schafer, 2015).

Frozen racism: Lentin (2016) describes crude forms of obvious racism as 'frozen' because they are constructed as a thing of the past that is, or ought to be overcome in the post-WWII world. Moreover, this 'frozen' racism associated with slavery, the Holocaust, and other such events of the past, is considered an evident 'real' form of racism. Lentin argues that frozen racism is used to obscure the shifting ways in which race operates.

Identitarianism: The idea that cultures should remain pure and not mixing with others, or if necessary only mix with similar others while retaining their core identity. The three core ideas of identitarianism are: (i) Ethnopluralism, i.e. the idea that cultures should be allowed to retain their uniqueness and dynamism, but separately from one another; ethnopluralism does not make explicit claims of cultural superiority. (ii) Post-ideology, the positioning beyond left and right and primarily as a cultural movement. (iii) 'Retorsion', the idea that the majority is now somehow a threatened minority in its own territory and this needs to be resisted (Ahmed and Pisoiu, 2017).

Informational libertarianism: the belief that a free market of ideas and information should operate on the internet and that this can only happen in a minimal regulatory context. This view is closely associated with some of the first and most influential online civil rights organisations, such as the Electronic Frontier Foundation (EFF) (Jordan, 2001; see also, Barbrook and Cameron, 1996).

Machine learning: Machine learning is a field of computer science that applies statistical analyses of commonalities in data. The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest (Murphy, 2012, page xxvii).

Motile racism: the converse of 'frozen' racism. Lentin (2016) follows Back and Solomos (1996) in understanding racism as a scavenger ideology, borrowing and using ideas and concepts developed elsewhere and for different purposes. Motile racism refers to the various mutations and shifts within racism, which may not necessarily be linked to racism as an ideology of biological superiority of one race over other, but which nevertheless are used to subjugate and inferiorise certain groups of people.

Neural network: This is a method of machine learning that is inspired by the biological neural networks in the human brain. Rather than performing operations sequentially, neural network techniques explore many competing hypotheses simultaneously, using parallel architectures composed of simple computational elements connected by links with variable weights (Karayiannis and Venetsopoulos, 2013: 2-3).

Primordial nationalism: Smith (1998) defines as primordial nationalism the claim that ethnies or nations exist because of 'primordial' bonds that connect their members, either through 'blood' or genetic ties (as in the socio-biological paradigm of Pierre van den Berghe) or through perceived cultural similarities of language, religion, territory and kinship (found in the culturalist approach of Clifford Geertz).

Racialisation: The ascription of 'racial', i.e. biological, immutable and unchanging characteristics to certain groups of people, including the association of phenotypical characteristics such as skin or hair colour with certain behaviours (see Barot and Bird, 2001; Goldberg, 1993). Carr and Haynes (2015: 22) understand this as "an ideological process utilised to justify or explain social stratification, inclusion or exclusion.".

Racially-loaded toxic contents: We define as racially-loaded toxic contents and kind of content that creates a division between 'them' and 'us', whereby the former is constructed as inferior; reinforce stereotypes; spread myths and disinformation; justify the exclusion, stigmatisation, and inferiorisation of particular groups; and reinforce exclusivist notions of national belonging and identity. Racially-loaded toxic language typically uses expressions and arguments that make certain words/concepts/images and the negative emotions they evoke – fear, disgust, or distrust – 'stick' to particular bodies (Ahmed, 2004).

Spreadability: The potential for digital media users to share contents online (Jenkins, Ford and Green, 2013). Specifically, Jenkins et al. use this term to encompass: the technical attributes of digital media that make it possible and easy to share contents, the economic structures that enable or restrict circulation, the attributes of the texts/contents shared, and the social networks that link people.

Trigger events: Topics or themes that tend to elicit reactions including racially-loaded toxic comments. It should be noted that often it is not the events themselves but their coverage by the media that trigger reactions and comments. This draws on the work of sociologist Joscha Legewie (2013), who found that events that construct an out-group as threatening or events that direct attention towards potential sources of inter-group conflict cause negative attitudes in response at least in the short term.

Bibliography

Ahmed, Reem and Pisoiu, Daniela (2017), The New, the Old and the Grey: Ascertaining the Discursive and Social Overlaps in the (Extreme) Right Spectrum Online, paper presented at the ECPR General Conference 2017, Oslo, 6-9 September

Ahmed, Sara (2004) *The Cultural Politics of Emotions*, Edinburgh: Edinburgh University Press.

Back, Les and Solomos, John (eds) (2000) *Theories of Race and Ethnicity*. A Reader, London: Routledge.

Balibar, Etienne and Wallerstein, Immanuel (1991) *Race, Nation, Class: Ambiguous Identities*. London: Verso.

Banks, James (2011) 'European Regulation of Cross-Border Hate Speech in Cyberspace: The Limits of Legislation', *European Journal of Crime, Criminal Law and Criminal Justice*, 19: 1-13.

Barbrook, Richard and Cameron, Andy (1996) The Californian ideology. *Science as Culture*, *6*(1), pp.44-72.

Barker, Martin (1981) *The New Racism*, London: Junction Books

Barot, Rohit, and John Bird. "Racialisation: the genealogy and critique of a concept." *Ethnic and Racial Studies* 24, no. 4 (2001): 601-618.

Bartlett, Jamie, Reffin, Jeremy, Rumball, Noelle and Williamson, Sara (2014) Anti-Social Media, available at https://www.demos.co.uk/project/anti-social-media/

Benesch, Susan (2011) 'Election-related violence: the role of dangerous speech' speech given at Centre for Human Rights and Global Justice, New York University, 25th March 2011

Benesch, Susan (2012) 'Dangerous Speech: A Proposal to Prevent Group Violence', available at http://www.worldpolicy.org/sites/default/files/Dangerous%20Speech%20Guidelines%20Benesch%20January%202012.pdf

Bonilla-Silva, Eduardo (2002) 'The linguistics of colour-blind racism: how to talk nasty about blacks without sounding "racist" ', *Critical Sociology*, 28(1–2): 41–64.

Bonilla-Silva, Eduardo (2003) *Racism without Racists: Colour-Blind Racism and Racial Inequality in Contemporary America*, New York: Rowman & Littlefield.

Brindle, Andrew (2009) A Linguistic Analysis of a White Supremacist Web Forum, PhD dissertation (unpublished). Lancaster University.

British Institute of Human Rights (2012) Mapping study on projects against hate speech online, Council of Europe, available at https://rm.coe.int/16807023b4

Butler, Judith (2017) 'Sovereign Performatives in the Contemporary Scene of Utterance', *Critical Inquiry*, 23 (2): 350-377.

Carr, James (2015) *Experiences of Islamophobia: Living with Racism in the Neoliberal Era*, London: Routledge.

Carr, James (2017) Islamophobia, Anti-Muslim Racism and Conceptions of Irish Homogeneity, in Haynes, A., Schweppe, J., and S. Taylor (eds.) *Critical Perspectives on Hate Crime: Contributions from the Island of Ireland*, pp. 253-274, Basingstoke: Palgrave.

Carr, James and Haynes, Amanda, 2015. A clash of racialisations: The policing of 'race'and of anti-Muslim racism in Ireland. *Critical Sociology*, *41*(1), pp.21-40.

Citron, Danielle K. (2009) 'Cyber Civil Rights', *Boston Law Review*, 89: 61-125.

Citron, Danielle K (2014), *Hate Crimes in Cyberspace*, Harvard University Press.

Cohen-Almagor, Raphael (2011) 'Fighting Hate and Bigotry on the Internet', *Policy and Internet,* 3 (3): 1–26.

Cole, Teju (2016) 'A Time for Refusal', The New York Times Magazine, 11 November, available at https://www.nytimes.com/2016/11/11/magazine/a-time-for-refusal.html?_r=0

Coliver, Sandra (1992) *Striking a Balance: Hate Speech, Freedom of Expression and Non-discrimination*, London: University of Essex press.

Daniels, Jessie (2008) "Race, Civil Rights, and Hate Speech in the Digital Era", in Anna Everett (eds) *Learning Race and Ethnicity: Youth and Digital Media*, The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, Cambridge, MA: The MIT Press.

Davis, Angela (2008) 'Recognising Racism in the Era of Neoliberalism', Vice Chancellor's Oration at Murdoch University, available at http://www.truth-out.org/opinion/item/16188-recognising-racism-in-the-era-of-neoliberalism

Dawkins, Richard (1976). The Selfish Gene. Oxford: Oxford University Press.

Dikaiakos, Marios D., Pallis George., and Markatos, Evangelos P. (2016) Mandola: Monitoring and detecting online hate speech, ERCIM News 2016, 107.

Essed, Philomena (1991) *Understanding Everyday Racism*, London: Sage

European Commission (2014), Report on the implementation of Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law, available at: https://publications.europa.eu/en/publication-detail/-/publication/ea5a03d1-875e-11e3-9b7d-01aa75ed71a1/language-en

Fredrickson, George M. (2002) *Racism: A Short History*, Princeton, N. J.: Princeton University Press.

Gabielkov, Maksym, Ramachandran, Arthi, Chaintreau, Augustin and Legout, Arnaud (2016) 'Social clicks: What and who gets read on Twitter?', ACM SIGMETRICS Performance Evaluation Review, 44(1), pp.179-192.

Gagliardone, Iginio, Patel, Alisha, Pohjonen, Matti (2014) Mapping and Analysing Hate Speech Online: Opportunities and Challenges for Ethiopia, available at http://pcmlp.socleg.ox.ac.uk/wp-content/uploads/2014/12/Ethiopia-hate-speech.pdf

Gagliardone, Iginio, Gal, Danit, Alves, Thiago, and Martinez, Gabriela (2015). *Countering online hate speech*. UNESCO Publishing.

Gallagher, Charles A. (2003) 'Playing the white ethnic card: using ethnic identity to deny contemporary racism', in Doane Ashley W. and Bonilla-Silva Eduardo (eds) *White Out: The Continuing Significance of Racism*, New York: Routledge, pp. 145–158.

Garner, Steve (2004) *Racism in the Irish Experience*, London: Pluto Press.

Garton Ash, Timothy (2016) *Free Speech: Ten Principles for a Connected World*, Yale: Yale University Press

George, Cherian (2016), *Hate Spin*, MIT Press.

Gilroy, Paul (2000) *Between Camps: Nations, Cultures and the Allure of Race*, London: Routledge

Goldberg, David T. (1993) *Racist Culture: Philosophy and the Politics of Meaning*, Oxford: Blackwell.

Goldberg, David T. (2008) *The Threat of Race: Reflections on Racial Neoliberalism*, Oxford: Wiley/Blackwell.

Goldberg, David T. (2015) *Are We All Postracial Yet?*, Cambridge: Polity Press.

Griffin, Dan and McMahon, Aine. 2016. ' RTÉ receives 1,300 complaints over Katie Hopkins interview', Irish Times, November 11. Available at: https://www.irishtimes.com/news/ireland/irish-news/rt%C3%A9-receives-1-300-complaints-over-katie-hopkins-interview-1.2864436, accessed 11/10/2017.

Habermas, Jurgen (1991), *The Structural Transformation of the Public Sphere*, Cambridge: Polity Press.

Hanes, Emma and Machin, Stephen (2014) 'Hate Crime in the Wake of Terror Attacks: Evidence from 7/7 and 9/11', *Journal of Contemporary Criminal Justice*, 30 (3):247-267.

Heinze, Eric (2016) *Hate Speech and Democratic Citizenship*, Oxford: Oxford University Press.

Herz, Michael and Molnar, Peter (2012) *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, Cambridge: Cambridge University Press.

Hesse, Barnor (2004) 'Im/plausible deniability: racism's conceptual double bind', *Social Identities*, 10 (1): 9-29.

Holmes, Seth M., and Heide Castañeda. (2016) 'Representing the "European refugee crisis" in Germany and beyond: Deservingness and difference, life and death.' *American Ethnologist* 43, no. 1: 12-24.

Hughey, Matthew W. (2011) 'Backstage discourse and the reproduction of white masculinities', *Sociological Quarterly* 52(1): 132–153.

Jenkins, Henry, Ford, Sam, and Green, Joshua (2013) *Spreadable media: Creating value and meaning in a networked culture*. NY: NYU Press.

Jordan, Tim (2001). Language and libertarianism: The politics of cyberculture and the culture of cyberpolitics. *The Sociological Review*, *49*(1), pp.1-17.

Joseph, Ebun. (2017) 'Whiteness and racism: Examining the racial order in Ireland'. *Irish Journal of Sociology* (2017): 0791603517737282.

Joyce, Sindy, Kennedy Margaret, and Haynes, Amanda, (2017)

Travellers and Roma in Ireland: Understanding Hate Crime Data through the Lens of Structural Inequality, in Haynes, A., Schweppe, J., and S. Taylor (eds.) *Critical Perspectives on Hate Crime: Contributions from the Island of Ireland*, pp. 325-354, Basingstoke: Palgrave.

Jubany, Olga and Roiha, Malin (2016) "Backgrounds, Experiences and Responses to Online Hate Speech: A Comparative Cross-Country Analysis", available at http://www.unicri.it/special_topics/hate_crimes/Backgrounds_Experiences_and_Responses_to_Online_Hate_Speech_A_Comparative_Cross-Country_Analysis.pdf

Karayiannis, Nikolaos and Venetsanopoulos, Anastasios, (2013) *Artificial neural networks: learning algorithms, performance evaluation, and applications* (Vol. 209). Springer Science & Business Media.

King, Ryan D. and Sutton, Gretchen M. (2014) 'High Times for Hate Crimes: Explaining the Temporal Clustering of Hate Motivated Offending', *Criminology*, 51 (4):871-894.

Krizhevsky, Alex, Sutskever, Ilya., and Hinton, Geoffrey E. (2012) 'Imagenet classification with deep convolutional neural networks', in *Advances in Neural Information Processing Systems* 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., pp. 1097–1105.

Legewie, Joscha (2013) 'Terrorist events and attitudes toward immigrants: A natural experiment', *American Journal of Sociology*, 118 (5):1199–245.

Lentin, Alana (2016) 'Racism in Public or Public Racism: doing anti-racism in post-racial times', *Ethnic and Racial Studies*, 39(1): 33-48.

Lentin, Alana (2004) *Racism and Anti-Racism in Europe*, London: Pluto Press.

Lentin, Alana and Humphry, Justine (2017) 'Anti-racism apps: framing understandings and approaches to anti-racism education and intervention', *Information, Communication, and Society*, 20:10, 1539-1553

Lentin, Alana and Lentin, Ronit (eds) (2006) *Race and State*, Newcastle, UK: Cambridge Scholars Press.

Lentin, Alana and Titley, Gavan (2011) *The Crises of Multiculturalism: Racism in a Neoliberal Age,* London: Zed Books.

Lentin, Ronit and McVeigh, Robbie (2006) *After Optimism? Ireland Racism and Globalisation*, Dublin: Metro Éireann Publications.

Linehan, Hugh (2017) 'Digital safety watchdog could prove a milestone online', *Irish Times*, 8 February.

Linehan, Thomas (2012) 'Comparing Anti-Semitism, Islamophobia, and asylophobia: the British case', *Studies in Ethnicity and Nationalism*, 12 (2): 366-386.

McGinnity, Frances, Grotti, Rafaelle, Kenny, Oona and Russell, Helen, (2017), Who experiences discrimination in Ireland?, ESRI Research Series, available at: https://www.ihrec.ie/app/uploads/2017/11/Who-experiences-discrimination-in-Ireland-Report.pdf

McGinnity, Frances, Grotti, Rafaelle, Russell, Helen, and Fahey Éamonn (2018), Attitudes to Diversity in Ireland, ESRI Research Series, available at: https://www.ihrec.ie/app/uploads/2018/03/Attitudes-to-diversity-in-Ireland.pdf

Matsuda, Mari J., Lawrence Charles R., Delgado, Richard and Crenshaw, Kirnberle. . (1993) *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, Boulder, CO: Westview Press.

Michael, Lucy, 2017. Anti-Black Racism: Afrophobia, Exclusion and Global Racisms. In *Haynes, A., Schweppe, J., and S. Taylor (eds.) Critical Perspectives on Hate Crime: Contributions from the Island of Ireland, Basingstoke: Palgrave,* pp. 275-299, Basingstoke: Palgrave.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, and Dean, Jaffrey (2013) Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, Curran Associates Inc., pp. 3111–3119.

Miles, Robert (1982) *Racism and Migrant Labour*, London: Kegan Paul.

Miles, Robert (1989) *Racism*, London: Routledge

Miller, Carl, Smith, Joshua, and Dale, Jack (2016) 'Islamophobia on Twitter: March to July 2016', Centre for Analysis of Social Media, Demos, UK, available at https://www.demos.co.uk/project/islamophobia-on-Twitter/

Mosse, George (1985) *Towards The Final Solution: A History of European Racism*, Madison: University of Wisconsin Press.

MRAP (Mouvement contre le racisme et pour l'amitié entre les peuples) (2009) Internet, enjeu de la lutte contre le racism, available at http://ancien.mrap.fr/documents-1/rapport-mrap2009.pdf

Mullaly, Una (2017) 'Why The Irish Times should not have published Nicholas Pell', *Irish Times*, 5 January.

Murphy, Kevin (2012) *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Nakamura, Lisa (2013) 'Glitch Racism: Networks as Actors within Vernacular Internet Theory', *Culture Digitally*, available at http://culturedigitally.org/2013/12/glitch-racism-networks-as-actors-within-vernacular-internet-theory/

Nakamura, Lisa (2015) The unwanted labour of social media: women of colour call out culture as venture community management. *New Formations* 86, no. 86: 106-112.

O'Mahony, J. (2011) 'Man cleared of online hatred against Travellers', *Irish Examiner* [online], 1 October, available at: http://www.irishexaminer.com/ireland/man-cleared-of-online-hatred-against-Travellers-169325.html, accessed 11 October 2017.

O'Regan, Veronica and Riordan, Elaine, (2018) Comparing the representation of refugees, asylum seekers and migrants in the Irish and UK press, Journal of Language and Politics, DOI: http://dx.doi.org/10.1075/jlp.17043.ore

Omi, Michael and Winant, Howard (1986) *Racial Formation in the United States*, London: Routledge.

Pohjonen, Matti and Udupa, Sahana (2017) 'Extreme Speech Online: An Anthropological Critique of Hate Speech Debates', *International Journal of Communication*, 11: 1173-1191.

Prentice, Sheryl, Rayson, Paul and Taylor, Paul J. (2012) "The language of Islamic extremism: Towards an automated identification of beliefs, motivations and justifications", *International Journal of Corpus Linguistics*, 17(2), 259-286.

Rainey, Bernadette, Elizabeth Wicks, and Ovey, Claire (2014). *The European convention on human rights*. Oxford: Oxford University Press.

Reisigl, Martin and Wodak, Ruth (2001) *Discourse and Discrimination: Rhetorics of Racism and Antisemitism*, London: Routledge.

Riggs, Damien W. and Augoustinos, Martha (2004) 'Projecting threat: managing subjective investments in whiteness', *Psychoanalysis, Culture and Society*, 9 (2): 219–236.

Schafer, Mike, (2015) 'Digital Public Sphere', in Mazzoleni, Gianpietro et al. (2015, Eds.): The International Encyclopedia of Political Communication, pp. 322-328. London: Wiley Blackwell.

Sharma, Sanjay (2017), Theorising Online Racism: the stream, affect & power laws, paper presented at the Association of Internet Researchers (AoIR) conference, 19-21 October, University of Tartu, Estonia.

Sharma, Sanjay and Brooker, Philip (2016) '#notracist: Exploring racism denial talk on Twitter', in Jessie Daniels, Karen Gregory and Tressie McMillan Cotton (eds) *Digital Sociologies*, Bristol: Policy Press, pp. 463-485.

Shifman, Limor (2014) Memes in Digital Culture. Cambrige (MA): MIT press.

Siapera, Eugenia, Viejo-Otero, Paloma, Moreo, Elena (2017), Hate Speech: Genealogies, Tensions and Contentions, paper presented at the Association of Internet Researchers (AoIR) conference, 19-21 October, University of Tartu, Estonia.

Smith, Anthony D (1998) *Nationalism and modernism: a critical survey of recent theories of nations and nationalism*. London: Routledge.

Song, Miri (2013) 'Challenging a culture of racial equivalence', Malmö Institute for Studies of Migration, Diversity and Welfare (MIM), Working Papers Series 13:5, available at https://www.mah.se/upload/Forskningscentrum/MIM/Publications/WPS%2013.5%20Miri%20Song.pdf. Accessed March 12, 2018.

Southern Poverty Law Centre (2016) 'Anti-Muslim sentiment dominated extremist Twitter accounts after the election', available at https://www.splcenter.org/hatewatch/2016/12/15/anti-muslim-sentiment-dominated-extremist-Twitter-accounts-after-election . Accessed March 12, 2018.

Standing, Lionel, Conezio, Jerry and Haber, Ralph. (1970) Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. In Psychonomic science 19, no. 2: 73-74.

Stanton 1, Gregory H. (2004) "Could the Rwandan genocide have been prevented?." *Journal of Genocide Research* 6.2: 211-228.

Steyn, Melisa and Foster, Don (2008) 'Repertoires for talking white: resistant whiteness in post-apartheid South Africa', *Ethnic and Racial Studies*, 31(1): 25–51.

Titley, Gavan (2016) 'The debatability of racism. Networked participative media and post-racialism', available at https://raster.fi/2016/02/17/thedebatabilityofracismnetworked-participativemediaandpostracialism/. Accessed May 5, 2017.

Titley, Gavan, (2017a) Introduction: Becoming symbolic: From *Charlie Hebdo* to 'Charlie Hebdo', in Titley, G., Freedman, D., Khiabany, G., and Mondon, A. (eds) *After Charlie Hebdo: Racism and Free Speech*, London: Zed.

Titley, Gavan (2017b) 'The Question of fake news (and hate speech)', presentation given at the 'Avoiding the hate trap in online journalism' NEAR media co-op conference, 10 April, 2017

Twomey, Aisling (2017) 'A Civil Society Perspective on Anti-Traveller and Anti-Roma Hate: Connecting Online to On the Street', in Haynes A., Schweppe J., Taylor S. (eds) *Critical Perspectives on Hate Crime*, London: Palgrave Macmillan.

Van den Berghe, Pierre L. (1978). Race and ethnicity: a socio-biological perspective. *Ethnic and racial studies*, *1*(4), pp.401-411.

Van Dijk, Teun A. (1993) *Elite Discourse and Racism*, London: Sage.

Virchow, Fabian, (2015), The 'Identitarian Movement': What Kind of Identity? Is it Really a Movement? In Simpson, Patricia and Helga Druxes (eds) *Digital Media Strategies of the Far Right in Europe and the United States*, pp. 177-190, Lanham: Lexington.

Waldron, Jeremy (2012) *The Harm in Hate Speech* ¸ Cambridge: Harvard University Press.

Walker, Samuel (1994) *Hate Speech: The History of an American Controversy*, Nebraska: University of Nebraska.

Warner, William and Hirschberg, Julia (2012) 'Detecting Hate Speech on the World Wide Web', Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012), pp. 19-26, Montreal, Canada. Association for Computational Linguistics, available at https://dl.acm.org/citation.cfm?id=2390377.

Wentraub-Reiter, Rachel (1998) 'Note: Hate Speech over the Internet: A Traditional Constitutional Analysis or a New Cyber Constitution?', *Boston Public Interest Law* Journal, 8: 145-173.

Williams, Matthew L. and Burnap, Pete (2015) 'Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data', *British Journal of Criminology*, 56, 211–238.

York, Jillian (2017) 'Google's Anti-Bullying AI Mistakes Civility for Decency. The culture of online civility is harming us all ', in *Motherboard*, August 18, available at https://motherboard.vice.com/en_us/article/qvvv3p/googles-anti-bullying-ai-mistakes-civility-for-decency, accessed 11/10/2017