Trajectories of Islamophobic hate amongst far right actors on Twitter

Bertie Vidgen*[,a,b], Taha Yasseri[a,b], Helen Margetts[a,b]

[a]The Oxford Internet Institute, University of Oxford, Oxford OX1 3JS

[b]The Alan Turing Institute, London NW1 2DB

[*]Corresponding author. Email: bvidgen@turing.ac.uk

Abstract

Far right actors use the Internet for myriad purposes, such as forming communities, sharing information and attracting support. Concerns have been raised about their use of social media to spread hateful messages by both academics and policymakers. Given the potentially dangerous effects of hate speech, which can inflict harm on targeted victims, create a sense of fear amongst communities and pollute civic discourse, there is a pressing need to understand at a granular level how it manifests amongst far right actors online. In this paper we investigate the dynamics of Islamophobia amongst followers of a far right political party on Twitter, the British National Party (the BNP). Using a new dataset of 5.2 million tweets, collected over a period of one year, we identify 7 distinct trajectories of Islamophobia, which capture heterogeneity in users' behaviour. The 7 trajectories reflect qualitative, quantitative and temporal differences in users' behaviour. We analyse the data using a classifier for Islamophobic content (which distinguishes between None, Implicit and Explicit Islamophobia), and latent Markov modelling with k-modes clustering. The findings provide a new level of granular insight into Islamophobic behaviour amongst the far right on social media. They both deepen existing knowledge and inform policy discussions regarding how far right extremism and hate speech can be analysed and tackled. We make our dataset of 5.2 million tweets publicly available.[1]

Keywords: Far right, Islamophobia, hate speech, social media, latent markov modelling, Twitter, the BNP.

Introduction

Far right actors were early adopters of digital technologies in the early 2000s [1]–[3] and have since used the Internet for myriad purposes, including forming communities [4]–[6], building international alliances [7]–[9], sharing information and broadcasting messages [1], [10], and organising and attracting support [2], [10]–[12]. In particular, social media has been widely

[1] Data will be released on publication of the final article.

adopted by the far right actors as it offers them the opportunity to bypass traditional media 'gatekeepers' and access very large audiences [13], [14]. Despite this growing body of research into far right politics online, relatively little attention has been paid to how digital far right actors produce and share hateful content.

The BNP (founded in 1982) is a far right party in the UK, founded in the early 1980s. Its electoral successes have challenged the widely held view that the UK is a case of 'far right failure' [15]. During the 2000s several BNP councillors and a London assembly member were elected, as well as two Members of the European Parliament in 2009. However, since the late 2000s, the party has suffered setbacks as the far right landscape has diversified and the party has faced internal troubles and legal proceedings [13], [16]. Its longstanding leader, and one-time MEP, Nick Griffin received considerable media exposure during the early 2010s in mainstream TV programs and national newspapers [17], [18], but was replaced in 2015 by the little-known Adam Walker. At the 2010 general election the BNP received 563,743 votes or 1.9% of the total (although no BNP Members of Parliament were elected due to Britain's first past the post system); in 2015, it received just 1,667 votes. However, determining the true level of support for the BNP is difficult given that its constituency of potential or 'latent' supporters may be far greater than its number of actual voters and party members [19]. The BNP remains an important focus of far right research it has much in common with other far right groups, many of whom also are unlikely to achieve electoral success in the near future but have considerable impact within the far right landscape and rely heavily on social media. This includes many of the newer, less public and less organisationally stable groups, such as Generation Identity, the 'Casuals' and National Action.

Xenophobic nationalism, or 'nativism', is widely viewed as a constitutive feature of the contemporary far right, alongside populism and authoritarianism [20]–[22]. Many minority groups have received prejudicial abuse from far right parties; the BNP has been variously described as racist [23], anti-Semitic [24], homophobic [25], anti-Immigrant [26] and sexist [27]. Since the the 9/11 Islamist terrorist attacks in America, many far right groups have directed their prejudice against Muslims [28]–[30]. As Zúquete puts it, 'the threat that the Crescent will rise over the continent and the spectre of a Muslim Europe have become basic ideological features and themes of the European extreme right' [31]. This is evidenced by far right narratives online. Awan describes how the English Defence League 'us[es] social networking sites like Twitter to post malicious statements […] promoting online hate' [32, p. 145] and elsewhere, with Zempi, writes that the far right 'exploit the virtual environment and world-wide events to incite hatred towards Islam and Muslims' [33]. Several studies also show how far right groups use online spaces to create deeply affective anti-Muslim discursive frames [34]–[36]. However, whilst there is consensus that Islamophobia is a core part of the contemporary far right's ideology and discourse, existing research provides competing accounts of how such behaviour manifests online.

Much of the traditional 'offline' literature suggests that Islamophobia will manifest fairly evenly online as, in many studies, the far right is treated as a homogenous block of like-minded prejudiced individuals. For instance, Trilling describes supporters of the BNP as 'bloody nasty people' [37] and Biggs and Knauss use membership of the BNP as a proxy for holding prejudicial beliefs [38]. Similarly, in a review of research into the far right, Rydgren describes the 'ethnic competition thesis' as a key explanation of voting for far right parties because 'even if not all voters who hold anti-immigration attitudes vote for a new radical right-wing party, most voters who do vote for such parties hold such attitudes' [39, p. 250]. This position is supported by Golder in a subsequent review, who discusses how ethnic competition drives far

right support through economic and cultural grievances [22, pp. 483–485]. These arguments have some support within online-specific research. In a measurement study, Chandrasekharan et al. suggest that all of the content posted by members of certain banned subreddits, including r/fatpeoplehate and r/CoonTown, is toxic [40]. Awan also describes how far right actors use social media 'to inflame religious and racial tensions' by creating 'walls of hate' [41]. These accounts suggest that most far right actors are deeply and vocally prejudiced individuals.

At the same time, other studies suggest that the far right is far more heterogeneous. Research into far right voters suggests they can be motivated by economic deprivation and a desire to 'protest' against mainstream parties [42], [43]. Qualitative investigations of far right supporters also suggest that many have myriad motivations for joining far right parties, and sometimes express ambiguous views on immigration and ethnic outgroups [44]. Reflecting on the fluctuating and dispersed nature of the online far right, Ganesh argues that it should be conceptualized as a 'swarm', comprising members with constantly shifting allegiances, priorities and levels of commitment [8]. He proposes that the far right is not a single fixed entity but a complex set of loosely affiliated individuals who are attracted to far right organisations for a range of reasons rather than a single goal of spreading hate.

Research into the dynamics of online hate more broadly suggest that how it manifests varies across users, time, context and geography [45]. For instance, in a study of Islamophobic tweets sent during 2016/2017, researchers at DEMOS found that of all the hateful tweets they collected ~15% were sent by 1% of the users and 50% were sent by just 6% [46]. Alrababa et al. investigate the role of celebrities in reducing prejudice amongst followers of football clubs on Twitter, and find that the prevalence of Islamophobic tweeting can drop substantially after individuals are exposed to positive celebrity role models [47]. Burnap and Williams also show that online hate follows temporal dynamics, exhibiting peaks and troughs around contentious events, such as terrorist attacks [48], [49].

At present, there is a lack of evidence regarding how Islamophobic behaviour manifests amongst far right actors on social media. Partly, this is because much previous work has focused on the prejudicial attitudes amongst far right actors rather than prejudicial behaviours. Behaviours and attitudes are not necessarily concomitant and may have a complex relationship with each other. There is a pressing need for research which refines our understanding of far right *behaviour* on social media, and which can be used to inform appropriate policy responses. To address this research gap, we investigate the dynamics of Islamophobia amongst followers of the BNP on Twitter over a period of one year, using a newly collected dataset of 5.2 million tweets.

Data and Methods

Data

All tweets sent by followers of the BNP's Twitter account (@bnp) were collected from 1st April 2017 to 1st April 2018. This period covers several important political events in the UK, including the General Election on 8th June 2017, Local Elections on 4th May 2017, Manchester Arena bombing on 22nd May 2017, London Bridge terror attack on 3rd June 2017 and the progression of the European Union (Withdrawal) Act of 2018 through the UK parliament. Tweets are collected using Twitter's Search API, which allows a maximum of 3,200 tweets to be collected from each user's timeline (including retweets). We collect data on a weekly basis and, as such, only miss tweets from users who exceed this high weekly limit.

At the start of the period (1st April 2017) there were 13,002 followers of the BNP and at the end (31st March 2018) there were 13,951. Of the original 13,002 users, 11,785 (90.6%) were still followers at the end (1,217 ceased following). Given that it is easy to start and stop following accounts on social media, often indicating a lack of genuine interest, we only include users in the dataset who follow the BNP across the entire period. 5,310 of these 11,785 users (45%) tweet at least once during the period. We remove tweets which are sent in languages other than 'English' or 'Undetermined'. This reduces the number of users by 68 to 5,242.

We remove bots, which are defined as accounts who send more than 40 tweets per day on average. This is based on the work of Kollanyi et al., who consider accounts which post at least 50 times per day to be highly automated [50]. Kollanyi et al.'s cutoff of 50 is arbitrary, and some bot-owners have responded to limits by setting their bots to tweet just below the limits.[2] As such, we opt for a lower threshold of 40 tweets per day per user (14,600 in total during the period studied). This approach can be understood as a way of removing high-activity users, including both bots and genuine users with idiosyncratic or semi-automated tweeting patterns [51]. 114 users meet this bot-detection criterion and are removed.[3] This reduces the number of users to 5,128. The datasets consists of 5,221,256 tweets.

Measurement of Islamophobia

Islamophobia is a deeply contested term, with many competing definitions available [52]–[54]. Bleich's definition of Islamophobia has been widely adopted: 'indiscriminate negative attitudes or emotions directed at Islam or Muslims' [55]. It is well-suited to the study of social media content and is easily operationalized for empirical research. We use a machine learning classifier to detect Islamophobia, which assigns tweets to one of three classes on an ordinal scale: None, Implicit and Explicit Islamophobia. The Explicit/Implicit distinction is widely used in other research detecting abusive content online [56] and enables granular insight into the different ways in which Muslims, and Islam, are associated with negative traits [57]. The classifier is described further in [58] and more information is available in the Methods appendix.

Latent Markov modelling

Latent Markov (LM) modelling is an extension to the traditional Markov chain model. It assumes the existence of K latent states, where K must be defined in advance [59]. The LM model then estimates both the behaviours associated with each latent state, the transitional

[2]This point was made to one of the authors by Sam Woolley, one of the authors of the Kollanyi et al. paper, in a private conversation.

[3]We repeat all of our analyses on the full dataset, without any users removed for high volume tweeting, and report similar results.

probabilities between states and the state which each user is assigned to in each time period. Parameters are estimated using maximum likelihood estimation via the expectation-maximization algorithm [60]. We fit our model with time homogeneous transitional probabilities (i.e. the transition probabilities are constant over all time periods).

Studying users' behaviour on Twitter longitudinally is difficult because users tweet at different times. As such, the actual timestamps of tweets cannot be used as this would create a LM model with millions of different 'events', few of which line up with each other. One solution is to measure tweets within a pre-defined time window, such as 1 day. However, this risks introducing considerable biases because users send different volumes of tweets over time (i.e. on some days the volume of tweets is high and on others it is low). As such, we scale the time period by the total number of tweets (5,221,256). This is divided into 100 periods, each of which consists of 52,213 tweets. The amount of linear time that each time period covers range from 1.7 days to 8.7 days. This approach is counter-intuitive but ensures that (i) for each user, the number of time periods without a value is minimized and (ii) users are compared across the same time intervals; t_x covers the same time period for every user – it is just that the linear length of t_x is not the same as the linear length of t_{x+1}. Given that the choice of 100 periods is arbitrary, we run our models with 10, 25 and 50 time periods and report similar results. Details of model fitting for the LM model are given in the Methods appendix.

For the LM model, we measure Islamophobia for each user in each time period by taking just the highest class of tweeting they exhibit. This utilizes the three levels of the ordinal Islamophobia variable, assigned by the classifier: None, Implicit and Explicit Islamophobia. For instance, if a user sends at least one tweet that is Explicit Islamophobic during t_x then that is how their behaviour is characterised in t_x. If they send at least one Implicit Islamophobic tweet but no Explicit tweets then their behaviour is characterised as Implicit. It is only characterised as None if they send no Implicit or Explicit tweets. This strategy ensures that Islamophobic tweets are well represented in the LM model. It is also theoretically robust since what is of greatest interest is whether users have engaged in Islamophobic behaviour rather than whether, for instance, the *majority* of their behaviour is Islamophobic. Finally, even though we use a varying time period scaled by the overall volume of tweets, some users do not send any tweets in some time periods. Rather than treating time periods when users do not tweet as missing, we assign them a value of None Islamophobic.

Results

Prevalence of Islamophobia

Of the 5.2 million tweets in the dataset, 4.4 million are non-Islamophobic (83.8%), 0.57 million are Implicitly Islamophobic (10.8%) and 0.28 million are Explicitly Islamophobic (5.3%). Surprisingly, twice as much of the Islamophobia expressed by followers of the BNP is Implicit (i.e. subtle and nuanced) rather than Explicit (i.e. aggressive and overt). This is shown in Figure 1(a). The prevalence of Islamophobia fluctuates considerably during the period studied. There are several peaks, most noticeably at the start when several terror attacks took place. Previous research indicates these are likely to drive spikes in online hate, and may explain some of the variation observed here [48]. The prevalence of Islamophobia over time across the whole cohort of users is shown in Figure 1 (b).

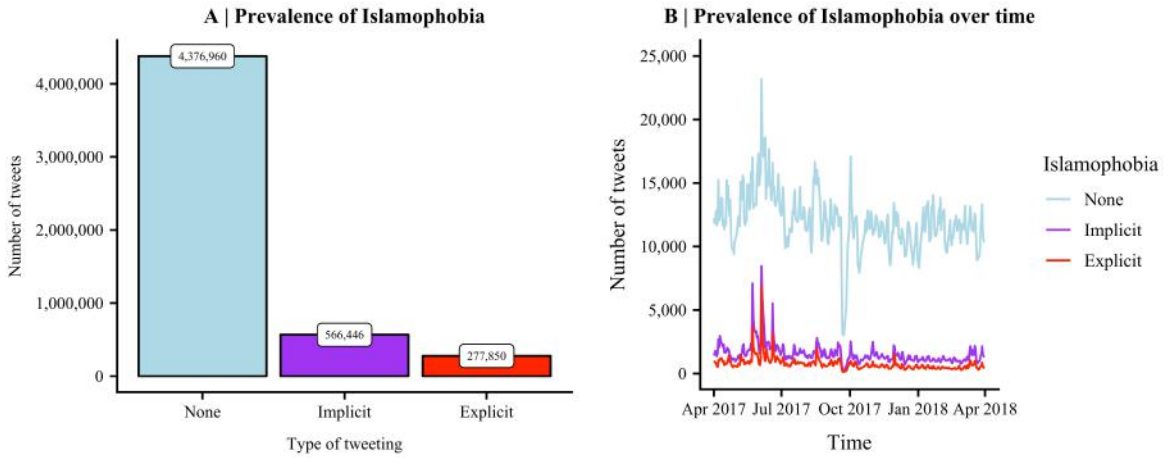

Figure 1, (a) Prevalence of Islamophobic tweets within 5.2 million tweet dataset, (b) Prevalence of Islamophobic tweets over time for all users in cohort.

The distribution of tweets per user is long-tailed, as shown in Figure 2(a). The maximum number of tweets is curtailed at 14,600 because during the sampling process high volume tweeters are removed. The distribution of Islamophobic tweets per user (combining Implicit and Explicit) is shown in Figure 2(b) and is also long-tailed. The Gini coefficients for the distribution of Implicit Islamophobia among users is 0.812, Explicit Islamophobia is 0.803, and both combined is 0.822. The Gini coefficient for the distribution of all tweets is very similar, 0.806. Overall, a small number of users are responsible for most of the Islamophobic tweets in the dataset. Figure 2(c) shows the number of Islamophobic tweets versus the total number of tweets sent by each follower.

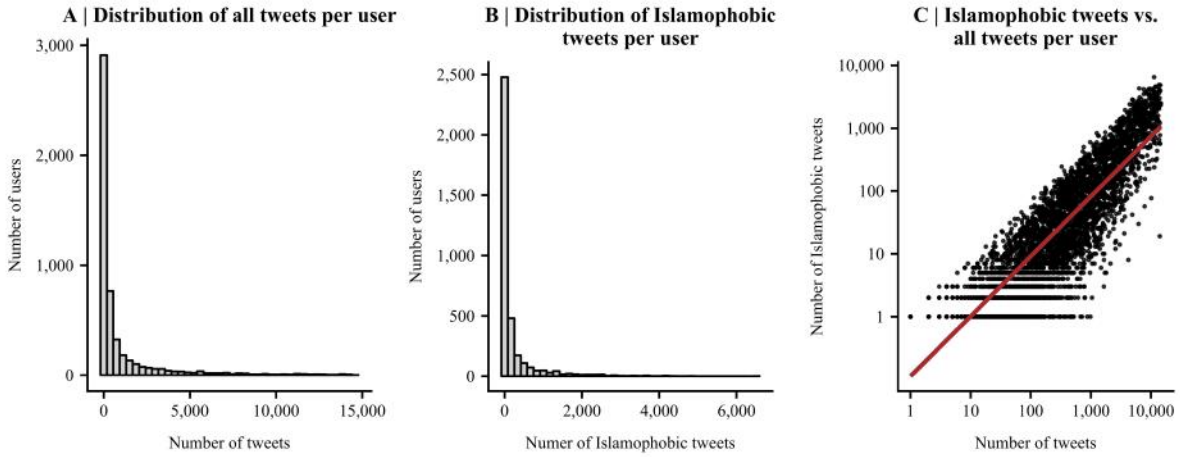

Figure 2, (a) Distribution of all tweets for all users in cohort, (b) Distribution of Islamophobic tweets (implicit and explicit combined) for all users in cohort, (c) Scatter plot of the number of Islamophobic tweets against the total number of tweets per user. In (c) axes are logarithmic and users who do not send any Islamophobic tweets (n = 1,484) are not shown.

These analyses suggest that the overall prevalence of Islamophobia reported in Figure 1(a) offers only a very coarse characterisation of users' behaviour. Figure 1(b) shows that there are considerable variations over time and Figure 2 that there are variations in terms of how much Islamophobia each user sends, both as an absolute value and as a proportion of the total number of tweets sent (see Figure 2(c)). Taken together, these findings suggest that patterns of

Islamophobic behaviour are highly heterogeneous, and that time is likely to be an important aspect for understanding how users differ. For instance, some users may follow the overall temporal trend shown in 2(b) whilst others will diverge from it. These variations are investigated further in the next section.

Trajectories of Islamophobic hate

To investigate variations in users' behaviour over time we identify and investigate distinctive trajectories in the data. A trajectory can be understood as a typified pattern of behaviour followed by a subset of users in the cohort over time. It is akin to a pathway, as has been widely examined in studies of terrorism [61], and a customer journey, as studied in business and management studies [62]. To model the existence of trajectories, we fit an LM model with $K = 3$ over the 100 time periods, each of which comprises a fixed interval of 52,213 total tweets sent (see Data and Methods). Based on initial exploratory analyses, we separate two groups of users before fitting this model: those who send only None Islamophobic tweets (n = 1,484) and those who send only None and Implicit Islamophobic tweets (n = 718). The LM model is fit on the remaining users (n = 2,926), all of whom send tweets across the three classes of None, Implicit and Explicit Islamophobia .

The three latent states in the LM model reflect different propensities to engage in each type of tweeting. State 1 has a 0.95 probability of None Islamophobic and low probabilities for both Implicit and Explicit Islamophobia (0.03 and 0.02): when users are in this latent state they are overwhelmingly likely to not engage in any Islamophobia. State 2 is the most evenly distributed across the three types, with probabilities which range from 0.22 to 0.45: users in this latent state will exhibit highly varied behaviour. State 3 has a 0.81 probability for Explicit Islamophobia, 0.13 for Implicit and just 0.06 for None: users in this latent state are highly likely to send an Islamophobic tweet, particularly an Explicit one. These probabilities are shown in

Table 1.

Table 1, Behavioural probabilities for latent states

Islamophobia	State 1	State 2	State 3
None	0.95	0.45	0.06
Implicit	0.03	0.32	0.13
Explicit	0.02	0.22	0.81
TOTAL	1	1	1

Each state has a transitional probability, which captures how likely users are to either stay in the same state or move states. These are shown in Table 2. The probabilities are all very high for staying in the same state (0.95 to 0.99). However, interestingly, users in State 3 are more likely to shift back to State 1 and 2 than users are to switch into State 3 (there is a probability of 0.95 of staying in State 3 whilst there is a probability of 0.99 of staying in State 1). Given that State 3 is most strongly associated with Explicit Islamophobic behaviour, and State 1 with None Islamophobic behaviour, this suggests that users are more likely to engage in Explicit Islamophobic behaviour and then return to None Islamophobic than the opposite way round:

Explicit Islamophobic tweeting is a less stable behaviour compared with None Islamophobic tweeting.

Table 2, Transitional probabilities for latent states

		State at t_{i+1}		
		1	2	3
State at t_i	1	0.99	0.01	0.00
	2	0.03	0.96	0.01
	3	0.02	0.03	0.95

The LM model provides a simplified representation of the underlying data, in which each user is represented as a vector of length 100 (in line with the 100 time periods), each value of which is a latent state. We cluster these vectors using the k-modes clustering algorithm. Through fitting to minimize within sum of squares, and manual inspection, we identify that five clusters are optimal (see the Methods appendix). Each of the five clusters represents a distinct trajectory; users are assigned to just one and cannot move between them. In addition, there are 2 further trajectories, which comprise the two groups separated at the start: (1) users who send only None Islamophobic tweets (n = 1,484) and users who send only None and Implicit Islamophobic tweets (n = 718). As such, we identify a total of 7 trajectories within the entire cohort of users.

To analyse the differences between trajectories, and to account for the underlying trend in the data shown in Figure 1(b), we calculate a metric called Trajectory Score (*TScore*). For each cohort, in each time period, we take the average number of tweets for each type of tweeting (None, Implicit and Explicit Islamophobia) and divide by the average number of tweets across the whole cohort. This is shown in Equation 1, where p is the time period, l is the type of tweeting, U is the number of users, u is each user, T is the trajectory, and n is the number of tweets.

Equation 1

$$\mu_{pl} = \frac{\sum_{u=1}^{U} n_{upl}}{U}$$

$$T_{pl} = \frac{\sum_{u=1}^{U_T} n_{upl}}{U_T}$$

$$TScore_{pl} = \left(\frac{T_{pl}}{\mu_{pl}}\right) \times 100$$

TScore can be interpreted as a coefficient where 100 indicates the trajectory is in line with the average of the whole cohort and any other value is a multiple of the cohort average. For instance, a value of 25 for Implicit Islamophobia indicates that users in this trajectory send 25% of the average amount of Implicitly Islamophobic tweets across the whole cohort. A value of 200 indicates that 200% of the average has been sent. There is no seasonality to account for

with this metric because we split the data into 100 periods based on the number of tweets rather than linear time (see **Error! Reference source not found.**).

The behavioural patterns of the 7 trajectories are shown in Figure 3. In each panel, the average prevalence of tweeting across the whole cohort (for each time period and within each type of tweeting) is depicted by the horizontal grey dashed line, which is at a constant of 100. We name the seven trajectories: None, Very Low, Low, High, Very High, Escalating and De-escalating. The seven trajectories capture differences in the volume of tweets which users send, the strength of those tweets, and their temporality, and are described in Table 3. Noticeably, we do not identify a Moderate trajectory as none of the trajectories has a TScore which is consistently close to 100.

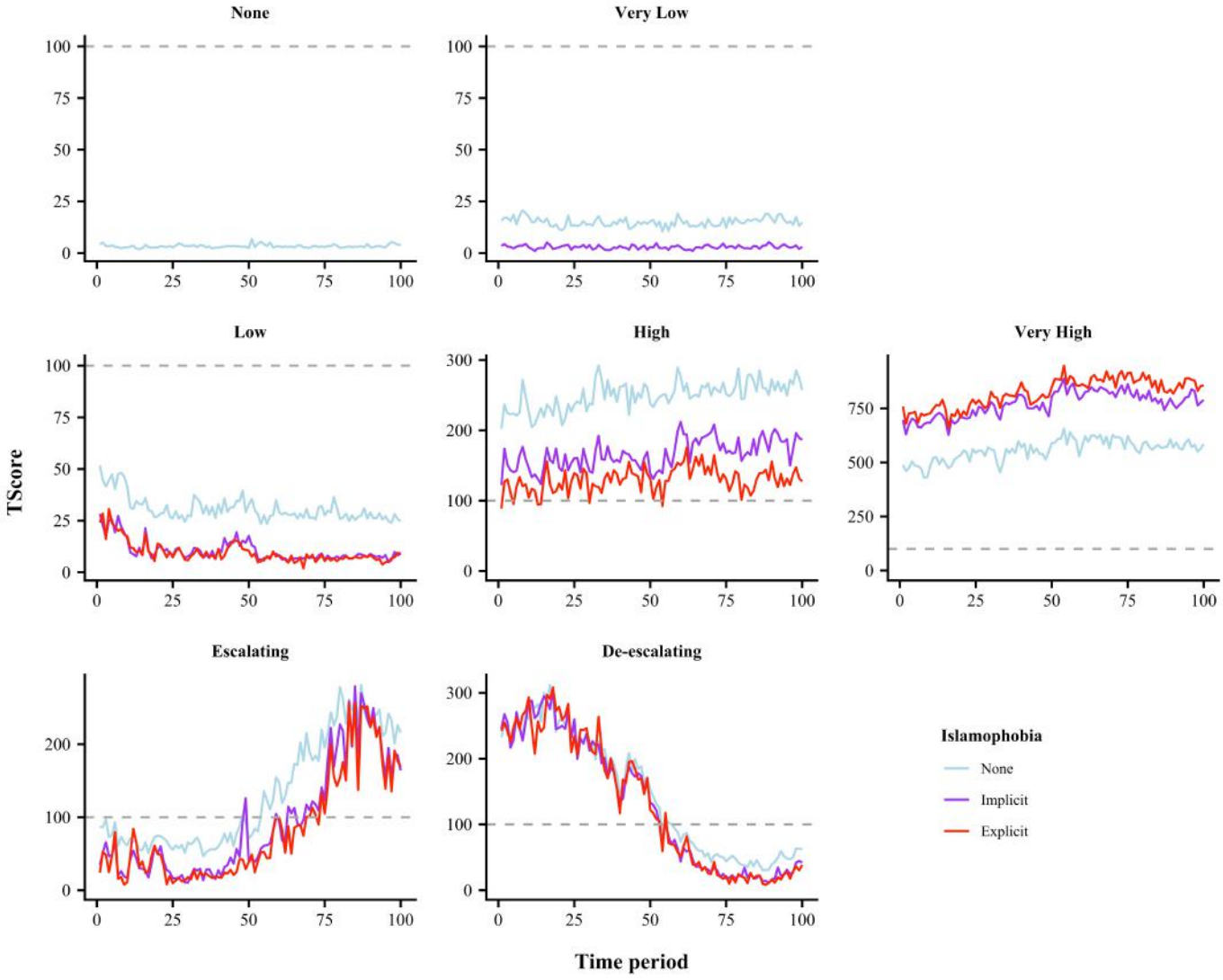

Figure 3, TScores for the 7 trajectories of Islamophobia over the 100 time periods of 52,213 tweets. The grey dotted line shows the average tweeting for the entire cohort. Scales are free to vary.

The TScores for five of the seven trajectories (None, Very Low, Low, High and Very High) are broadly stable over time, with only very weak trends observed (i.e. for Low the TScores decrease slightly over time and for Very High they slightly increase). These trajectories primarily capture differences in the prevalence of tweeting, but also reveal some qualitative

differences. For instance, the Very High trajectory is the only one in which the TScore for Implicit and Explicit Islamophobia is greater than the TScore for None. This indicates that users in this trajectory engage in more Islamophobic behaviour both in absolute terms and as a proportion of their total behaviour. The relationship between Implicit and Explicit Islamophobic tweeting differs. For Low, the TScores are closely aligned, for High, the TScore for Implicit is greater than for Explicit, and for Very High the TScore for Explicit is greater than Implicit. The consistency of users' behaviour also varies between trajectories. The High trajectory has greater variance over time than the Low and Very High trajectories. This suggests these users are more susceptible to exogenous shocks, which could be driving short-term changes. The Escalating and De-escalating trajectories differ from the other trajectories because the users show a clear change in behaviour. For both Implicit and Explicit Islamophobia, the TScores for the Escalating trajectory start from below average (~30) and finish far above (~170). In contrast, in the De-escalating trajectory the TScores for Implicit and Explicit tweeting are high at the start (~240) and far lower at the end (~40). This analysis shows that the trajectories are a useful way of identifying quantitative, qualitative and temporal differences between users.

Table 3, Descriptions of the 7 trajectories of Islamophobia

Name	Description
None	Users who never engage in any form of Islamophobia (whether Implicit or Explicit).
Very Low	Users who engage in very little Implicit Islamophobia and no Explicit Islamophobia.
Low	Users who engage in both Implicit and Explicit Islamophobia, far below the average level.
High	Users who consistently engage in an above average level of Implicit and Explicit Islamophobia.
Very High	Users who consistently engage in a high level of Islamophobia and are comparatively more likely to engage in Explicit.
Escalating	Users whose Islamophobia is increasing over time.
De-escalating	Users whose Islamophobia is decreasing over time.

Quantitative differences between the trajectories are shown in Figure 4, which depicts the mean and dispersion of the number of each type of tweets for each trajectory. The figures are provided in the Methods appendix. The differences between the trajectories are highly statistically significant. We conduct non-parametric omnibus tests of statistical significance: ANOVA type, Wilks' Lambda type, Lawley Hotelling type and Bartlett Nanda Pillai type, as well as permutation variations [63]. These are all statistically significant ($p < 0.000001$). We further verify this with Kruskall-Wallis tests on each of the three types of tweeting. In all cases, the differences are statistically significant ($p < 0.000001$). We then conduct pairwise Wilcoxon

rank sum tests on each pair of trajectories, and all differences are significant ($p < 0.000001$). These results provide strong evidence that differences between trajectories are significant.

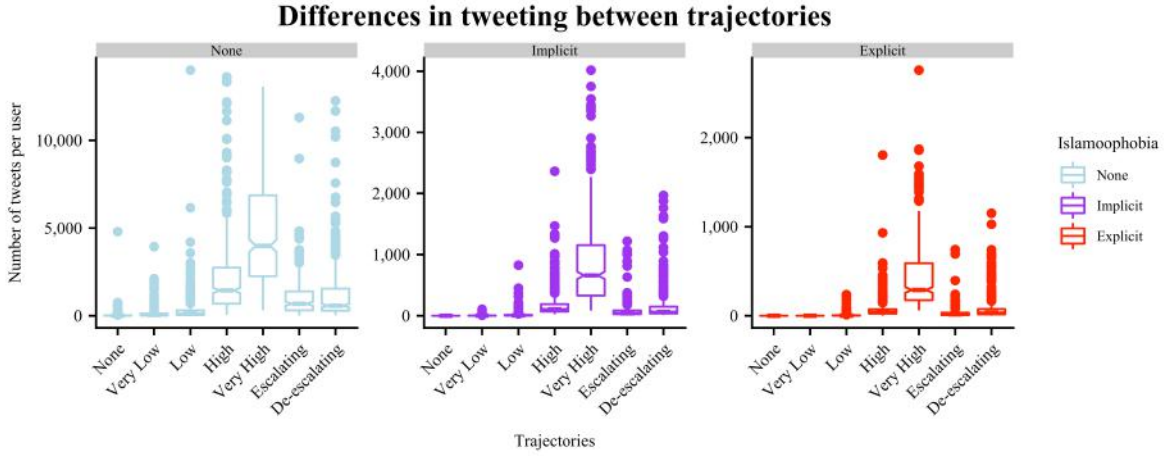

Figure 4, The mean and dispersion of each type of tweeting (None, Implicit and Explicit) for the 7 trajectories of Islamophobia.

The number of users in each trajectory varies considerably, as shown in Figure 5. The most prevalent is None, which accounts for 28.9% of users. Very Low comprises 14.0%. It is plausible that many of the users in these trajectories (total, 42.9%) are less committed to far right politics. They may be followers of other political parties, journalists or academics. The Low trajectory comprises a further 27.0% of users. The two most concerning trajectories are the users in High and Very High (9.2% and 8.8% respectively); these 18.0% are perpetually engaging in considerable levels of Islamophobia. Finally, the two trajectories with the most noticeable temporal trends, Escalating and De-escalating, comprise 4.8% and 7.4% respectively. There are 50% more De-escalating compared to Escalating users. The greater proportion of De-escalating most likely reflects the fact that several Islamist terrorist attacks occurred at the start of the period, which might have motivated some users to send many Islamophobic tweets in response [49].

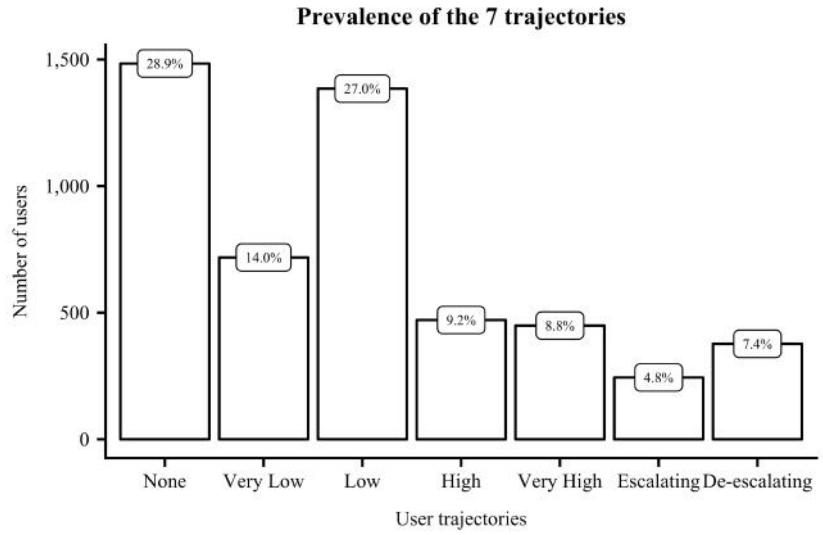

Figure 5, The number and percentage of users assigned to each of the 7 trajectories of Islamophobia.

Discussion and Conclusion

We have investigated the dynamics of Islamophobia amongst followers of the BNP on Twitter and have identified seven distinct behavioural trajectories, which capture quantitative, qualitative and temporal differences in users' behaviour. Our results offer a new way of understanding and characterising the far right which can, in turn, be used inform future areas of research, such as investigating the causes and self-perception of far right actors. Overall, our findings support the view that the followers of the BNP on Twitter comprise a shifting, complex assemblage of individuals, as proposed by Ganesh's concept of the 'swarm', rather than just a homogenous group of 'bloody nasty people', as proposed by Trilling. We anticipate that other far right groups, on other platforms, might exhibit different behavioural trajectories and that the prevalence of those trajectories is likely to differ. Nonetheless, we anticipate that the key argument would hold; the far right is not a single homogeneous group of Islamophobes, which can be easily represented through summary statistics, but a heterogeneous mix of individuals, exhibiting many different behavioural trajectories. These findings could be generalised by identifying trajectories across different types of hate, such as xenophobia, racism and homophobia. They can also be used to increase understanding of the dynamics of Islamophobia on social media more widely, including amongst followers of other political parties and non-political users. To enable future researchers to use and develop our findings, we have made the ID strings of our 5 million tweet dataset publicly available.

This work not only contributes to empirical knowledge but also informs policy discussions regarding how to tackle the rise of far right extremism and online hate. First, the results show that platforms, such as Twitter, need to adopt a more holistic approach which focuses on the behavioural patterns of users rather than just single bits of content. Second, the results could be used by policymakers to prioritise how resources to counter hate speech and provide support to victims should be allocated. Resources should be allocated to reflect the differing challenges posed by the infrequent and more nascent Islamophobia expressed by users in the Low trajectory compared with the constant and aggressive Islamophobia expressed by users in Very High. Third, this research could provide a more nuanced way of planning and targeting interventions against hateful users. Users in different trajectories may respond differently to available policies, such as bans, filters and demonetisation and counter-speech. Finally, fourth, the Escalating and De-escalating trajectories present opportunities for policymakers to both better understand and tackle far right extremism. Further research would benefit by investigating why these individuals change their behaviour and, in the case of the Escalating trajectory, developing ways of intervening at an early stage. Potentially, an advanced predictive model could flag users who are likely to escalate at an early stage in the process and ethical efforts could be made to address their actions. These policy implications could be evaluated further in future work, such as by investigating a more diverse range of online spaces, including niche platforms like 8chan, Discord and some communities on Reddit, to see whether similar behavioural patterns are observed.

There are several limitations of the current research. The LM model is fit on a simplified representation of the data, comprising just the strongest expression that each user tweets in each time period. As such, the volume of tweets is not modelled directly. Nonetheless, the model performs well at capturing differences in not only the strength but also the volume of Islamophobic tweets. This is because there is an underlying association between the two, which the LM model picks up on as we fit a reasonably large number of time periods (100). In the future, this could be included addressed through using, for instance, a continuous multivariate LM model. Evaluating performance with an unsupervised method is inherently difficult, as

many standard evaluative metrics, such as the R^2, cannot be calculated. The results presented here indicates the existence of several distinct trajectories, which have been verified by statistical significance testing. Future work should aim to increase the robustness of the modelling and to verify the findings. A further area of investigation is whether these *behavioural* differences we observe between trajectories align with *attitudinal* differences, as these are not necessarily concomitant. Investigating the extent to which these are associated would require a larger multi-methodological research design to verify.

This research provides a new way of characterising far right actors online and also directly informs policy discussions around how their hateful behaviour can be tackled. Using a mixture of social and computational sciences, our primary contribution is to identify the internal heterogeneity of the far right and establish the need for more nuanced assessments of far right behaviour online.

Trajectories of Islamophobia | Appendix

1. Islamophobia classifier

The classifier for Islamophobic content is described in detail in (Vidgen & Yasseri 2018). It was trained on a newly annotated dataset of 4,000 tweets and achieves balanced accuracy of 0.83 and a micro-F1 score of 0.78 when tested on an unseen 300 tweet dataset. Precision is 0.78, which is far above the 0.7 minimum recommended by van Rijsbergen for empirical research [64]. The main sources of classification error are (1) confusing implicit and explicit Islamophobia and (2) confusing non-Islamophobic with implicit Islamophobic, specifically content which (a) discusses Muslims and Islam but in a non-hateful way and (b) is hateful against another target, such as immigrants or minority ethnic groups. Overall, the classifiers' performance compares well with other ternary multi-class classifiers for abusive content and is suitable for empirical research [65], [66].

The Implicit/explicit Islamophobia distinction in the classifier is analytical and does not capture moral distinctions or directly inform whether statements should be considered permissible on social media platforms, given the protections afforded to freedom of expression. Consider a news report about Islamist terrorist activity, e.g. 'Muslim terrorist attacks London bridge'. On the one hand, this is a factual statement. On the other, it frames Muslims in a negative way by associating them with a reviled trait, terrorism. Speakers may have myriad motivations in making this statement; they may want to report a news event of nationwide significance, spread negative views about Muslims and thereby stir opposition, or use it as a starting point to discuss multicultural integration. Irrespective of these differing intentions, all have served the same purpose: to reproduce a negative framing of Muslims.

2. Latent Markov model fitting

LM models are tested for 1 to 12 latent states and evaluated with AIC and BIC. Both measures are closely aligned and indicate that a range of between 3 and 7 latent states is optimal, as shown in Figure 6. Fitting a number of latent states towards the top-end of the indicated range (e.g. 5 to 7) is problematic because each latent state accounts for a specific range of behaviours. Because the states are highly tailored to small subsets of users, it is less likely that users will transition between states. As such, the transitional probabilities become very high for remaining in the same state. For instance, in a model with 5 latent states, the transitional probabilities for remaining in the same state are all over 0.98. This makes it harder for the models to capture longitudinal changes in behaviour. As such, we set K to a value at the lower end of the range (K = 3).

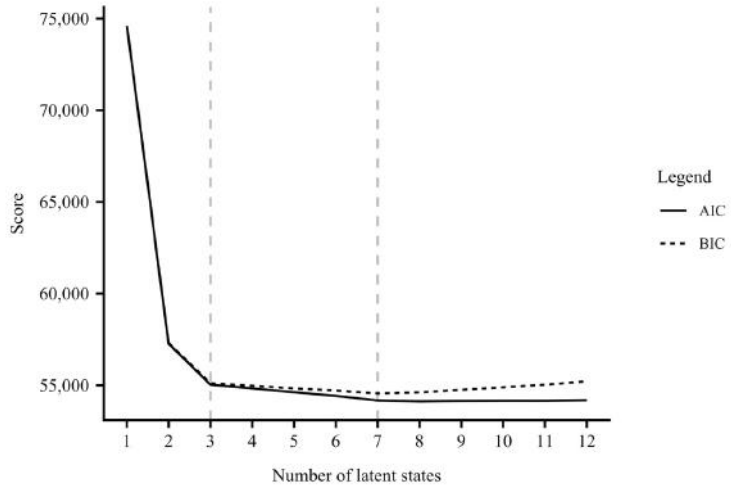

Figure 6, Results of fitting the optimal number of latent states in the LM model, using AIC and BIC.

3. Number of clusters: k-modes fitting

From the output of the LM model, we test for between 2 and 20 clusters with the k-modes algorithm (referred to as 'trajectories' in the main body of the paper), evaluated by measuring the Within Sum of Squares. The results indicate a range of between 5 and 8 clusters is optimal. This is shown in Figure 7.

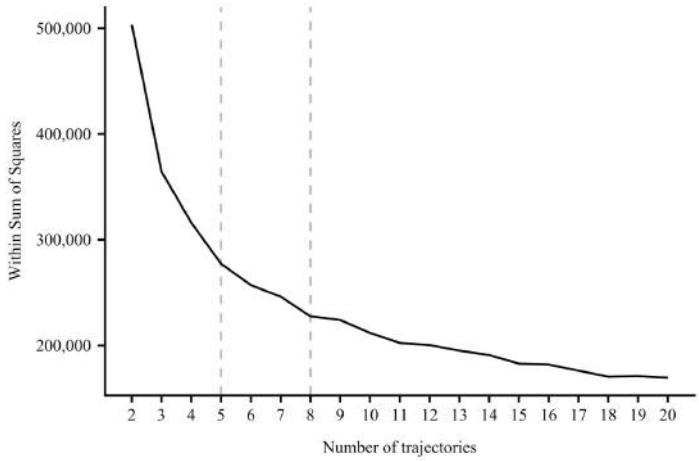

Figure 7, Results of fitting for the optimal number of trajectories, using Within Sum of Squares.

4. Quantitative differences between the 7 trajectories

Table 4, Quantitative differences in None, Implicit and Explicit Islamophobic tweeting between the 7 trajectories of Islamophobia.

	None		Implicit		Explicit		All
Trajectory	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Total number of tweets
None	28 (100%)	139	0 (0%)	0	0 (0%)	0	28
Very Low	128 (97%)	266	3 (3%)	7	0 (0%)	0	131
Low	263 (94%)	541	12 (4%)	32	6 (2%)	13	281
High	2,137 (89%)	2,240	182 (8%)	223	70 (3%)	119	2,389
Very High	4,762 (79%)	3,084	844 (14%)	692	439 (7%)	386	6,045
Escalating	1,140 (90%)	1,335	92 (7%)	166	37 (3%)	77	1,269
De-escalating	1,211 (83%)	1,668	160 (11%)	271	83 (6%)	144	1,454

References

[1] C. Atton, “Far-right media on the Internet: culture, discourse and power.,” *New Media Soc.*, vol. 8, no. 4, pp. 573–587, 2006.

[2] R. Engström, “The Online Visual Group Formation of the Far Right: A Cognitive-Historical Case Study of the British National Party,” *Public J. Semiot.*, vol. 6, no. 1, pp. 1–21, 2014.

[3] P. Webster, “Religious discourse in the archived web: Rowan Williams, Archbishop of Canterbury, and the sharia law controversy of 2008,” in *The Web as History*, N. Brügger and R. Schroeder, Eds. London: UCL Press, 2017, pp. 190–203.

[4] L. Bowman-Grieve, “Exploring stormfront: A virtual community of the radical right,” *Stud. Confl. Terror.*, vol. 32, no. 11, pp. 989–1007, 2009.

[5] W. De Koster and D. Houtman, “‘Stormfront Is Like a Second Home To Me,’” *Information, Commun. Soc.*, vol. 11, no. 8, pp. 1155–1176, 2008.

[6] L. Figea, L. Kaati, and R. Scrivens, “Measuring online affects in a white supremacy forum,” *IEEE Int. Conf. Intell. Secur. Informatics Cybersecurity Big Data, ISI 2016*, pp. 85–90, 2016.

[7] C. Froio and B. Ganesh, “The transnationalisation of far right discourse on Twitter,” *Eur. Soc.*, vol. 0, no. 0, pp. 1–27, 2018.

[8] B. Ganesh, “The Ungovernability of Digital Hate Culture,” *J. Int. Aff.*, vol. 71, no. 2, pp. 30–49, 2018.

[9] P. Jackson and M. Feldman, “The EDL: Britain’s ‘New Far Right’ social movement,” 2011.

[10] N. Hatakka, “When logics of party politics and online activism collide: The populist Finns Party’s identity under negotiation,” *New Media Soc.*, vol. 19, no. 12, pp. 2022–2038, 2017.

[11] J. A. Schafer, “Spinning the web of hate: web-based hate propagation by extremist organizations,” *J. Crim. Justice Pop. Cult.*, vol. 9, no. 2, pp. 69–88, 2002.

[12] M. Wojcieszak, “‘Don’t talk to me’: Effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism,” *New Media Soc.*, vol. 12, no. 4, pp. 637–655, 2010.

[13] Hope Not Hate, *Hope Not Hate: State of Hate 2017*. London: Hope Not Hate, 2017.

[14] G. E. Hine *et al.*, “Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan’s Politically Incorrect Forum and Its Effects on the Web,” in *ICWSM*, 2017, pp. 92–101.

[15] P. Ignazi, “Extreme Right Parties in Western Europe,” *Extrem. Right Parties West. Eur.*, no. April 2016, pp. 1–270, 2005.

[16] M. Goodwin, “Forever a False Dawn? Explaining the Electoral Collapse of the British National Party (BNP),” *Parliam. Aff.*, vol. 67, no. 4, pp. 1–20, 2013.

[17] G. O. Edwards, “A comparative discourse analysis of the construction of ‘in-groups’ in the 2005 and 2010 manifestos of the British National Party,” *Discourse Soc.*, vol. 23, no. 3, pp. 245–258, 2012.

[18] F. P. Su, "Patent Priority Network : Linking Patent Portfolio," *J. Am. Soc. Inf. Sci.*, vol. 60, no. 1999, pp. 2353–2361, 2009.

[19] P. John and H. Margetts, "The Latent Support for the Extreme Right in British Politics," *West Eur. Polit.*, vol. 32, no. 3, pp. 496–513, 2009.

[20] E. Carter, "Right-wing extremism / radicalism: reconstructing the concept," *J. Polit. Ideol.*, vol. 23, no. 2, pp. 157–182, 2018.

[21] C. Mudde, *Populist Radical Right Parties in Europe*. Cambridge: Cambridge University Press, 2007.

[22] M. Golder, "Far Right Parties in Europe," *Annu. Rev. Sociol.*, vol. 19, no. 1, pp. 477–497, 2016.

[23] J. E. Richardson and R. Wodak, "Recontextualising fascist ideologies of the past: right-wing discourses on employment and nativism in Austria and the United Kingdom," *Crit. Discourse Stud.*, vol. 6, no. 4, pp. 251–267, 2017.

[24] N. Copsey, "Changing course or changing clothes? Reflections on the ideological evolution of the British National Party 1999-2006," *Patterns Prejudice*, vol. 41, no. 1, pp. 61–82, 2007.

[25] G. J. Severs, "The 'obnoxious mobilised minority': homophobia and homohysteria in the British National Party, 1982 – 1999," *Gend. Educ.*, vol. 29, no. 2, pp. 165–181, 2017.

[26] R. Ford and M. J. Goodwin, "Angry white men: Individual and contextual predictors of support for the british national party," *Polit. Stud.*, vol. 58, no. 1, pp. 1–25, 2010.

[27] J. V Gottlieb, "Women and British Fascism Revisited: Gender, the Far-Right, and Resistance," *J. Womens. Hist.*, vol. 16, no. 3, pp. 108–123, 2017.

[28] E. Bayrakli and F. Hafez, *European Islamophobia Report 2017*. Istanbul: SETA, 2018.

[29] C. Allen, *Islamophobia*. Surrey: Ashgate, 2011.

[30] R. Eatwell and M. J. Goodwin, *The New Extremism in the 21st Century*. Abingdon: Routledge, 2010.

[31] J. P. Zúquete, "The European extreme-right and Islam: New directions?," *J. Polit. Ideol.*, vol. 13, no. 3, pp. 321–344, 2008.

[32] I. Awan, "Islamophobia and Twitter: A typology of online hate against muslims on social media," *Policy and Internet*, vol. 6, no. 2, pp. 133–150, 2014.

[33] I. Awan and I. Zempi, "'I will blow your face off' - Virtual and physical world anti-muslim hate crime," *Br. J. Criminol.*, vol. 57, no. 2, pp. 362–380, 2017.

[34] G. Evolvi, "Hate in a Tweet: Exploring Internet-Based Islamophobic Discourses," *Religions*, vol. 9, no. 10, p. 307, 2018.

[35] C. Froio, "Race, religion, or culture? Framing Islam between racism and neo-racism in the online network of the French far right," *Perspect. Polit.*, vol. 16, no. 3, pp. 696–709, 2018.

[36] J. Bartlett and M. Littler, *Inside the EDL: Populist politics in a digital age*. London: DEMOS, 2011.

[37] D. Trilling, Bloody Nasty People: the rise of Britain's Far Right. London: Verso, 2012.

[38] M. Biggs and S. Knauss, “Explaining membership in the British National Party: a multilevel analysis of contact and threat,” *Eur. Sociol. Rev.*, vol. 28, no. 5, pp. 633–646, 2012.

[39] J. Rydgren, “The Sociology of the Radical Right,” *Annu. Rev. Sociol.*, vol. 33, no. 1, pp. 241–262, 2007.

[40] E. Chandrasekharan, U. Pavalanathan, E. Gilbert, A. Srinivasan, A. Glynn, and J. Eisenstein, “You Can’t Stay Here,” *Proc. ACM Human-Computer Interact.*, vol. 1, no. 2, pp. 1–22, 2017.

[41] I. Awan, “Islamophobia on Social Media: A Qualitative Analysis of the Facebook’s Walls of Hate,” *Int. J. Cyber Criminol.*, vol. 10, no. 1, pp. 1–20, 2016.

[42] W. Van Der Brug, M. Fennema, and J. Tillie, “Anti-immigrant parties in Europe: Ideological or protest vote?,” *Eur. J. Polit. Res.*, vol. 37, no. 1, pp. 77–102, 2000.

[43] D. Cutts, R. Ford, and M. J. Goodwin, “Anti-immigrant, politically disaffected or still racist after all? Examining the attitudinal drivers of extreme right support in Britain in the 2009 European elections,” *Eur. J. Polit. Res.*, vol. 50, no. 3, pp. 418–440, 2011.

[44] J. Rhodes, “‘It’s not just them, it’s whites as well’: Whiteness, class and BNP support,” *Sociology*, vol. 45, no. 1, pp. 102–117, 2011.

[45] A. M. Bliuc, N. Faulkner, A. Jakubowicz, and C. McGarty, “Online networks of racial hate: A systematic review of 10 years of research on cyber-racism,” *Comput. Human Behav.*, vol. 87, no. May, pp. 75–86, 2018.

[46] Demos, *Anti-Islamic hate on Twitter*. London: DEMOS, 2017.

[47] A. Alrababa, W. Marble, S. Mousa, and A. Siegel, “Can Exposure to Celebrities Reduce Prejudice ? The Effect of Mohamed Salah on Islamophobic Behaviors and Attitudes,” Zurich, 2019.

[48] M. Williams and P. Burnap, “Cyberhate on social media in the aftermath of Woolwich: a case study in computational criminology and big data,” *Br. J. Criminol.*, vol. 56, no. 1, pp. 211–238, 2016.

[49] P. Burnap *et al.*, “Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack,” *Soc. Netw. Anal. Min.*, vol. 4, no. 1, pp. 1–14, 2014.

[50] B. Kollanyi, P. N. Howard, and S. C. Woolley, “Bots and Automation over Twitter during the First U.S. Election,” *Comprop Data Memo*, no. 4, pp. 1–5, 2016.

[51] A. Larsson and M. Hallvard, “Bots or journalists? News sharing on Twitter,” *Communications*, vol. 40, no. 3, pp. 361–370, 2015.

[52] Z. Iqbal, “Understanding Islamophobia: Conceptualizing and Measuring the Construct,” *Eur. J. Soc. Sci.*, vol. 13, no. 4, pp. 55–62, 2010.

[53] T. Kayaoğlu and T. Kayaoglu, “Three takes on Islamophobia,” *Int. Sociol.*, vol. 27, no. 5, pp. 609–615, 2012.

[54] S. Sayyid, “A Measure of Islamophobia,” *Islam. Stud. J.*, vol. 2, no. 1, pp. 10–25, 2014.

[55] E. Bleich, “What is Islamophobia and how much is there? theorizing and measuring an emerging comparative concept,” *Am. Behav. Sci.*, vol. 55, no. 12, pp. 1581–1600, 2011.

[56] B. Vidgen, R. Tromble, A. Harris, S. Hale, D. Nguyen, and H. Margetts, "Challenges and frontiers in abusive content detection," in *3rd Workshop on Abusive Language Online*, 2019.

[57] R. Benford and D. Snow, "Framing Processes and Social Movements: An Overview and Assessment," *Annu. Rev. Sociol.*, vol. 26, no. 1, pp. 611–639, 2000.

[58] B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media," *ArXiv Prepr. 1812.10400*, pp. 1–6, 2018.

[59] G. Spedicato and M. Signorelli, "The markovchain Package: A Package for Easily Handling Discrete Markov Chains in R," *R CRAN*, pp. 1–67, 2013.

[60] F. Bartolucci, A. Farcomeni, and F. Pennoni, "An overview of latent Markov models for longitudinal categorical data," 2010.

[61] C. McCauley and S. Moskalenko, "Mechanisms of political radicalization: Pathways toward terrorism," *Terror. Polit. Violence*, vol. 20, no. 3, pp. 415–433, 2008.

[62] J. J. Marquez, A. Downey, and R. Clement, "Walking a Mile in the User's Shoes: Customer Journey Mapping as a Method to Understanding the User Experience," *Internet Ref. Serv. Q.*, vol. 20, no. 3–4, pp. 135–150, 2015.

[63] A. Ellis, W. Burchett, S. Harrar, and A. Bathke, "Nonparametric Inference for Multivariate Data: The R Package npmv," *J. Stat. Softw.*, vol. 76, no. 4, pp. 1–18, 2017.

[64] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.

[65] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*, vol. 6, pp. 13825–13835, 2018.

[66] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in *ICWSM*, 2017, pp. 1–4.