D.Rad | DeRadicalisation in Europe and Beyond: Detect, Resolve, Reintegrate

Radicalisation Foresights

An analysis of Toxic tweets related to climate change, Covid-19, and immigration.

D 6.3

June 2023

Isabel Holmes – Brunel University London, UK

Marcus Nicolson, Himanshu Sharma, Amrullah Haleemi – Glasgow Caledonian University, UK

Reference: D.RAD [D6.3]

This research was conducted under the Horizon 2020 project ‘De-Radicalisation in Europe and Beyond: Detect, Resolve, Re-integrate’ (959198).

Any enquiries regarding this publication should be sent to us at: isabel.holmes@brunel.ac.uk

This document is available for download at https://dradproject.com.

Co-funded by the Horizon 2020 programme of the European Union

TABLE OF CONTENTS

About D.Rad

D.Rad is a comparative study of radicalisation and polarisation in Europe and beyond. It aims to identify the actors, networks, and broader social contexts driving radicalisation, particularly among young people in urban and peri-urban areas. D.Rad conceptualises this through the I-GAP spectrum (injustice-grievance-alienation-polarisation) so as to move towards measurable evaluations of de-radicalisation programmes. Our intention is to identify the building blocks of radicalisation, which include a sense of being victimised; a sense of being thwarted or lacking agency in established legal and political structures; and coming under the influence of "us vs them" identity formulations.

D.Rad benefits from an exceptional breadth of backgrounds. The project spans national contexts, including the UK, France, Italy, Germany, Poland, Hungary, Finland, Slovenia, Bosnia, Serbia, Kosovo, Israel, Iraq, Jordan, Turkey, Georgia, Austria, and several minority nationalisms. It bridges academic disciplines ranging from political science and cultural studies to social psychology and artificial intelligence. Dissemination methods include D.Rad labs, D.Rad hubs, policy papers, academic workshops, visual outputs and digital galleries. As such, D.Rad establishes a rigorous foundation to test practical interventions geared to prevention, inclusion and de-radicalisation.

With the possibility of capturing the trajectories of seventeen nations and several minority nations, the project will provide a unique evidence base for the comparative analysis of law and policy as nation-states adapt to new security challenges. The process of mapping these varieties and their link to national contexts will be crucial in uncovering strengths and weaknesses in existing interventions. Furthermore, D.Rad accounts for the problem that processes of radicalisation often occur in circumstances that escape the control and scrutiny of traditional national frameworks of justice. The participation of AI professionals in modelling, analysing, and devising solutions to online radicalisation will be central to the project's aims.

Executive Summary

This report presents an analysis of polarisation in online debates. It seeks to unpack the ways through which toxic tweets can stimulate polarisation. Specifically, we trace communication patterns on Twitter to examine levels of toxicity in online exchanges related to the three themes: climate change, Covid-19, and immigration. We used the google-generated Perspective API Toxicity tool to conduct an analysis of the perceived toxicity levels of tweets from both pro-

and anti- faction groups within these thematic categorisations. An initial analysis reveals that the most toxic tweets identified were related to the issue of immigration, with strong overlaps to debates surrounding COVID-19 and the vaccination. Accordingly, we proceeded with an analysis of the dataset to open a wider academic discussion of the links between toxic tweets, polarisation, and potential for radicalisation. These findings are related to a discussion of ontological security, which is a theoretical perspective that can help us begin to understand the rationale that may lead users to engage in toxic tweeting. The report concludes with a summary of these theoretical conclusions on the significance of social media for viewing polarisation and political ideologies.

Literature Review

Social media platforms allow for the sharing of ideas and content across international boundaries and connect people in unprecedented ways. However, these platforms are also used for the propagation of polarising political discourse, misinformation, and hate speech. Ozduzen et al. (2021, p.3350) state: "the digital space serves as a ventriloquist for the obscene that cannot be said in person." Online hate speech is a serious issue which requires critical analysis to create a better understanding of how polarising ideas and discourses are shared. Recent studies have examined polarization in online debates, particularly on social media platforms including Twitter and Facebook.

Ausserhofel and Maireder (2013) have emphasised that political conversations on Twitter can be conceptualised as spheres of communication. These spheres point to the emergence of the Twittersphere, a social circle which is controlled by a group of political elites. Lee and Cho (2023) have examined methods to combat online political polarisation and discovered that cross cutting exposure (exposing groups to contrasting political opinions) may be one tool to reduce polarisation and find commonalities between groups. Rathnayake and Suthers (2019) unpack the momentary connectedness that is formed between users through the sharing of issue-response hashtags, which are described as "issue-publics". Urbaniak et al. (2022) have discovered correlations between username toxicity and subsequent online toxic behaviour (personal attacks, sexual harassment among others). It was found that username toxicity was a significant predictor for toxic behaviour on the Reddit platform. Algorithms have frequently been used by researchers in social media research, with varying levels of success.

Algorithms have become trusted information tools in the public domain, but often without due critical evaluation and scrutiny prior to their inclusion in research processes. Gillespie (2014)

has called for a critical analysis of algorithms in social research and illustrates how criteria used in algorithms are often unspecified or vague, including within Twitter's own 'Trending' categorisation. Critical social researchers should therefore aspire to illuminate the workings and impact of algorithms in research. Su et al., (2017) have also critiqued computer coding methods for their inability to decipher the meanings and context of opinions expressed online. Algorithms should therefore be used in research with care, and in acceptance of these limitations.

Toxicity In Response to Terror Attacks

Bruns and Burgess (2011) emphasise that political tweets are often shared in response to specific events, which are used to build response communities. Fischer-Preßler et al. (2019) found that people used Twitter as part of sense-making process following Berlin terrorist attack, to justify their own worldviews, and to voice hostility to opposing groups. In Finland, Sumiala et al. (2023) have traced how the image of the 'bad Muslim' has been constructed discursively through print media and online communication channels in the aftermath of an attack perpetrated by an asylum seeker. Salehabadi et al. (2022) have found that toxic Twitter conversations are typically long and have a larger engagement from individual Twitter users than other threads and conversations. Therefore, toxicity is often positively related to user participation. BBC News (2022) recently published the results of a toxicity study into the number of offensive tweets sent to MPs in the UK, which has identified the proliferation of hate speech directed at politicians. Toxicity has also been found in tweets related to COVID-19.

Pascual-Ferrá et al. (2021) have found that tweets expressing anti-mask sentiment during the Covid-19 pandemic were more toxic than pro-mask tweets, and that this toxicity increased during peak holiday periods such as Thanksgiving. #WearADamnMask was the most popular pro-mask sentiment hashtag used by pro-mask users, while #NoMasks was the most used anti-mask sentiment hashtag during the same period. Among the anti-mask sentiment hashtags, #MasksDontWork had the highest mean toxicity scores in measures of toxicity, severe toxicity, identity attack, insult, profanity, and sexually explicit. Xie et al. (2023) have investigated individual characteristics that would lead to Covid-19 vaccine hesitancy and polarization through a process of 'biased assimilation'. Elsewhere, Beaunoyer at al. (2020) have drawn attention to the unequal access to the digital sphere among populations and how this has impacted upon messages shared through the Covid crisis. There are strong correlations between polarising tweets related to both Covid-19 and immigration.

Prasad (2020) emphasised how Muslims in India have been blamed for the spreading of Covid-19 as part of larger misinformation campaigns surrounding the virus. This research emphasises how ethnic minorities and migrant groups were often scapegoated for spreading the virus. In relation to immigration, Ekman (2018) has underlined how the Swedish far-right have used social media to both share racist messages, as well as build communities around these political sentiments. Murthy and Sharma (2019) found evidence of networked racialised hostility on the YouTube comment space, where users were able to create and sustain their own racist networks. Awan (2014) has argued that online islamophobia should be treated with the same gravity afforded to its street-level variant. Lingiardi et al. (2020) found that women were more likely to be victims of hate speech on Twitter. Other studies have examined polarization surrounding climate change, which is highly divisive and politicized issue.

Elgesem and Brüggemann (2022) examined how Facebook users in Norway, Germany, and the UK discuss the climate activist Greta Thunberg. The study foregrounds that German online discourse surrounding Thunberg is more polarised than in the UK and Sweden. Furthermore, the same study emphasises how polarizing discursive practices of supporters and leaders of the extreme right-wing German political party AFD have contributed to polarisation in German Facebook discourse. Schweinberger, Haugh, and Hames (2021) explored polarization in online debates surrounding Covid-19 in Australia on Twitter. The study demonstrates how emotive language can contribute to polarization in online debates. Overall, these studies shed light on the complex nature of toxicity in online debates, particularly on social media platforms, and identify the factors that contribute to polarisation. The extant literature has helped to inform the direction of the study. However, previous studies are limited in attempts to understand the ways in which toxic tweets can stimulate polarisation.

Methodology

This study adopts a mixed methods methodology that draws on both quantitative and qualitative research methods. First, we collected a sample of 8,659 tweets from three larger datasets (see Table 1 for full numbers), which were open access data sets that are readily available for researchers to utilise online. We used three existing datasets containing tweets on the topics of climate change, Covid-19 (vaccinations) and immigration (Chen, Chen & Pang, 2022; Effrosynidis et al., 2022; Rowe et al., 2021). This initial sample was collected at random, regardless of author stance on the issue (both pro, anti, and neutral positions). However, we also gathered a second sample of 9,600 tweets from the two datasets that labelled authors by

their attitudes towards Covid-19 (vaccinations) and climate change, in order to compare pro and anti-groups for each issue. Table 1 shows basic information about each dataset.

To collect tweets related to Covid-19 vaccinations, Chen and colleagues (2022) first constructed a list of users who were active in the online discourse regarding Covid-19. The authors focused on Western Europe. The authors used keywords related to Covid-19 vaccines to filter tweets sent by users. In terms of the immigration dataset, tweets were found by using the Twitter Premium API to perform keyword and hashtag searches. In this case, individual user activity in the debate was not considered. The climate change dataset was created by combining two existing climate change related datasets with related tweets found by the authors on the Internet Archive (Littman & Wrubel, 2019; Samantray & Pin, 2019).

The Covid-19 and climate change datasets both contained information on author stance. For example, "denier" of man-made climate change or "believer", pro covid-19 vaccination or anti. For the first part, our analysis examined differences between pro- and anti- groups, we selected an equal number of tweets at random from each camp, ignoring neutral tweets. We found 110 users who had a tweet labelled as both pro- and anti- vaccine in the Covid-19 dataset. As we could not be sure of the stance of these users, tweets by them were removed from the dataset. When creating the datasets to compare each theme as a whole, ignoring stance, we sampled completely at random, meaning the full range of opinions were represented.

Finally, we created word clouds of the most common 50 words used in the tweets of each dataset. In preparation for this exercise, some words which have the same meaning were simplified so that they would be recognised by the algorithm. Stopwords were also removed, in addition to the cleaning that was performed to the dataset as a whole. To prepare the tweets for analysis, we removed features that might affect model performance. For example, html artefacts and Twitter-specific references such as '#' and 'RT (Re-Tweets).

Table 1: Dataset Information

Dataset Name	**Number of Tweets**	**Sample Number (Pro and Anti dataset)**	**Date Range (Pro and Anti dataset)**	**Sample Number (Balanced dataset)**	**Date Range (Balanced dataset)**
Vaccine	17,934*	3,600	03/04/2020 - 23/09/2020	2,738	03/04/2020 - 30/09/2020

Immigration	174,267**	N/A	N/A	3,572	03/04/2020 - 25/03/2020
Climate Change	15,789,411	6,000	19/04/2019 - 25/09/2019	2,349	12/04/2019 - 18/09/2019

**Tweets labelled with stance only*

***UK tweets only*

For the climate dataset, we also only included tweets posted in 2019, which was the most recent year in the dataset. Author stance is not given in the immigration dataset, so tweets were selected wholly at random.

Covid-19 Tweets - Set Languages

Language	Number of Tweets (Pro and Anti dataset)	Number of Tweets (Balanced dataset)
French	1884	1376
German	767	617
English	764	604
Dutch	88	60
Spain	32	38
Italian	8	9
Portuguese	3	4
Turkish	6	3
Romanian	5	2
Japanese	3	0
Czech	1	0
Russian	1	0
Greek	1	0
Arabic	1	0
Danish	1	0
Polish	1	0
Thai	1	0
Misc (hashtags only, emojis, not specified by twitter)	33	25

Climate Set Languages

Language	Number of Tweets (Pro and Anti dataset)	Number of Tweets (Balanced dataset)
English	5936	2258
French	20	5
Spanish	5	5
Japanese	3	1

Portuguese	2	2
Romanian	2	0
Italian	2	2
Hindi	1	0
German	1	2
Danish	1	1
Thai	0	54
Dutch	0	2
Urdu	0	1
Arabic	0	1
Misc (hashtags only, emojis, not specified by twitter)	27	15

Immigration Set Languages

Language	Number of Tweets
English	3569
Polish	1
Portuguese	1
Misc (hashtags only, emojis, not specified by twitter)	1

Error Rates (Perspective)

Dataset	Number of Errors
Climate (Pro and Anti)	3
Climate (Balanced)	62
Immigration	6
Covid (Pro and Anti)	35
Covid (Balanced)	22

Algorithm

In the next step of the analysis, we used the toxicity tool, Perspective API, offered by Google Jigsaw to analyse tweets. Perspective provides information on the probability that a comment could be perceived as fitting one of the categories in Table 2 below. Categorisations for the threats are listed under the Type column.

Table 2: Definitions of Perspective categories (Attributes and Languages, n.d.)

Type	Definition
Toxicity	"A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion."
Severe Toxicity	"A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words."
Identity Attack	"Negative or hateful comments targeting someone because of their identity."
Insult	" Insulting, inflammatory, or negative comment towards a person or a group of people."
Profanity	"Swear words, curse words, or other obscene or profane language."
Threat	"Describes an intention to inflict pain, injury, or violence against an individual or group."

Accessible through an API, Perspective is trained on a large corpus made up of comments from sources such as Wikipedia webpages, with each comment having been annotated by between 3 and 10 annotators. The likelihood of a comment being perceived as toxic, insulting, or otherwise problematic is represented by the number of annotators who marked the comment as such. For example, if 3 out of 10 annotators marked a comment as toxic, it would receive a score of 0.3 on this metric (Training Data, n.d.). BERT based models are then trained on this text and then distilled into Convolutional Neural Networks (CNN) that allow for faster results for the end user (Model Cards, n.d.). The reported accuracies for the tool are high, however it should be noted that some researchers have found it is vulnerable to bias and error (Model Performance, n.d.). For example, Hosseini and colleagues (2017) found that making simple spelling errors in comments lowered toxicity scores significantly (e.g. from 84% to 20%). Despite these limitations, the tool is a useful, if blunt, instrument for analysing online discussion without the time and financial costs associated with training a model from a base level. In addition, it should be noted that the model architecture continues to be updated to address these issues.

Results

We carried out several t-tests to assess the differences on each class given by Perspective for deniers and believers, in the Covid-19 (pro and ant-vaccination groups) and Climate datasets. A t-test is a statistical test that compares the means of two groups and whether the differences between these are significant. We used a P-value of 0.5 is used to test whether the test was

significant or not. We found that across all categories in the Covid-19 dataset, those who were anti-vaccine posted tweets with higher levels of problematic content. See Table 3 for a full breakdown of results. In the Climate set, as shown in Table 4, man-made climate change deniers were found to post more problematic content for all categories apart from profanity, where there was no significant difference, and severe toxicity, where believers had higher levels.

Table 3: t-test results comparing anti and pro vaccine users on Perspective variables

Variable	Pro-Vaccine		Anti-Vaccine		
	M	SD	M	SD	*t*-test
Toxicity *l*	6.844	10.95	11.584	14.407	10.047***
Threats *l*	5.473	10.489	7.589	15.495	4.334***
Identity Attack *l*	4.321	9.969	5.713	11.717	3.473**
Insult *l*	5.769	10.576	9.966	14.291	9.052***
Profanity *l*	5.53	11.177	6.672	12.798	2.581*
Severe Toxicity *l*	2.523	6.923	3.88	9.569	4.405***

*$p<.05$, **$p<.01$, ***$p<.001$

Note. M = Mean. SD = Standard Deviation. *l* indicates Levenes test was significant. In this case Welch's *t*-test was used

Table 4: t-test results comparing believers and deniers on Perspective variables

Variable	Believers		Deniers		
	M	SD	M	SD	*t*-test
Toxicity *l*	13.689	22.247	20.328	20.318	11.938***
Threats *l*	1.972	4.161	2.342	5.273	2.978**
Identity Attack *l*	1.97	4.367	4.09	7.415	13.326***
Insult *l*	7.866	14.876	14.449	20.032	14.279***
Profanity	10.115	21.711	10.442	16.549	0.648
Severe Toxicity *l*	2.003	6.592	1.665	5.537	-2.131*

*$p<.05$, **$p<.01$, ***$p<.001$

Note. M = Mean. SD = Standard Deviation. *l* indicates Levenes test was significant. In this case Welch's *t*-test was used

As we did not have details on author stance categorisation for immigration, we could not compare the results given by the Perspective tool across groups in the same way. However, we were able to compare the three different datasets to each other to learn about which topic contained the most problematic tweets using a one-way ANOVA. Results are shown in Table 5 below.

Table 5: One-Way ANOVA to test between group differences on perspective variables

Variable	F	Sig
Toxicity	81.882	0.0
Identity Attack	219.571	0.0
Insult	5.154	0.006
Profanity	36.605	0.0
Threat	204.211	0.0
Severe Toxicity	42.602	0.0

The ANOVA showed significant differences between at least two groups for all variables. Therefore, we followed up with Tukey's HSD Test. Results are shown in Table 6 below. Non-significant results are shaded in grey.

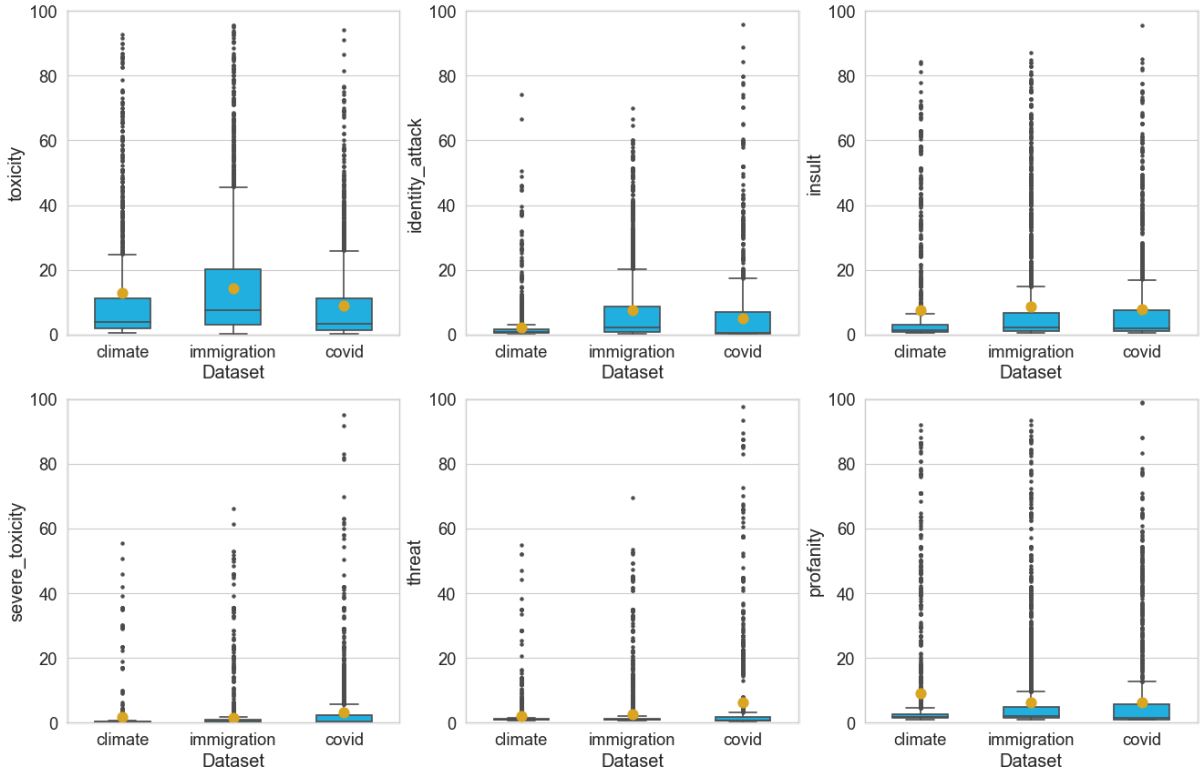
Table 6: Tukey HSD Test Results

	Comparison	**Mean Difference**	**P-adjusted**
Toxicity	Climate - Covid		0.0
	Climate - Immigration		0.002
	Covid - Immigration		0.0
Identity Attack	Climate - Covid		0.0
	Climate - Immigration		0.0
	Covid - Immigration		0.0
Insult	Climate - Covid		0.758
	Climate - Immigration		0.008
	Covid - Immigration		0.051
Profanity	Climate - Covid		0.0
	Climate - Immigration		0.0
	Covid - Immigration		0.983
Threat	Climate - Covid		0.0
	Climate - Immigration		0.02
	Covid - Immigration		0.0
Severe Toxicity	Climate - Covid		0.0
	Climate - Immigration		0.6522

	Covid - Immigration		0.0

To allow for easy interpretation of these results, we created boxplots for all variables (Figure 1). Group means are plotted in yellow. This also allows us to better understand the distribution of problematic tweets. Immigration themed tweets were frequently rated as more problematic and were found to be measure higher on levels of toxicity, identity attack, and insults. However, Covid-19 (vaccination) themed tweets scored higher on the categorisations of severe toxicity and threats. Climate themed tweets, on the other hand, were significantly higher for profanity.

Figure 1: Boxplot Results of variables



As shown in Table 7, the standard deviations for many of the variables are large. For example, in the climate topic there appears to be a lot of variation in toxicity and profanity levels. Interestingly, for the climate and immigration themes, the STD for threat and severe toxicity were lower than for Covid themed tweets.

Table 7: Means and Standard Deviations for all variables by group

Dataset	**Variable**	**Mean**	**STD**
Climate	toxicity	12.84	20.78
	identity attack	2.10	5.18
	insult	7.38	14.29

	profanity	9.13	20.00
	threat	1.91	4.28
	severe toxicity	1.70	5.99
Covid	toxicity	8.92	13.05
	identity attack	4.87	10.20
	insult	7.67	12.95
	profanity	6.17	12.17
	threat	6.10	12.38
	severe toxicity	2.98	7.79
Immigration	toxicity	14.37	16.79
	identity attack	7.60	11.72
	insult	8.51	15.01
	profanity	6.11	11.81
	threat	2.50	5.84
	severe toxicity	1.54	5.62

Word Clouds of Most Toxic Tweets (over 30)

Covid-19 Word Cloud

Immigration Word Cloud

Climate Change Word Cloud

Discussion

The overall picture presented in our analysis is of relatively low Toxicity scores across all three datasets of tweets. However, the large Standard Deviations between the datasets suggests a high level of variance between user posts. The results point to a small but significant group of polarising Twitter users who engage in highly problematic tweeting behaviour. The boxplot charts show tight groupings for low scores with long tails of more extreme results. This signals that most users in the dataset are posting non-toxic tweets. However, the large standard deviations point to a significant group of antisocial users. A closer analysis of the most toxic-tweeting user profiles revealed that these are not single-issue posters. Rather, toxic tweet posting users shared views on a variety of, often controversial, issues. In the figures below we present screenshots of two prolific users who often engaged in toxic tweeting.

Figure 3: Toxic Twitter Users Screenshots

The word clouds presented above, and twitter bios of the users (Figure 3), illustrate that these are not single-issue tweeters. Rather, these users engage in debates surrounding several controversial issues and topics. References are made to immigration, national values, political establishments, and dissatisfaction in the ruling class. One feature which recurs across most twitter user profiles we examined is a distrust of political elites and state institutions. This is reflected in the hashtag #notmyking in the user profile above, as well as alluding to so-termed Canadian values and the Canadian Prime Minister. This leads us to consider what drives these users to share such polarising views. One explanation for this behaviour can be found in ontological security theory (OST).

Ontological security (OS) is the sense of existential security that allows individuals to establish and maintain their sense of identity and self, even in the face of adversity. Ontological security was first developed by Scots psychiatrist R. D. Laing to detail how his patients lost touch with reality through an absence of social trust and routine social interactions. Giddens (1991) has later theorised that ontological security is dependent upon feelings of social trust and the predictability of everyday routines. Being able to predict the outcome of everyday social interactions and having a strong self-narrative, knowing where you are from and the story you have, are important prerequisites for ontological security. Giddens has stated that individual coping mechanisms may be used to bracket out any existential doubts that could lead to ontological (in)security, which is understood as a debilitating sense of dread. These coping mechanisms can build on the use of shared narratives and routines, even if these are not factually accurate or do not equate with personal experience (Nicolson, 2023).

In our study it appears that users engage in toxic tweeting on a broad range of political issues which may help individuals to be accepted by the online sub-group in question. An example of

this process is through posting an anti-vaccination tweet in an attempt to form part of the Anti-Vax social movement. Users assert their dissatisfaction with everyday politics, the perceived elite, and engage in the sharing of counter-narratives surrounding the three themes in question. These practices allow twitter users to illustrate their sense of belonging to the sub-group through a collective experience of alienation and distrust in the state establishment.

The experiences shared in the content of the toxic tweets predominantly relate to ontological (in)security, which would most often lead to a debilitating anxiety and the inability to cope with the pressures of everyday life. However, it appears that the users are able to use the platform to share their anxieties, distrust, and dissatisfaction in a manner that upholds their sense of identity. This correlates with the findings of recent research by Zimdars et al. (2023), and demonstrates how, paradoxically, twitter users may engage in toxic tweeting (relating to ontological (in)security) in order to establish their own sense of self, build a continuous biographical narrative, and uphold their everyday sense of belonging. Toxic tweets in this sense, may be considered as a coping mechanism, or everyday routine, for processing unexpected life events and political crisis.

Ontological security has been conceptualised as an ongoing process of becoming secure (Kinnvall, 2004), and one which is not possible to complete. When analysed from this perspective we may consider toxic tweeting as part of an ongoing process to find certainty and establish credentials for group membership. When analysed from an ontological security theoretical point of view, we can begin to understand some of the underlying rationale which may prompt users to engage in this activity. This leads us to consider why it is that the debate surrounding the Covid-19 vaccination appears to be the most polarised, in comparison with the issues of climate change and immigration. Tweets related to Covid showed the highest mean scores for the likelihood of threats and severe toxicity significantly worse than for other topics. Therefore, we turn our attention to consider the underlying factors which make debates on the vaccination issue a more toxic space on Twitter.

We speculate that the reasons for this finding could be explained by the vaccination, and Covid-19, being perceived as a larger immediate threat to life than the issues of immigration and climate change among the wider public. While climate change and immigration are seen as threats to life, these are interpreted as on-going and long-process crises. This finding may also be related to the time at which the tweets were shared and collected, at the outset of the Covid-19 pandemic in early 2020. Studies have shown that anxieties over Covid-19 decreased as the

pandemic went on and vaccine hesitancy increased in the intervening time (Phillips et al., 2022). Covid-19, and the development and implementation of the vaccine, dominated news headlines and online debates during the year of 2020. This again goes to explain the proliferation of these tweets, but not necessarily the underlying reasons for toxicity within these. Furthermore, users tended to engage in tweeting about multiple controversial issues, and shared conspiracy theories and beliefs that were not limited to only one of the themes. This may also be related to a need for cognitive closure (Webster and Kruglanski, 1994) through the expressed beliefs shared online. This can go towards explaining intense discomfort with any challenge to beliefs manifesting as toxicity and rage in online communications. Thus, the sharing of toxic tweets can be thought to provide some degree of cognitive closure to individual users, as well as upholding a sense of ontological security which allows for the maintenance of identities and group membership.

Concluding Remarks

This study has analysed a dataset of tweets on the three issues of Covid-19, climate change, and immigration in order to understand polarisation within online debates and Twitter users. We have employed the use of an algorithm, Perspective API, to analyse the datasets and gauge levels of toxic content within these. While not without its limitations, we believe the algorithm holds value when coupled with qualitative analysis in order to critically assess the content of toxic tweets.

A theoretical perspective of ontological security can provide some answers for understanding the user motivations behind the sharing of toxic and polarising tweets. This is because users are able to share their own feelings of anxiety and distrust of state mechanisms in order to establish their own feelings of security. While this is not the focus of the present study, we should also like to mention that online platforms can foster unity between users and allow for the emergence of counter-narratives that can also play a positive role in democratic processes.

The main findings of the report are summarised as follows:

- Social media and Twitter are prominent sites for the sharing of political ideologies and sites of polarisation.
- Most tweets analysed found low Toxicity scores among users.
- There were high standard deviation results in Toxicity levels, which points to a small but significant number of antisocial users.

- An analysis of word clouds and user profiles illustrates that these are not single-issue posters, but rather that users tweet about multiple controversial issues.
- Covid-19 vaccination related tweets were found to have higher Toxicity levels than immigration and climate change themed tweets.
- Users shared feelings of anxiety and distrust in political elites and state institutions.
- These tweets can be thought to foster ontological security and credentials for group membership.
- These findings do not detract from the dangers of social media and the potential for the sharing of misinformation and potentially radicalising political ideologies.

We authors propose to develop this report into a future research journal publication, building on a further mixed-methods analysis of the datasets. This continuation of the present report will include a larger sample of tweets and include a lengthier qualitative analysis of Toxic tweets. This report marks a strong starting point from which to build on this interdisciplinary research.

References

Ausserhofer, J., & Maireder, A. (2013). National politics on Twitter: Structures and topics of a networked public sphere. *Information, communication & society*, *16*(3), 291-314.

Awan, I. (2014). Islamophobia and Twitter: A typology of online hate against Muslims on social media. *Policy & Internet*, *6*(2), 133-150.

BBC News (2022) Scale of abuse of politicians on Twitter revealed. https://www.bbc.co.uk/news/uk-63330885

Beaunoyer, E., Dupéré, S., & Guitton, M. J. (2020). COVID-19 and digital inequalities: Reciprocal impacts and mitigation strategies. *Computers in human behavior*, *111*, 106424.

Bruns, A., & Burgess, J. (2011). The use of Twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European consortium for political research (ECPR) general conference 2011* (pp. 1-9). The European Consortium for Political Research (ECPR).

Chen, N., Chen, X., & Pang, J. (2022). A multilingual dataset of COVID-19 vaccination attitudes on Twitter. *Data in Brief*, *44*, 108503.

Effrosynidis, D., Karasakalidis, A. I., Sylaios, G., & Arampatzis, A. (2022). The climate change Twitter dataset. *Expert Systems with Applications*, 117541.

Ekman M (2018) Anti-refugee mobilization in social media: the case of Soldiers of Odin. *Social Media+ Society* 4(1): 1–11.

Elgesem, D. and Brüggemann, M. (2022) "Polarisation or just differences in opinion: How and why Facebook users disagree about Greta Thunberg," *European Journal of Communication*, p. 026732312211161. Available at: https://doi.org/10.1177/02673231221116179.

Fischer-Preßler, D., Schwemmer, C., & Fischbach, K. (2019). Collective sense-making in times of crisis: Connecting terror management theory with Twitter user reactions to the Berlin terrorist attack. *Computers in Human Behavior, 100*(5), 138–151. doi: 10.1016/j.chb.2019.05.012

Giddens, A. (1991). *Modernity and self-identity: Self and society in the late modern age*. Stanford university press.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), *Mediatechnologies: Essays on communication, materiality, and society* (pp. 167–194). Cambridge: MIT Press.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138.*

Kinnvall, C. (2004). Globalization and religious nationalism: Self, identity, and the search for ontological security. *Political psychology*, *25*(5), 741-767.

Lee, S., & Cho, J. (2023). Hearing and speaking the other side: The roles of expression and opinion climate perception in political polarization. *Computers in Human Behavior*, *143*, 107672.

Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2020). Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, *39*(7), 711-721.

Littman, J., & Wrubel, L. (2019). Climate change tweets Ids. Harvard Dataverse.

Murthy D, Sharma S (2019) Visualizing YouTube's comment space: online hostility as a networked phenomena. *New Media & Society* 21(1): 191–213.

Nicolson, M. (2023). Racial microaggressions and ontological security: exploring the narratives of young adult migrants in Glasgow, UK. *Social Inclusion*, *11*(2), 37-47.

Ozduzen, O., Korkut, U., & Ozduzen, C. (2021). 'Refugees are not welcome': Digital racism, online place-making and the evolving categorization of Syrians in Turkey. *New media & society*, *23*(11), 3349-3369.

Pascual-Ferrá, P. et al. (2021). Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic. *Big Data & Society,* Available at: https://doi.org/10.1177/20539517211023533.

Perspective Developers. (2023). *Attributes and Languages* (n.d.). https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

Perspective Developers. (2023). *Performance Overview* (n.d.). https://developers.perspectiveapi.com/s/about-the-api-model-cards?language=en_US&tabset-20254=3

Perspective Developers. (2023). *Model Cards* (n.d.). https://developers.perspectiveapi.com/s/about-the-api-model-cards?language=en_US

Perspective Developers. (2023). *Training Data* (n.d.). https://developers.perspectiveapi.com/s/about-the-api-training-data?language=en_US

Phillips, R., Gillespie, D., Hallingberg, B., Evans, J., Taiyari, K., Torrens-Burton, A., ... & Wood, F. (2022). Perceived threat of COVID-19, attitudes towards vaccination, and vaccine hesitancy: A prospective longitudinal study in the UK. *British Journal of Health Psychology*, *27*(4), 1354-1381.

Prasad, A. (2020). The organization of ideological discourse in times of unexpected crisis: Explaining how COVID-19 is exploited by populist leaders. *Leadership*, *16*(3), 294-302.

Rambukkana, N. (Ed.). (2015). Hashtag publics: The power and politics of discursive networks. New York, NY: Peter Lang

Rathnayake, C., & Suthers, D. D. (2019). 'Enclaves of exposure': A conceptual viewpoint to explore cross-ideology exposure on social network sites. *The Social Science Journal*, *56*(2), 145-155.

Rowe, F., Mahony, M., Graells-Garrido, E., Rango, M., & Sievers, N. (2021). Using Twitter to track immigration sentiment during early stages of the COVID-19 pandemic. *Data & Policy*, *3*.

Salehabadi, N., Groggel, A., Singhal, M., Roy, S. S., & Nilizadeh, S. (2022). User Engagement and the Toxicity of Tweets. *arXiv preprint arXiv:2211.03856*.

Samantray, A., & Pin, P. (2019). Credibility of climate change denial in social media. Palgrave Communications, 5(1), 1–8.

Schweinberger, M., Haugh, M. and Hames, S.C. (2021) "Analysing discourse around COVID-19 in the Australian Twittersphere: A real-time corpus-based analysis," *Big Data & Society*, 8(1), p. 205395172110214. Available at: https://doi.org/10.1177/20539517211021437.

Sumiala, J., Harju, A. A., & Palonen, E. (2023). Global Populism: Its Roots in Media and Religion| Islam as the Folk Devil: Hashtag Publics and the Fabrication of Civilizationism in a Post-Terror Populist Moment. *International Journal of Communication*, *17*, 19.

Su, L. Y. F., Cacciatore, M. A., Liang, X., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2017). Analyzing public sentiments online: Combining human-and computer-based content analysis. *Information, Communication & Society*, *20*(3), 406-427.

Urbaniak, R., Tempska, P., Dowgiałło, M., Ptaszyński, M., Fortuna, M., Marcińczuk, M., & Wroczyński, M. (2022). Namespotting: Username toxicity and actual toxic behavior on Reddit. *Computers in Human Behavior*, *136*, 107371.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of personality and social psychology*, *67*(6), 1049.

Xie, L., Wang, D., & Ma, F. (2023). Analysis of individual characteristics influencing user polarization in COVID-19 vaccine hesitancy. *Computers in Human Behavior*, 107649.

Zimdars, M., Cullinan, M. E., & Na, K. (2023). Alternative health groups on social media, misinformation, and the (de) stabilization of ontological security. *New Media & Society*, 14614448221146171.