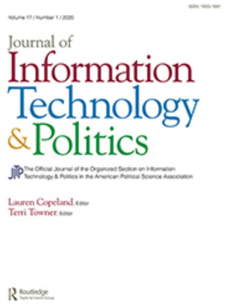

Journal of Information Technology & Politics

Lauren Copeland, Editor
Terri Towner, Editor

Journal of Information Technology & Politics

ISSN: 1933-1681 (Print) 1933-169X (Online) Journal homepage: https://www.tandfonline.com/loi/witp20

Detecting weak and strong Islamophobic hate speech on social media

Bertie Vidgen & Taha Yasseri

To cite this article: Bertie Vidgen & Taha Yasseri (2020) Detecting weak and strong Islamophobic hate speech on social media, Journal of Information Technology & Politics, 17:1, 66-78, DOI: 10.1080/19331681.2019.1702607

To link to this article: https://doi.org/10.1080/19331681.2019.1702607

Published online: 13 Dec 2019.

JOURNAL OF INFORMATION TECHNOLOGY & POLITICS
2020, VOL. 17, NO. 1, 66–78
https://doi.org/10.1080/19331681.2019.1702607

Detecting weak and strong Islamophobic hate speech on social media

Bertie Vidgen and Taha Yasseri

ABSTRACT

Islamophobic hate speech on social media is a growing concern in contemporary Western politics and society. It can inflict considerable harm on any victims who are targeted, create a sense of fear and exclusion amongst their communities, toxify public discourse and motivate other forms of extremist and hateful behavior. Accordingly, there is a pressing need for automated tools to detect and classify Islamophobic hate speech robustly and at scale, thereby enabling quantitative analyses of large textual datasets, such as those collected from social media. Previous research has mostly approached the automated detection of hate speech as a binary task. However, the varied nature of Islamophobia means that this is often inappropriate for both theoretically informed social science and effective monitoring of social media platforms. Drawing on in-depth conceptual work we build an automated software tool which distinguishes between non-Islamophobic, weak Islamophobic and strong Islamophobic content. Accuracy is 77.6% and balanced accuracy is 83%. Our tool enables future quantitative research into the drivers, spread, prevalence and effects of Islamophobic hate speech on social media.

KEYWORDS
Hate speech; Islamophobia; prejudice; social media; natural language processing; machine learning

Introduction

Islamophobia is a growing concern in contemporary Western politics and society, with research indicating that it is both widespread within the far right (Hope Not Hate, 2017) and has entered mainstream political discourse (Poynting & Briskman, 2018). In particular, the spread of Islamophobic hate speech on social media has received considerable attention from the government (All Party Parliamentary Group on British Muslims, 2018), Muslim community groups (Ingham-Barrow, 2018), academics (Williams & Burnap, 2016) and the platforms themselves (Twitter, 2018). Islamophobia on platforms such as Twitter is highly concerning given that they are used by elected representatives to communicate with the public, and for "political talk" more broadly (Jungherr, 2016). Much existing research into Islamophobia has been qualitative in nature (Ad-Dab'bagh, 2017; Awan, 2016; Mondon & Winter, 2017) and there is a pressing need for theoretically informed large-scale quantitative studies to deepen our understanding of its drivers and dynamics. This requires robust computational tools which can be deployed rapidly and systematically. In this paper, we report on the creation and performance of an automated software tool (specifically, a machine learning classifier), in order to address a single research aim:

> *To create a classifier for detecting Islamophobic content on social media which distinguishes between different strengths of Islamophobia.*

The code, annotation guidelines and word embeddings model used to develop this tool are available in the online supplement, thus enabling other researchers to use and extend the research.[1]

Islamophobia: exploring a concept

Online Islamophobic hate speech poses myriad problems for society: it can inflict harm on any victims who are targeted, create a sense of fear and exclusion amongst their communities, toxify public discourse and motivate other forms of extremist and hateful behavior through a cycle of "cumulative extremism" (Akgönül, Alibašić, Nielsen, & Račius, 2018; Bakali, 2019; Busher & Macklin, 2014; Walia, Khan, & Islam, 2019). However, there is considerable disagreement about what the term "Islamophobia" means. It is what Gallie terms an "essentially contested concept" – there are numerous definitions and descriptions of

CONTACT Bertie Vidgen bvidgen@turing.ac.uk The Oxford Internet Institute, University of Oxford, Oxford, United Kingdom

Islamophobia but little consensus as to what its core features are (Gallie, 1956). It has been described as a form of racism (Meer & Modood, 2009), stereotyping (Moosavi, 2015), prejudice (Imhoff & Recker, 2012), fear (Kunst, Sam, & Ulleberg, 2013), and exclusion (Bayrakli & Hafez, 2018). In the present work, we use Bleich's widely cited definition of Islamophobia, as this captures conceptual arguments posed by other leading theorists in the field (C. Allen, 2011) and is similar to definitions offered by the Runnymede trust and the All-Party Parliamentary Group on British Muslims (All Party Parliamentary Group on British Muslims, 2018; Runnymede Trust, 2017). Bleich defines Islamophobia as: "Indiscriminate negative attitudes or emotions directed at Islam or Muslims." (Bleich, 2011). We adapt Bleich's definition for social media content:

> *"Content which is produced or shared which expresses indiscriminate negativity against Islam or Muslims."*

Bleich's definition captures a wide variety of hateful behaviors which take place online, including threats, insults, dehumanizing language, derogative statements and the use of slurs. Its key strength is that it provides a conceptual orientation for understanding the core basis of Islamophobia, rather than just offering a prescriptive list of its main features (which could fast become outdated as linguistic practices change and may be unsuitable in certain contexts). In particular, Bleich's definition focuses on group-level rather than person-level negativity. For instance, a threat such as "I want to kick all Muslims out of the UK" does not express hatred against a particular person but, rather, expresses dislike of, and opposition to, Muslims in general.

A key part of conceptual debates about Islamophobia is whether it pertains to anti-Islamism (which can be understood as opposition against Islam *qua*religion/institution) or anti-Muslimism (which can be understood as opposition against Muslims *qua* social group) – or, indeed, both together. Research in social psychology often emphasizes that prejudice refers solely to the treatment of individuals/groups and not to institutions or ideologies (Brown, 2010; Pettigrew, Tropp, Wagner, & Christ, 2011). Some sociological researchers make a similar argument; in a discussion of hate speech Rosenfeld explicitly states that "disparaging religion cannot […] be equated with disparaging the religious." (Rosenfeld, 2012, p. 277). However, others who study Islamophobia, and in particular those who work closely with victims, argue that it should include both anti-Islamism and anti-Muslimism (Awan & Zempi, 2016; Chakraborti & Garland, 2012). Bleich argues that "Islam and Muslims are often inextricably intertwined in individual and public perceptions" (Bleich, 2011) and Allen similarly finds that Islamophobes often criticize Islam as a proxy for criticizing Muslims (C. Allen, 2017). Given this, we include both anti-Islamism and anti-Muslims within our definition of Islamophobia, although we recognize that this may not be appropriate for research in other settings.

Researching islamophobic hate speech on social media

Studying social media data poses unique challenges compared with studying traditional political science data sources, such as survey data and political manifestos (Margetts, 2017). Social media datasets often contain millions of data points and, in most cases, qualitative approaches cannot be scaled to handle such large volumes, even if work is conducted by research teams or uses crowdsourcing. This problem is particularly acute when studying online hate as its prevalence in "the wild" is very low. Studies which randomly sample tweets from Twitter's 1% stream and annotate them typically report that fewer than 1 in 100 tweets contain any form of hate, and sometimes as few as 1 in 1000 (Schmidt & Wiegand, 2017). Computational tools are therefore crucial in this domain to enable social science researchers to study complex issues in the wide and diverse landscape of social media. This is crucial for advancing the field of online Islamophobia research which remains at a nascent stage as many key questions have only been partially addressed (Bliuc, Faulkner, Jakubowicz, & McGarty, 2018; Vidgen et al., 2019).

Over the past decade, the social sciences have undergone a so-called "computational turn" (Blei

& Smyth, 2017; Conte et al., 2012). This has been received with considerable optimism: Lazer et al. argue that computational social science has led to a newfound level of granularity and breadth, with little trade-off between them (Lazer et al., 2009), Watts argues that big data can provide insight into age-old questions by "mak[ing] visible social processes that are much more difficult to study in conventional organizational settings" (Watts, 2007) and in political science Grimmer and Stewart argue that "automated content methods can make possible the previously impossible" (Grimmer & Stewart, 2013). However, at the same time, many researchers caution that big data plus computation does not always equal good scientific research. In relation to ethnography, Manovich argues that "algorithms used by computer scientists […] will never arrive at the same insights and understanding of people and dynamics in the community" (Manovich, 2011, p. 8). Boyd and Crawford similarly make the point that some stories and processes cannot be uncovered just "by farming millions of Facebook or Twitter accounts" but, instead, require in-depth qualitative and ethnographic work (Boyd & Crawford, 2012). Others have drawn attention to how computational social sciences often foreground the *computational* aspect more heavily than the *social* (Cowls & Schroeder, 2015); as Cihon and Yasseri note, "despite its name, [computational social science] has drawn from computer scientists, mathematicians and physicists far more than social scientists" (Cihon & Yasseri, 2016).

Thus, whilst computational methods can open new and innovative ways of studying behavior on social media, they also risk being overly reductive and not taking into account important social insights. This issue is particularly important when studying online hate speech because what is perceived to be hateful is deeply contested, and can vary depending on which group is targeted (many people are more attuned to racism than Islamophobia (Runnymede Trust, 2017)), the subjective outlook of the individual (which can be affected by their background, life experiences and values (Salminen, Veronesi, Almerekhi, Jung, & Jansen, 2018)) and the context (who speaks, when, with what authority, and to whom, can all affect the *meaning* of any bit of content (Leader Maynard & Benesch, 2016)).

Interventions in social psychology also point to the multifaceted nature of prejudice, which includes both explicit, overt and direct behaviors as well as those which are implicit, covert and indirect (Pettigrew & Meertens, 1995). Much recent research has focused on "everyday" prejudicial and hateful actions (Moosavi, 2015) as well as "micro-aggressions" (Haque, Tubbs, Kahumoku-Fessler, & Brown, 2018). However, these distinctions between different types of speech have not been fully explored with regard to online hate speech. This is surprising given that such distinctions offer considerable empirical and theoretical advantages over using a single category of "Islamophobia". Making more fine-grained distinctions would enable researchers to better understand the temporal, geographic and network dynamics of Islamophobia (which may differ across strengths), investigate radicalization processes (whereby individuals progress from being weakly to strongly Islamophobic) and understand how Islamophobia enters into mainstream discourses and gains widespread acceptance. It could also enable platforms and governments to better regulate and monitor social media and develop more nuanced content moderation and intervention strategies (Gillespie, 2018).

To better research Islamophobic hate on social media, we need tools and methods which address the two issues discussed here. First, tools need to be scalable and capable of handling large volumes of content. Second, they need to draw on theoretical insights, such as the varying strength of hate, so that they can be used for social scientific analysis. This is reflected in the aim of this paper, which was outlined in the Introduction:

> *To create a classifier for detecting Islamophobic content on social media which distinguishes between different strengths of Islamophobia.*

Defining weak and strong islamophobia

Drawing on Bleich's definition of Islamophobia, as well as work undertaken with victims of Islamophobia by relevant charities (Ingham-

Barrow, 2018), we define strong Islamophobia on social media as:

> *"Content which explicitly expresses negativity against Muslims."*

We define weak Islamophobia on social media as:

> *"Content which implicitly expresses negativity against Muslims."*

We use the term "negativity" rather than "hate" to capture the multifaceted ways in which Islamophobia manifests, from cold dismissal and pseudo-evidential bigotry to angry vitriol (Awan & Zempi, 2016). The distinction between explicit and implicit negativity builds on previous hate speech research, such as Wasseem et al. (Waseem, Davidson, Warmsley, & Weber, 2017) and Kumar et al. (Kumar, Ojha, Malmasi, & Zampieri, 2018). Note that the terms "implicit" and "explicit" simply capture how overt the negativity is. They do not contain any assessment of the intention of actors or the impact of speech on victims. Despite the analytical robustness of this two-part conceptualization, explicitness/implicitness is still highly subjective notions; any tweet can be interpreted in many different ways. Below we provide examples of both weak and strong Islamophobia to clarify the distinction. These examples are based on real data but are synthetic, due to the highly sensitive nature of the subject matter and the ethical challenges posed by reporting social media data verbatim (Williams, Burnap, & Sloan, 2017).

Examples of strong islamophobia

Strong Islamophobia includes expressing explicitly negative *views*, such as describing Muslims as barbarians, and calling for prejudicial *actions*, such as demanding that Muslims are forcibly banned from the UK, and expressing negative *emotions* about Muslims, such as anger and distrust, which are often articulated through the use of profanities. Examples of strong Islamophobic tweets include:

"Muslim men groom and rape children"

"Muslim mothers support FGM!"

"Typical, another bloody Muslim just blew himself up"

"Fuck alllll Muslims"

"Muslim invasion, they're going to take over the UK"

"Top European Lawyer says that Muslims don't obey rule of law and should not be allowed to remain in Europe whilst posing a threat"

"The Police target Muslims because they're a problem, new #evidence"

"Huge rally atm against Loughborough Mosque – let's take back our country"

In example 6, the speaker is supposedly reporting someone else's claims (a "top European lawyer" is referenced) – but nonetheless, it is still the speaker who is engaging in Islamophobia as s/he has shared the content. Note also that in determining whether the tweet is Islamophobic (whether strong or weak), the "truth" of its content is not evaluated. Even if a claim is supported by notional evidence, Islamophobia can still be expressed. In any given context, "truth" is always contested, and there is no neutral objective position from which to judge the epistemology validity of any claim (B. Allen, 1996). Thus, whilst intuitively it seems like many Islamophobic tweets contain falsehoods, this is not the conceptual basis on which we decide whether or not they are Islamophobic.

Examples of weak islamophobia

Weak Islamophobia is distinguished from strong based on whether the negativity is implicit. There are two main types of weak negativity. First, emphasizing perceived differences between Muslims and other members of society, such as attributing to Muslims strange or unusual practices. Such content excludes and marginalizes Muslims in an insidious fashion; Muslims are not explicitly targeted and attacked but, rather, their incompatibility is highlighted. This can be seen as implicitly negative as perceived differences are not celebrated but problematized (Bulmer & Solomos, 2015). Examples include:

(1) "Muslims are just different!"
(2) "Muslim food smells so weird"
(3) "Wearing a Burkha doesn't feel very #UK"

The second form of weak negativity is to take the tropes associated with strong negativity (such as claiming that Muslims are terrorists, barbarians or uneducated) and to ostensibly link them to only a small subset of Muslims (e.g. to just one individual terrorist or Muslims only living in one small geographical area, such as Rotherham) – and by doing so to implicitly forge a connection between the negative trope and *all* Muslims, by using the term "Muslims" or "Islam", even with caveats to heighten the specificity (such as "this Muslim terrorist" or "Muslim Men in Rotherham"), an *implicit* connection is established with all Muslims. The key point here is that discourses about pedophiles, terrorists or FGM practitioners can often be articulated without the need to reference Muslim identity. Examples of this type of weak Islamophobia are provided below. In all of the cases, the speaker appears to be commenting on a specific case but still implicitly creates an association with the negative trope and all Muslims.

(4) "Muslim terrorists attack London Bridge"

"Muslim radicals in the desert kill Christian hostage"

"Muslim paedo is a sick f*ck"

Previous work

Previous work in hate speech detection demonstrates the challenges of – but also potential for – creating a classification system which distinguishes between weak and strong Islamophobic speech. To our best knowledge, no previous research has focused specifically on this task, instead developing binary classifiers for Islamophobia (Vidgen et al., 2019). Researchers at DEMOS detect Islamophobic content on Twitter by creating several classifiers which each classifies a separate facet of Islamophobia. These are based on correlations between unigrams and bigrams, and when combined give a single measurement of Islamophobia within tweets (Demos, 2017). Burnap and Williams identify hate directed against Black and ethnic minority groups and religious groups in the wake of the Woolwich terror attack in 2013 (Burnap & Williams, 2015; Williams & Burnap, 2016). They use an ensemble classifier, which aggregates the results of several classifiers, each of which use dependency parsing and a dictionary of hateful terms as the main input. They achieve accuracy of 0.95 and recall of 0.95. Their high performance is partly due to their unbalanced dataset: out of 1,901 tweets, 1,679 (88%) are "benign" and just 222 (12%) are "hateful" – and their classifier performs better on "benign" tweets.

Classifier performance in multi-class hate speech detection tasks, looking at other targets of abuse, is often far lower. As Salminen et al. note, "existing works using multi-label classification for online hate speech are extremely rare, and we could not locate prior work that had achieved good results" (Salminen et al., 2018). Furthermore, most existing research into multiclass classification has focused on distinguishing between different *targets* of hate rather than different *strengths*. Classifying content based on strength rather than target poses additional challenges as there is less variation between classes.

Burnap and Williams train a classifier to distinguish between different levels of cyberhate (divided into "moderate" and "extreme" classes) targeted against Black Minority Ethnic (BME) and religious groups on Twitter (Burnap & Williams, 2016), achieving a precision of 0.77. Malmasi and Zampieri distinguish between "Hate" speech, "Offensive" speech and "OK" speech. They achieve 78% accuracy but on an unevenly weighted dataset – over half of their corpus is "OK". Their model struggles to distinguish between non-OK content; of 2,399 "Hateful" instances in their dataset, 1050 are categorized correctly, 1,113 are miscategorized as "Offensive" and 236 as "OK". They also do not test their model on unseen data, only reporting the results of cross-validation, which could risk overfitting (Malmasi & Zampieri, 2017).

Davidson et al. train a model to distinguish between hate speech and offensive speech, and non-offensive speech in tweets. They report impressive results, with a precision of 0.91, recall of 0.90 and an F1 score of 0.90. Their work demonstrates the potential for multiclass classification and makes an important theoretical argument apropos the need to separate different types of content. However, as they note, their model performs poorly with hate speech, of which almost 40% is misclassified. The high F1 score is large because their classes are very uneven (76% of the data is in the "Offensive" speech category). They also train and test their classifier on a single dataset.

Creating the classifier

We leverage the power of computational analyses, in particular machine learning, to create an automatic software tool which can distinguish between weak Islamophobic, strong Islamophobic and non-Islamophobic content. This involves five steps, which we report in the remainder of this section. First, collect relevant data. Second, annotate the data to create a training dataset. Third, extract input features. Fourth, identify the optimum algorithm. Fifth, test and validate the results.

Data collection

Given that the spread of Islamophobia within UK political parties is a key concern within current civic and political discourse, we train our classifier on a dataset which is situated in this domain. To create the dataset we follow four steps. First, we create a list of 50,000 Twitter users, each of which follows at least one out of six prominent UK political parties. These parties comprise a diverse ideological mix, including mainstream parties with nationwide support bases and broadly liberal policies (the Conservatives, the Liberal Democrats, and Labor), a right-wing populist party, known for articulating a stridently anti-EU message (UKIP), and two far-right parties (the BNP and Britain First) (Ford & Goodwin, 2014; Golder, 2016; John & Margetts, 2009). Given Islamophobia's close association with far-right politics, we ensure greater representation of tweets from far-right actors by also including 45 high-profile far-right accounts in our list of 50,000 users. These 45 far-right accounts are identified from reports by Hope Not Hate.[2] Second, we collect all of the tweets that these 50,000 accounts produced between January 2017 and June 2018 using Twitter's "Search" API. This creates a dataset with 140 million tweets. Third, we sampled from the 140 million tweets to create a training dataset of 4,000 tweets, stratifying the sample to include tweets from followers of each of the political parties. We mitigate the impact of bots and anonymous online users (who may have unusual tweeting patterns) by limiting the maximum number of tweets that each user can contribute to the dataset to just three.

Creating a training dataset with sufficient instances of hateful content is a time-consuming endeavor, not least because in most online contexts the prevalence of hate is relatively low overall (Schmidt & Wiegand, 2017). To ameliorate this problem, Waseem and Hovy recommend increasing the prevalence of hate speech by sampling data which is more likely to be relevant (Waseem & Hovy, 2016). This approach was partially adopted here; of the 4,000 tweets in the sample, 1,000 are identified using keyword searches for "Muslims" and "Islam".

Our sample is highly heterogeneous: it includes tweets from a large number of users which follow a diverse range of political accounts, stratified to ensure coverage over a full year of time. We have sought to maximize heterogeneity in order to increase the applicability of our classifier in different "real world" settings. This also reduces the risk of overfitting by ensuring a more diverse range of linguistic styles are included in the dataset (Waseem & Hovy, 2016). However, we caution that our dataset is not a "representative" sample of political actors online as we have optimized our classifier's performance by focusing on settings which previous research indicates are likely to contain Islamophobic hate (such as far-right actors). As such, our classifier may still not be applicable in all contexts.

Data annotation

All 4,000 tweets in the training dataset were annotated blind by three human annotators who are experts in UK politics and the study of prejudice. Based on the definitions of Islamophobia offered above we created guidelines for the annotators, and expanded them through two preliminary studies, each consisting of 200 tweets. Across the 4,000 tweets, inter-rater agreement was high. Percentage agreement is 89.9%, Fleiss' kappa is 0.837, and Krippendorf's alpha is 0.895. We also compute category-wise scores for Fleiss' kappa, which range from 0.737 for weak Islamophobia to 0.907 for strong Islamophobia. The consistency of these results shows the internal validity of the annotation guidelines.

In cases where annotators disagree, tweets are assigned to classes based on the majority decision. In the final dataset, 3,106 tweets are classed as non-Islamophobic, 484 tweets are classed as weak Islamophobic, 410 tweets are classed as strong Islamophobic. To create an evenly balanced dataset, the number of non-Islamophobic tweets is reduced to 470 instances through random removal. This creates a final training dataset of 1,364 tweets.

Feature selection

Feature selection refers to the choice of input variables used to train the classifier. For comparison, we first create a baseline model which uses only the count of each term in each tweet as an input. Second, we create a model using 50 surface-level and derived additional features. These include sentiment and polarity, count of swear words (Ipsos, 2016), parts of speech and named entities. We also derive two new input features: (i) mentions of Muslim names and (ii) mentions of the names of Mosques, both of which are taken from relevant Wikipedia pages. Third, we create a combined model that uses both one-hot encodings and all 50 of the non-text features. Fourth, we create a model using pre-trained gloVe word embeddings, trained on two billion tweets (Stanford, 2018). Fifth, we create a gloVe model using newly-trained word embeddings on the corpus of our own 140 million tweets, using the method described by Pennington et al. (Pennington, Socher, & Manning, 2014). Finally, sixth, we create a model which uses the newly-trained word embeddings as well as all 50 of the non-text features. For testing we implement 10-fold cross-validation on the annotated dataset, using the Naïve Bayes algorithm. The results are shown in Table 1. We measure performance with Accuracy, which is the percentage of tweets which are assigned to the correct class.

Table 1. Accuracy of models for classification

Input feature model	Accuracy
BASELINE MODEL: Text (Counts of each term)	30.07%
Model 1: 50 additional features	49.96%
Model 2: Text (Counts of each term0) + 50 additional features	30.36%
Model 3: Pre-trained word embeddings	63.20%
Model 4: Newly trained word embeddings (setting-specific)	69.13%
Model 5: Newly trained word embeddings (setting-specific) + 50 additional features	65.20%

The baseline model, which uses just the count of each term as an input, has an accuracy of only 30.7%, which is worse than either a random assignment (which would have 33.49% accuracy) or a zero-rule assignment (which would have 36.09% accuracy). All five models outperform it. This result shows the difficulty of extracting relevant signals to classify Islamophobic hate speech and justifies our choice of a more complex algorithmic architecture. The best performing model is the newly trained word embeddings alone (model 5). Interestingly, this model considerably outperforms the accuracy of the pre-trained word embeddings model (5.9 percentage points, 69.13% compared with 63.2%). This suggests that the benefits of training the word embeddings on tweets which are setting-specific outweighs the cost of having a smaller dataset. This is in line with previous work on word embeddings, such as Lai et al., who report that "corpus domain is more important than corpus size." (Lai, Liu, He, & Zhao, 2016). We create a final model (model 7) in which the newly trained word embeddings used in model 5 are optimized by including some of the additional features. Through testing every combination of additional features, we identify six which maximize accuracy in model 7:

> *Word embeddings + count of mentions of Mosques + presence of HTML + presence of RT + part of speech: 'conjunction' + named entity recognition: 'location' + named entity recognition: 'organization'*

Model 7 includes only one of the two new features which we engineered (the count of mentions of Mosques), and not the use of Muslim names. This could reflect the fact that during the period we collected data, several mosques became the target of anti-Muslim sentiment, including several "invasions" by far-right activists and a terrorist attack. In our dataset, tweets which contain mentions of Mosques are 5 times more likely to be classified as Islamophobic (either weak or strong). The presence of HTML and Retweets is also associated with the tweet being more likely to be classified as Islamophobic, which could reflect the multimedia and viral nature of negative social media content. Two named entities, location and organization, are included in the final model. Qualitative analysis shows that, in many cases, the identified organizations and locations relate to either the Islamic faith or far-right political leaders and parties. As such, these features' inclusion within the model can be theoretically justified. This is important as it means that the classifications produced by our classifier can be limitedly "explained". The inclusion of "Conjunction" is less theoretically justified, and more post-hoc analysis is needed in future work to understand why this feature is associated with hate. One avenue is to explore the grammatical structures used in hateful tweets, given previous research which shows that this can be leveraged for abusive content detection (*Alorainy, Burnap, Liu, & Williams, 2018*).

Choice of algorithm

For simplicity, we test the algorithms on model 5 (which consists of only the newly trained word embeddings). We test six different algorithms, which are selected based on previous research (Wainer, 2016): Naïve-Bayes, Random Forests (with trees = 10, 100 and 1,000), Logistic Regression, Decision Trees, Support Vector Machines (SVM), and Deep Learning. We implement a multi-class SVM with a one-against-one strategy (Hsu & Lin, 2002). We optimize the hyperparameters of the SVM classifier with a "radial" kernel. "C" is 2 and gamma is 0.01. We also optimize the Deep Learning model, testing for the activation function, optimization function, learning rate, and a number of epochs. The results, including optimized hyperparameters, are shown in Table 2.

All six algorithms perform well, with accuracy ranging from 61.23% to 72.17%. The two highest performing are SVM and Deep Learning (using only a "shallow" Multi-layer perceptron architecture) – the accuracy of SVM is 72.17%, which outperforms Deep Learning by 1.03 percentage points. Thus, contrary to our initial expectations, we use SVM for the classifier. This is in-line with work by MacAveney et al., who also find that SVM outperforms deep learning for hate speech detection (MacAvaney et al., 2019). The SVM hyperparameters are set to maximize generalizability (i.e. low "C" and gamma values), which makes the classifier highly suitable for applying in different empirical contexts.

Table 2. Results of algorithm testing

Algorithm	Accuracy
Naïve-Bayes	69.13%
Random Forests (trees = 10)	65.40%
Random Forests (trees = 100)	68.72%
Random Forests (trees = 1000)	67.94%
Logistic Regression	69.13%
Decision Trees	61.23%
SVM with kernel = 'radial' + 'C' = 2 + gamma = 0.01	72.17%
Deep Learning with epochs = 100 + activation function = 'relu' + optimization function = rmsprop, learning rate = 0.001	71.14%

Evaluation of performance and limitations

The final classifier consists of model 7 implemented with a tuned SVM. To evaluate performance, we cross-validate the classifier on the training data set (n = 1,341 tweets) using a ten-fold classification. The results are shown in Table 3.

For the accuracy, recall and precision and F1 scores, we use the macro-aggregation strategy described by Sokolova and Lapalme, in which values are calculated for each class and then the per-class agreement is averaged, with each class treated equally (Sokolova & Lapalme, 2009). The classifier performs similarly for recall and precision (0.741 and 0.739, respectively), and as such has a comparable F1 score (0.74). Balanced accuracy is 0.807. These results show that the classifier

Table 3. Performance of classifier over 10 folds

Fold	Accuracy	Balanced accuracy	Precision	Recall	F1 score
1	0.796	0.846	0.795	0.798	0.797
2	0.76	0.808	0.75	0.736	0.743
3	0.736	0.808	0.74	0.75	0.745
4	0.721	0.792	0.714	0.724	0.719
5	0.718	0.774	0.686	0.685	0.686
6	0.746	0.808	0.74	0.742	0.741
7	0.702	0.785	0.699	0.721	0.71
8	0.79	0.845	0.793	0.793	0.793
9	0.756	0.809	0.736	0.736	0.736
10	0.735	0.798	0.733	0.729	0.731
Mean	0.746	0.807	0.739	0.741	0.740

does well at balancing the need to identify relevant instances with minimizing misclassifications.

One hundred tweets are randomly sampled from tweets assigned to each of the three classes (non-, weak and strong Islamophobia) to create a dataset of 300 automatically classified unseen tweets. The same three annotators annotate these 300. As before, we take the majority decision to decide the annotation (in 95% of cases all three annotators are in agreement). The result performance of the classifier on this task is shown in Table 4.

The classifier performs better across all metrics on the unseen 300 tweets, with an accuracy of 77.3%. The uplift in performance and consistency of the results indicate the robustness of our approach and its generalizability, which is most likely due to our selection of theoretically informed input features. Importantly, these results suggest that the classifier is suitable for implementation in empirical research as precision is well above the 70% minimum recommended by van Rijsbergen (van Rijsbergen, 1979). The classifications of the classifier are shown in Table 5.

The classifier performs well at distinguishing non-Islamophobic from strong Islamophobic tweets. However, it struggles with distinguishing weak from both strong and non-Islamophobic tweets. For instance, out of 100 tweets which are labeled as strong Islamophobic, 23 are actually weak. Similarly, out of 100 tweets which are labeled as weak Islamophobic, 22 are actually non-Islamophobic.

Qualitative investigation of the unseen 300 tweets shows that, in many cases, the misclassified non-Islamophobic tweets express hatred and prejudice against groups other than Muslims, such as immigrants. This means that they share many of the same signals, such as use of offensive terminology and derogative statements, such as "Bloody …" or declaratives, such as "I hate …". Some also discuss Muslims and Islamic practices but without expressing any negativity. In these cases, the weak and non-Islamophobic tweets will share similar signals, such as terms relating to Mosques, Muslims and prayer. In effect, the middle category of weak Islamophobia has to contend with the greatest within-class variation and the most mixed signals, making it harder to separate from the other two. This work reflects similar challenges in prior computational work, such as the inability of most classifiers to handle instances of irony and humor or to explicitly model the role of users and social context (Schmidt & Wiegand, 2017). In future work, recent innovations in natural language could be used to address these classification errors through *contextual* word embeddings. Contextual word embeddings create a unique vectoral representation for each unique context in which each term is used. For instance, use of the term "Muslim" in an aggressive setting, such as "F**king Muslims …", has a different vector representation to, and as such can be separated from, use of the term "Muslim" in a more positive or neutral setting, such as "Muslims contribute to society …". The unique vectors for each term can either be used as input features, clustered to identify broad semantic differences (which can then be used as an input) or used as one input layer in conjunction with a non-contextual embedding, such as the GloVe model used here (Devlin, Chang, Lee, & Toutanova, 2018). Implementing such models can risk overfitting if used with small datasets, and as such a larger volume of data would need to be annotated than in the present work.

Table 4. Performance of classifier on unseen data

	Accuracy	Balanced accuracy	Precision	Recall	F1 score
Results on 300 unseen tweets	0.773	0.83	0.778	0.773	0.776
Difference with prior testing	+ 0.027	+ 0.023	+ 0.039	+ 0.032	+ 0.036

Table 5. Contingency table for classifier performance on unseen data

		Predicted Islamophobia			
		None	Weak	Strong	
Actual	None	91	22	4	117
	Weak	8	68	23	99
	Strong	1	10	73	84
		100	100	100	**300**

Application

To demonstrate the potential applications of our classifier for social scientific research, we apply it to an unseen dataset of 73,311 tweets produced by 45 far-right Twitter accounts during 2017. The results of this are shown below in Figure 1.

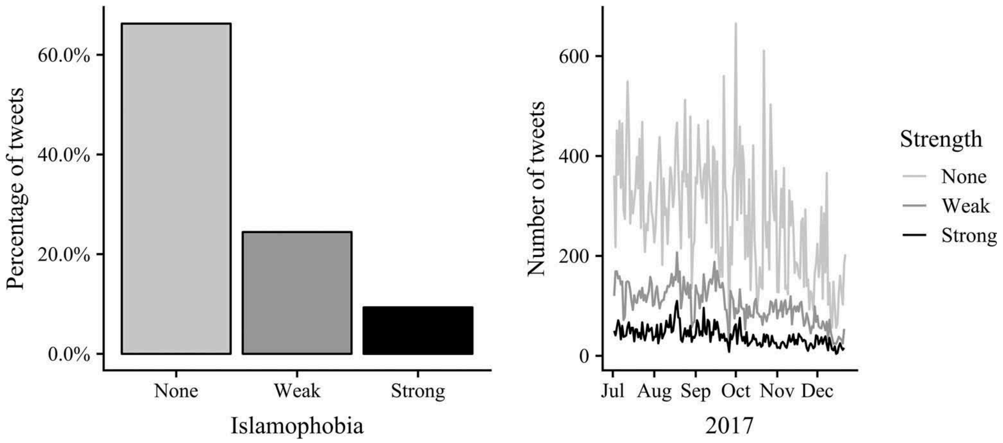

Figure 1. Panel A shows the number of tweets and Panel B shows the prevalence of tweets over time.

Conclusion

In this paper, we have sought to realize the research aim outlined at the start: *To create a classifier for detecting Islamophobic content on social media which distinguishes between different strengths of Islamophobia*. The multi-class Islamophobic hate speech classifier we have described marks a step forward in developing quantitative and theoretically informed tools to research Islamophobia on social media and meets this research aim. It can be used for large-scale quantitative analyses of online hate to investigate new topics and answer open questions in the extant literature, such as understanding the temporal dynamics of hate, as briefly shown in Section 3.6. The classification pipeline we have presented is also relevant for both (i) studying Islamophobic content from other social media platforms, such as Facebook or Gab and (ii) classifying and studying other forms of online hate, such as misogyny, racism, and anti-Semitism. However, we also caution that more work needs to be undertaken, particularly in making nuanced distinctions between the different strengths, and that hate detection is an ongoing area of research which will need to be constantly revisited as the nature of online abuse changes. This work also feeds into wider debates in the social sciences regarding the use of automated computational tools by providing a working example of a tool which is tailored toward a nuanced and challenging task. We welcome future work which further develops the tool described here.

Notes

1. Available at: https://zenodo.org/record/3463560#.XY5LKC2ZOu5.
2. This list is also available in our online supplement.

Declaration of Interest

The authors declare that there are no conflicts of interest.

Funding

This research is funded by an Award from the University of Oxford ESRC Social Sciences Doctoral Training Centre, Award number [ES/J500112/1].

Data availability statement

The authors confirm that the data supporting the findings of this study are available within the article's supplementary materials and in an online repository. The URL is: https://zenodo.org/record/3463560#.XY5LKC2ZOu5

The DOI is: 10.5281/zenodo.3463560

Notes on contributors

Bertie Vidgen is a Researcher at the Alan Turing Institute, a Visiting Researcher at the Oxford Internet Institute and Visiting Researcher at the Open University. He completed his PhD at the University of Oxford.

Taha Yasseri is a Senior Research Fellow in Computational Social Science at the OII, a Turing Fellow at the Alan Turing Institute for Data Science, and a Research Fellow in Humanities and Social Sciences at Wolfson College, University of Oxford.

ORCID

Bertie Vidgen http://orcid.org/0000-0002-7892-0814

References

Ad-Dab'bagh, Y. (2017). A twenty-first century scourge: Introducing the special issue on Islamophobia. *International Journal of Applied Psychoanalysis Studies*, *14* (1), 167–172. doi:10.1002/aps.1529

Akgönül, S., Alibašić, A., Nielsen, J., & Račius, E. (Ed.) (2018). *Yearbook of Muslims in Europe, Volume 9*. Boston: Brill. doi:10.1080/09596410.2012.71244

All Party Parliamentary Group on British Muslims. (2018). *Islamophobia defined: The inquiry into a working definition of Islamophobia*. London, UK: The House of Commons.

Allen, B. (1996). Feminist standpoint theory: A black woman's (re)view of organizational socialization. *Communication Studies*, *47*(4), 257–271. doi:10.1080/10510979609368482

Allen, C. (2011). *Islamophobia*. Surrey, UK: Ashgate.

Allen, C. (2017). Political approaches to tackling Islamophobia: An 'Insider/Outsider' analysis of the British coalition government's approach between 2010–15. *Social Sciences*, *6*(77), 1–19. doi:10.3390/socsci6030077

Alorainy, W., Burnap, P., Liu, H. A. N., & Williams, M. L. (2018). The enemy among Us: Detecting hate speech with threats based othering language embeddings,*ACM transactions on the Web, 13*(3), 1–26.

Awan, I. (2016). Islamophobia on social media: A qualitative analysis of facebook's walls of hate. *International Journal of Cyber Criminology*, *10*(1), 1–20. doi:10.5281/zenodo.58517

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behaviour*, *27*(1), 1–8. doi:10.1016/j.avb.2016.02.001

Bakali, N. (2019). *The far right's love affair with Islamophobia*. Texas, USA: Yaqeen Institute for Islamic Research.

Bayrakli, E., & Hafez, F. (2018). *European Islamophobia Report 2017*. Istanbul, Turkey: SETA.

Blei, D. M., & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences*, *114*(33), 8689–8692. doi:10.1073/pnas.1702076114

Bleich, E. (2011). What is Islamophobia and how much is there? theorizing and measuring an emerging comparative concept. *American Behavioral Scientist*, *55*(12), 1581–1600. doi:10.1177/0002764211409387

Bliuc, A. M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, *87*(1), 75–86. doi:10.1016/j.chb.2018.05.026

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15* (5), 662–679. doi:10.1080/1369118X.2012.678878

Brown, R. (2010). *Prejudice: Its social psychology*. Sussex, UK: Wiley-Blackwell.

Bulmer, M., & Solomos, J. (2015). *Multiculturalism, Social Cohesion and Immigration*. (M. Bulmer & J. Solomos, Eds.). Abingdon, UK: Routledge.

Burnap, P., & Williams, M. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 1–15. doi:10.1140/epjds/s13688-016-0072-6

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet*, *7*(2), 223–242. doi:10.1002/poi3.85

Busher, J., & Macklin, G. (2014). Interpreting "Cumulative Extremism": Six proposals for enhancing conceptual clarity. *Terrorism and Political Violence*, *27*(5), 884–905. doi:10.1080/09546553.2013.870556

Chakraborti, N., & Garland, J. (2012). Reconceptualising hate crime victimisation through the lens of vulnerability and 'difference.'. *Theoretical Criminology*, *16*(4), 499–514.

Cihon, P., & Yasseri, T. (2016). A biased review of biases in Twitter studies on political collective action. *Frontiers in Physics*, *4*(1), 1–10. doi:10.3389/fphy.2016.00034

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., … Helbing, D. (2012). Manifesto of computational social science. *European Physical Journal: Special Topics*, *214* (1), 325–346. doi:10.1140/epjst/e2012-01697-8

Cowls, J., & Schroeder, R. (2015). Causation, correlation, and big data in social science research. *Policy & Internet*, *7*(4), 447–472. doi:10.1002/poi3.100

Demos. (2017). *Anti-Islamic hate on Twitter*. London, UK: Demos.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv:1810.04805v2*, 1–16. Retrieved from http://arxiv.org/abs/1810.04805

Ford, R., & Goodwin, M. (2014). *Revolt on the right: Explaining support for the radical right in Britain*. London, UK: Routledge. doi:10.4324/9781315859057

Gallie, W. B. (1956). Essentially contested concepts. *Proceedings of the Aristotelian Society*, *56*, 167–198.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation and the hidden decisions that shape social media*. Yale: Yale University Press.

Golder, M. (2016). Far right parties in Europe. *Annual Review of Sociology*, *19*(1), 477–497. doi:10.1146/annurev-polisci-042814-012441

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 1–31. doi:10.1093/pan/mps028

Haque, A., Tubbs, C. Y., Kahumoku-Fessler, E. P., & Brown, M. D. (2018). Microaggressions and Islamophobia: Experiences of Muslims across the United States and clinical implications. *Journal of Marital and Family Therapy*, *44*, 2.

Hope Not Hate. (2017). *Hope Not Hate: State of Hate 2017*. London, UK: Hope Not Hate. .

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425. doi:10.1109/72.991427

Imhoff, R., & Recker, J. (2012). Differentiating Islamophobia: Introducing a new scale to measure Islamoprejudice and secular Islam critique. *Political Psychology*, *33*(6), 811–824. doi:10.1111/j.1467-9221.2012.00911.x

Ingham-Barrow, I. (2018). *More than words: Approaching a definition of Islamophobia.* . London, UK: MEND.

Ipsos MORI. (2016). *Attitudes to potentially offensive language on TV and radio*. London, UK : Ipsos MORI.

John, P., & Margetts, H. (2009). The latent support for the extreme right in British politics. *West European Politics*, *32* (3), 496–513. doi:10.1080/01402380902779063

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, *13*(1), 72–91. doi:10.1080/19331681.2015.1132401

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying*, Santa Fe, USA (pp. 1–11).

Kunst, J. R., Sam, D. L., & Ulleberg, P. (2013). Perceived Islamophobia: Scale development and validation. *International Journal of Intercultural Relations*, *37*(2), 225–237. doi:10.1016/j.ijintrel.2012.11.001

Lai, S., Liu, K., He, S., & Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems*, *31*(6), 5–14. doi:10.1109/MIS.2016.45

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., … Alstyne, M. V. (2009). Computational social science. *Science*, *323*(February), 721–724.

Leader Maynard, J., & Benesch, S. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, *9*(3), 70–95. doi:10.5038/1911-9933.9.3.1317

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *Plos One*, *14*(8), 1–16. doi:10.1371/journal.pone.0221152

Malmasi, S., & Zampieri, M. (2017, December). *Detecting hate speech in social media*. Retrieved from http://arxiv.org/abs/1712.06427

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the Digital Humanities*, *1*(1), 1–17. http://www.manovich.net/DOCS/Manovich_trending_paper.pdf

Margetts, H. (2017). Political behaviour and the acoustics of social media. *Nature Human Behaviour*, *1*(86), 1–3. doi:10.1038/s41562-017-0086

Meer, N., & Modood, T. (2009). Refutations of racism in the 'Muslim question'. *Patterns of Prejudice*, *43*(3), 335–354. doi:10.1080/00313220903109250

Mondon, A., & Winter, A. (2017). Articulations of Islamophobia: From the extreme to the mainstream? *Ethnic and Racial Studies*, *40*, 2151–2179. doi:10.1080/01419870.2017.1312008

Moosavi, L. (2015). The racialization of Muslim converts in Britain and their experiences of Islamophobia. *Critical Sociology*, *41*(1), 41–56. doi:10.1177/0896920513504601

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha,Qatar, (pp. 1532–1543).

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and Blatant Prejudice in Western Europe. *European Journal of Social Psychology*, *25*(1), 57–75. doi:10.1002/ejsp.2420250106

Pettigrew, T. F., Tropp, L. R., Wagner, U., & Christ, O. (2011). Recent advances in intergroup contact theory. *International Journal of Intercultural Relations*, *35*(3), 271–280. doi:10.1016/j.ijintrel.2011.03.001

Poynting, S., & Briskman, L. (2018). Islamophobia in Australia : From Far-Right Deplorables to Respectable Liberals. 1–17. doi:10.3390/socsci7110213

Rosenfeld, M. (2012). Hate speech in constitutional jurisprudence: A comparative analysis. In M. Herz & P. Molnar (Eds.), *The content and context of hate speech: Rethinking regulation and responses*. Cambridge: Cambridge University Press, (pp. 242–289).

Runnymede Trust. (2017). *Islamophobia: Still a challenge for us all*. London, UK: Runnymede Trust.

Salminen, J., Almerekhi, H., Milenković, M., Jung, S.-G., An, J., Kwak, H., & Jansen, B. J. (2018). *Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media*. Proceedings of the Twelfth International AAAI Conference on Web and Social Media (pp. 330–339). Retrieved from www.aaai.org

Salminen, J., Veronesi, F., Almerekhi, H., Jung, S., & Jansen, B. J. (2018). *Online hate interpretation varies by country, but more by individual*. Proceedings of SNAMS (pp. 1–7). doi:10.1109/SNAMS.2018.8554954

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *International Workshop on NLP for Social Media* (pp. 1–10). Valencia, Spain. doi:10.18653/v1/w17-1101

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. doi:10.1016/j.ipm.2009.03.002

Stanford. (2018).*GloVe,* Stanford University. Available at: https://nlp.stanford.edu/projects/glove/.

Twitter. (2018).*Hateful conduct policy*, Twitter. Available at: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.

Vidgen, B., Tromble, R., Harris, A., Hale, S., Nguyen, D., & Margetts, H. (2019). Challenges and frontiers in abusive content detection in *Proceedings of the ThirdWorkshop on Abusive Language Online*, Florence, Italy, (pp. 80–93).

Wainer, J. (2016). Comparison of 14 different families of classification algorithms on 115 binary datasets. *ArXiv preprint:1606.00930v1*, (2014), pp. 1–36.

Walia, K., Khan, S., & Islam, N. (2019). *Terrorism, hate crimes and western politics: Islamophobia in the context of globalization and the media*. Istanbul, Turkey):INSAMER Humanitarian and Social Research Center.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). *Understanding abuse: A typology of abusive language detection subtasks*. 1st Workshop on Abusive Language Online (pp. 78–84). doi:10.1080/17421770903114687

Waseem, Z., & Hovy, D. (2016). *Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. NAACL-HLT* (pp. 88–93). doi:10.18653/v1/n16-2013

Watts, D. J. (2007). A twenty-first century science. *Nature*, *445*(February), 2007.

Williams, M., & Burnap, P. (2016). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, *56*(1), 211–238. doi:10.1093/bjc/azv059

Williams, M., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 1–20. doi:10.1177/0038038517708140