TOPLINES: ANTI-ISLAMIC HATE ON TWITTER

This document contains the top-line results of independent research conducted by the Centre for the Analysis of Social Media at Demos. The research has measured the volume of messages on Twitter algorithmically considered to be derogatory towards Muslims over a year, from March 2016 to March 2017. This is part of a broad effort to understand the scale, scope and nature of uses of social media that are possibly socially problematic and damaging.

(1) TWITTER IS USED FOR ALL KINDS OF BENEFICIAL ACTIVITIES. BUT A SMALL MINORITY OF UK TWEETS ACROSS THE YEAR WERE IDENTIFIED AS CONTAINING LANGUAGE THAT IS DEROGATORY TOWARDS MUSLIMS

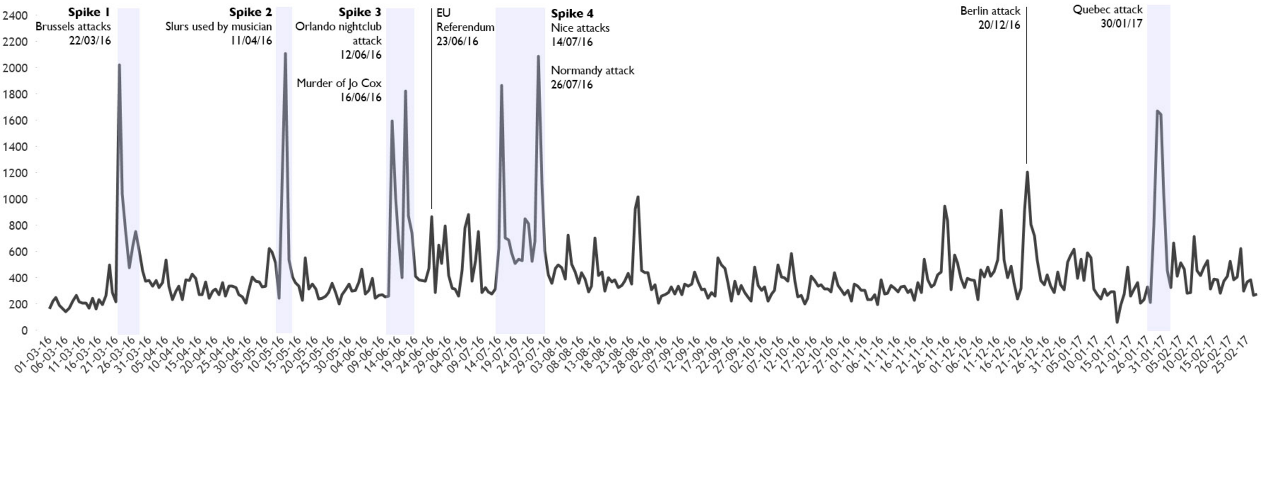
- Over a year, researchers detected 143,920 Tweets sent from the UK considered to be derogatory and anti-Islamic.
- This is about 393 a day.
- These Tweets were sent from over 47,000 different users.

(2) THESE TWEETS FELL INTO A NUMBER OF DIFFERENT CATEGORIES, FROM DIRECTED, ANGRY INSULTS TO BROADER POLITICAL STATEMENTS.

- A random sample of hateful Tweets were manually classified into three broad categories:
  - **'Insult' (just under half):** Tweets used an anti-Islamic slur in a derogatory way, often directed at a specific individual.
  - **'Muslims are terrorists'(around one fifth)** Derogatory statements generally associating Muslims and Islam with terrorism.
  - **'Muslims are the enemy' (just under two fifths):** Statements claiming that Muslims, generally, are dedicated toward the cultural and social destruction of the West.

(3) KEY EVENTS, ESPECIALLY TERRORIST ATTACKS, DRIVE LARGE INCREASES IN THE VOLUME OF MESSAGES ON TWITTER CONTAINING THIS LANGUAGE

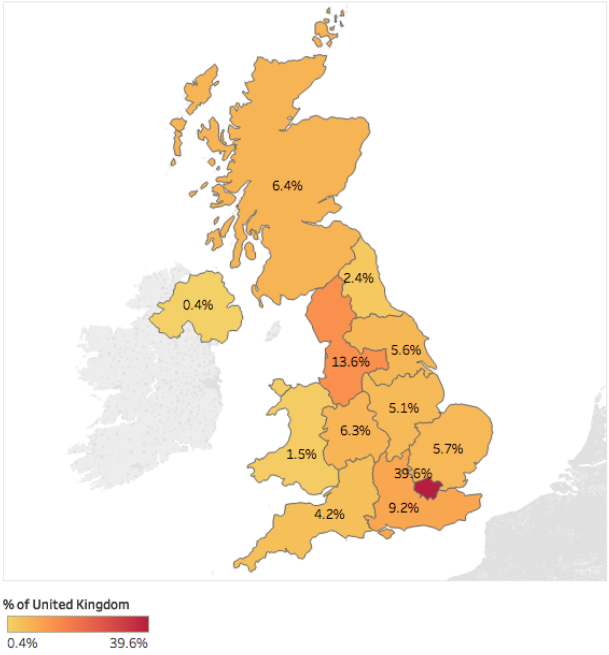
The Brussels, Orlando, Nice, Normandy, Berlin and Quebec attacks all caused large increases. There was a period of heightened activity over Brexit, and sometimes online 'Twitter storms' (such as the use of derogatory slurs by Azealia Banks toward Zayn Malik) also drove sharp increases.



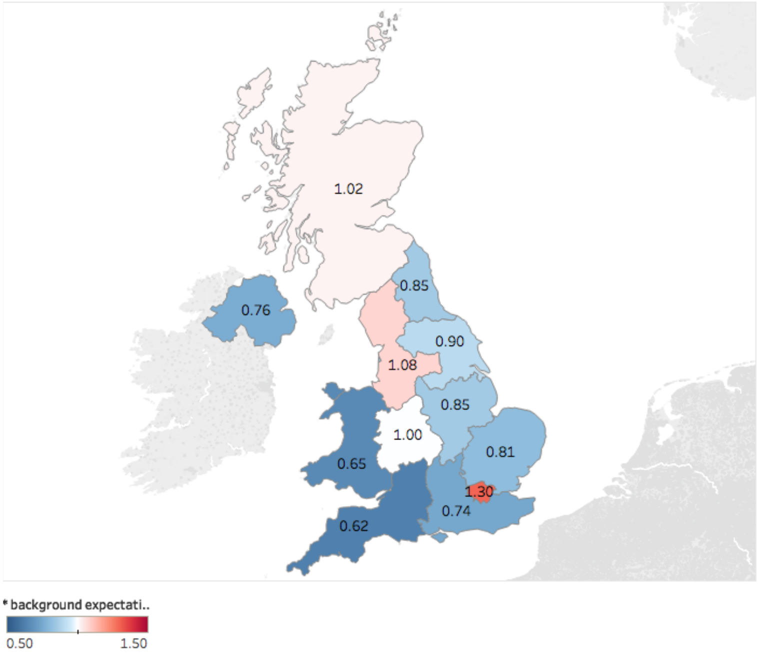
(4) TWEETS CONTAINING THIS LANGUAGE WERE SENT FROM EVERY REGION OF THE UK, BUT THE MOST OVER-REPRESENTED, COMPARED TO GENERAL TWITTER ACTIVITY, WERE LONDON AND THE NORTH WEST.

Of the 143,920 Tweets containing this language and classified as being sent from within the UK, 69,674 (48%) contained sufficient information to be located within a broad area of the UK. To measure how many Tweets each region generally sends, a random baseline of 67 Million Tweets were collected over 19 days over late February and early March. The volume of Tweets containing derogatory language towards Muslims was compared to this baseline. This identified regions where the volume was higher or lower than the expectation on the basis of general activity on Twitter.

Anti-Islamic: Tweet Volume (%) - NUTS 1



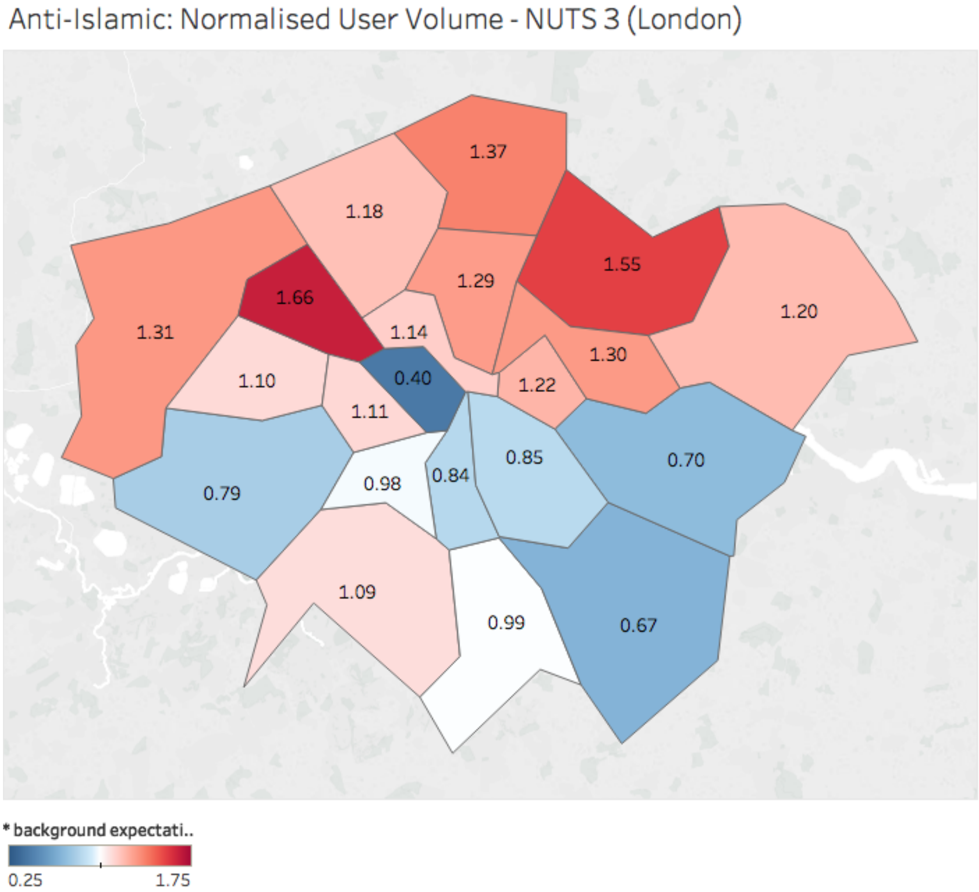
Anti-Islamic Normalised User Volume - NUTS 1



(5) NORTH LONDON SENT MARKEDLY MORE TWEETS CONTAINING LANGUAGE CONSIDERED DEROGATORY TOWARDS MUSLIMS THAN SOUTH LONDON (COMPARED TO GENERAL TWITTER ACTIVITY)

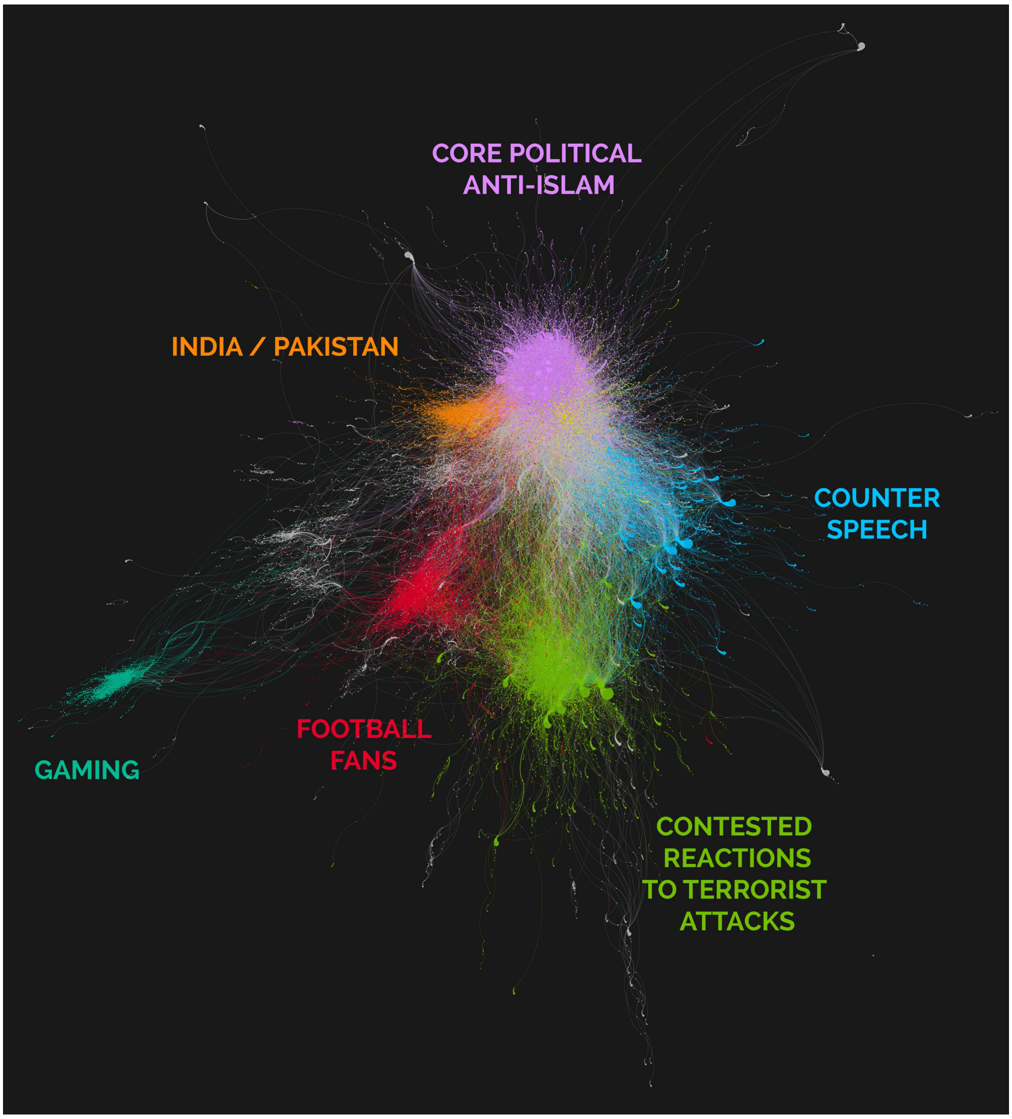
27,576 (39%) were sent from Greater London. Of these, 14,953 Tweets (about half) could be located to a more specific region within London (called a 'NUTS-3 region'; typically either a London Borough or a combination of a small number of London Boroughs).[1]

- Brent, Redbridge and Waltham Forest sent the highest number of derogatory, anti-Islamic Tweets relative to their baseline average of general Twitter activity.
- Westminster and Bromley sent the least number of derogatory, anti-Islamic Tweets relative to their baseline average of general Twitter activity.



[1] An important caveat is that the volumes associated with each of these regions are obviously smaller than the total number of Tweets in the dataset overall

(6) SIX DIFFERENT ONLINE TRIBES WERE IDENTIFIED[2]


CORE POLITICAL ANTI-ISLAM
INDIA / PAKISTAN
COUNTER SPEECH
GAMING
FOOTBALL FANS
CONTESTED REACTIONS TO TERRORIST ATTACKS

These were:

Core political anti-Islam. The largest group of about 64,000 users including recipients of Tweets. Politically active group engaged with international politics.

- Hashtags employed by this group suggest engagement in anti-Islam and right wing political conversations: (#maga #tcot #auspol #banIslam #stopIslam #rapefugees)

[2] A caveat here is that this network graph includes Tweets that are misclassified and also includes the recipients of abuse. It is also important to note that not everyone who shares Tweets does so with malicious intent; they can be doing so to highlight the abuse to their own followers.

- In aggregate, words in user descriptions emphasise nationality, right-wing political interest and hostility towards Islam (anti, Islam, Brexit, UKIP, proud, country)

Contested reactions to Terrorist attacks. The second largest group, of about 18,000 users, including recipients of tweets.

- Aggregate overview of user descriptions imply a relatively young group (sc, snapchat, ig, instagram, 17,18,19,20, 21)

- User descriptions also imply a mix of political opinion (blacklivesmatter, whitelivesmatter, freepalestine)

- Hashtags engage in conversations emerging in the aftermath of terrorist attacks (#prayforlondon, #munich, #prayforitaly, #prayforistabul, #prayformadinah, #orlando)

- Likewise, hashtags are a mix of pro- and anti-Islamic (#britainfirst, #whitelivesmatter, #stopislam, #postrefracism, #humanity)

The counter-speechers. A group of 8,700 people; although of course the data collection, by design, only detected the part of the counter-speech conversation containing language that can be used in a way derogatory towards Muslims. It is therefore likely that it did not collect the majority of counter-speech activity.[3]

The shape of the cluster shows a smaller number of highly responded to-/retweeted comments.

- Hashtags engage predominantly with anti-racist conversations (#racisttrump, postrefracism, #refugeeswelcome, #racism, #islamophobia)

- In aggregate, user descriptions show mix of political engagement and general identification with left-wing politics (politics, feminist, socialist, Labour).

- Overall they also show more descriptions of employment than the other clusters (writer, author, journalist, artist).

The Football Fans. 7,530 users are in this cluster, including recipients of Tweets.

- The bio descriptions of users within his cluster overwhelmingly contain football-related words (fan, football, fc, lfc, united, liverpool, arsenal, support, club, manchester, mufc, chelsea, manutd, westham)

- No coherent use of hashtags. This cluster engaged in lots of different conversations.

India/Pakistan. Just under 5,000 users are in this cluster (including recipients).

- Hashtags overwhelmingly engage in conversation to do with India-Pakistan relations or just Pakistan (#kashmir, #surgicalstrike, #pakistan, #actagainstpak).

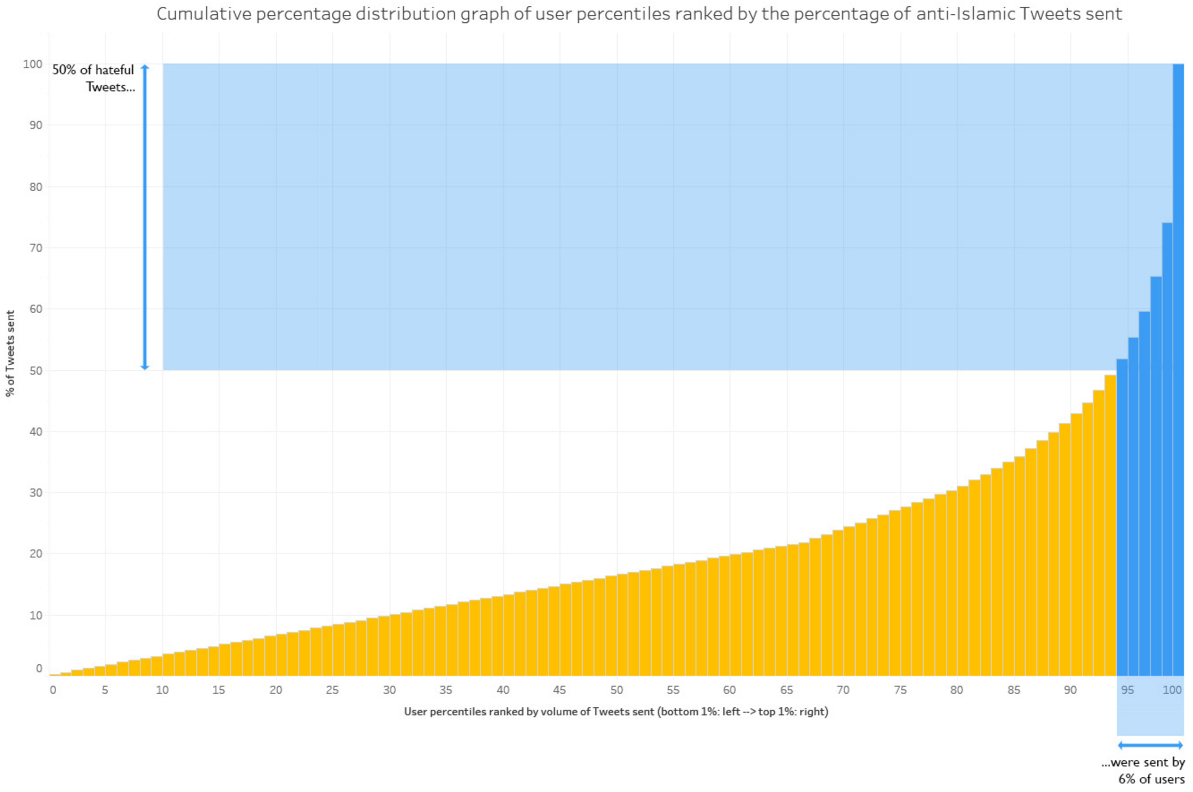
- In aggregate, words in user descriptions relate to Indian/nationalist identity and pro-Modi identification (proud, Indian, hindu, proud indian, nationalist, dharma, proud hindu, bhakt,)

[3] In other work on the subject we have found there are usually more posts about solidarity, support for Muslims than attacks on them.

The Gamers. 2,813 users are in this cluster (including Tweet recipients).

- There is no coherent use of hashtags.
- Overall, aggregate comments in user descriptions either imply young age (16,17,18) or are related to gaming (player, cod [for 'Call of Duty'], psn)

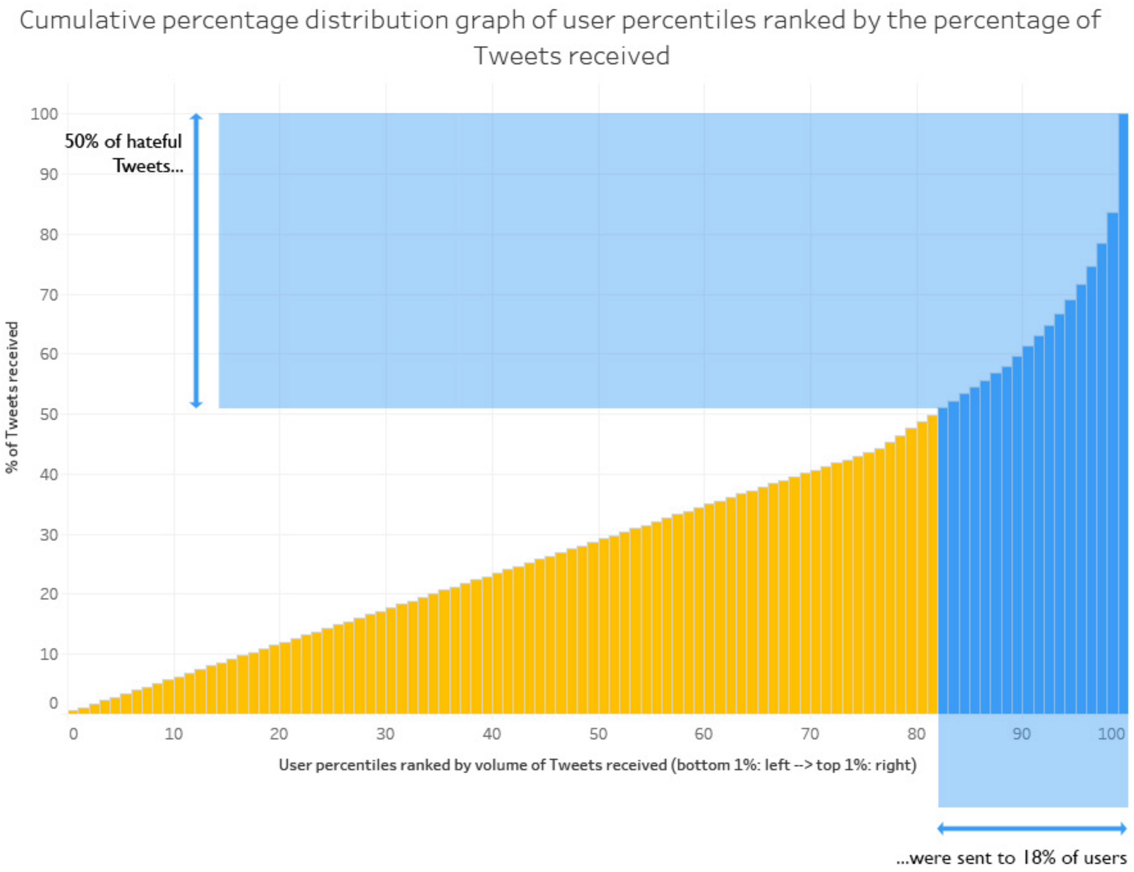
(7) A SMALL NUMBER OF ACCOUNTS ARE RESPONSIBLE FOR MANY OF THE TWEETS CONTAINING LANGUAGE GENERALLY CONSIDERED DEROGATORY TOWARDS MUSLIMS



Similar to many other kinds of online life, a small number of accounts create many of the tweets containing this language.

- 50% of Tweets classified as containing language considered anti-Islamic and derogatory are sent by only 6% of accounts
- 25% of Tweets classified as containing language considered anti-Islamic and derogatory were sent by 1% of accounts.

(8) LIKEWISE, A SMALL NUMBER OF ACCOUNTS RECEIVED A LARGE AMOUNT OF THE DEROGATORY, ANTI-ISLAMIC ACTIVITY THAT WAS DIRECTED AT A PARTICULAR PERSON



ETHICS

At Demos we believe it is important that the principle of internet freedom should be maintained; and that it should be a place where people feel they can speak their mind openly and freely. However, racist, xenophobic, and derogatory anti-Islamic abuse can curtail freedom, and the capacity to speak and act freely online, as much as it can be an expression of these values. It is important, as society confronts the ways that social media acts as a new platform for the expression and dissemination of these of kinds of views, to understand as best as possible the scale, scope, nature and severity of these kinds of practices: when they happen, who they happen to, and why. This is what this research hopes to contribute to.

CASM has conducted extensive work on the ethics and public acceptability of social media research.[4] An ethical framework has been applied to this project, such that:

- The research only uses publicly available data, viewable and visible to any Twitter user;

- The research conducted is aggregated and anonymous: the research does not identify any specific user or users, but aims to understand the overall scale and nature of Islamophobic abuse on Twitter;

[4] See Demos' recent paper with Ipsos MORI *#socialethics: A Guide to Embedding Ethics in Social Media Research* , https://www.ipsos-mori.com/Assets/Docs/Publications/im-demos-social-ethics-in-social-media-research-summary.pdf

- Where quotations are used as examples and elaborations they have been altered to prevent the retrospective identification of any Twitter user on the basis of the quotation, while still maintaining the overall meaning;

- There is no suggestion of any illegality of any of the content measured: the purpose of the research was not to look for content that was illegal, and it does not suggest that the content that was found was illegal. This research is not seeking to inform how laws should be enforced on social media. This research, and Demos' broader research agenda, seeks instead to inform the broader question of how people from different races, religions, sexualities and genders are spoken about on social media, and the extent that people from different backgrounds face abuse and hostility.

METHOD

Twitter data is often challenging to analyse. Data drawn from social media are often too large to fully analyse manually, and also are often not amenable to the conventional research methods of social science. The research team used a technology platform called Method52, developed by CASM technologists based at the Text Analytics Group at the University of Sussex.[5] It is designed to allow non-technical researchers to analyse very large datasets like Twitter.

Defining 'Derogatory anti-Islamic' Messages

This paper is predicated on the training of a machine to be able to distinguish between an expression that is derogatory and anti-Islamic, and an expression that is not.

As described below, this process involves human analysts categorising messages that a series of algorithms then learn from. The categorisation of social media content is inherently impressionistic. An analyst must make judgements about whether a statement fits a given definition or not.

The way this process happens is that analysts look at the data that was collected, and then inductively generate the categories and analytical frames based on what they found. This led to a number of different categories, some that were regarded as derogatory and anti-Islamic, and some that were not.

The categories that were regarded as derogatory and anti-Islamic were:

'**Insult':** This was the largest category of hateful Tweets, and involved Tweets expressing hatred towards Muslims or Islam. This category was primarily composed of Tweets which used slur terms in an abusive way; for example:

> *Every fkin TV ad has a foreign brown or black muzzie in, is there no white Brits in the UK . #UTV #CH4 #BBC #SKY*
>
> *Went into a paki store the other day to get beer raghead there had fucking nerve to I D me so I fucking decked him.*
>
> *London elected a Muslim Sadiq Khan as Mayor, Defended 9/11 Terrorists #svpol islam complete evil dressed up as religion*

[5] This group is led by Professor David Weir and Dr Jeremy Reffin. More information is available about their work at: http://taglaboratory.org

@[XXX] somebody needs to start spraying place with industrial weed killer. That's for the grass. I suggest napalm for the muzzies.

'Muslims are the enemy': This category contains discussion which claims that Islam as a religion is an invasive force attempting to destroy or 'take over' Western countries and institutions, and also contains Tweets which equate Muslims generally with sexual abuse.

U know Muslims are winning in their segregation Jihad when these signs appear in airports. Well done hijabis. [link to picture of a sign saying 'Muslim toilets']

When Islam grows, it moves from praying towards jihad. It makes you submit.

@[XXX] @[XXX] @[XXX] @[XXX] Exactly. 1400 yrs of jihad pedophilia, rape, incest, necrophilia & hasn't evolved the tiniest bit.

'Muslims = Terrorists': This category contains discussion which suggests Muslims are synonymous with terrorism or violent jihad, or which implies that terrorism is exclusively a problem within Islam.

Not every muslim is a terrorist but all terrorists are Muslim

Germany is experiencing full Islam: Immigration Jihad Welfare Jihad Birth Jihad Violent Jihad Cultural Jihad Rape Jihad

The following categories were not judged to be derogatory towards Muslims:

'Quotes and parody': Some of the Tweets incorrectly identified as hate involved users quoting slurs which had been directed at them, or using slur terms to parody people expressing hatred. These cases are particularly difficult for natural language classifiers to identify correctly, as they often contain phrases which, taken by themselves, are likely to be considered hateful.

"Remember: it's can only be a terrorist if it's a Muslim"
"But he said kill a Muslim..."
"Them's the rules"

@[XXX] meanwhile
muslim walks the streets peacefully
white people: islam scum terrorist terrorizes whole town leaving ppl terrorised

When I was a kid, they used to call me paki and attack me & family and smashed our home. Now, one of us, a LONDONER, is Mayor of London.

'Non-hateful use of potential slur terms': This category included users who were judged to be using slur terms in a way which suggests appropriation rather than abuse - who had reclaimed potentially hateful terms and were using them in everyday conversation.

@[XXX] @[XXX] @[XXX] Wowsers Halima how did you not tell that to your fellow paki

@[XXX] @[XXX] We're automatically the best because we are Pakis.

These cases overwhelmingly involved the term 'paki', which was also found to be used as a simple shorthand for Pakistan

Barbara Kay: A female leader did not improve women's lives in Paki…

'Not all muslims are terrorists, and not all terrorists are muslim': In direct opposition to the category, above, which equates Islam with terrorism, (but often involving similar language) a number of Tweets were found to be pointing out that not all terrorists were muslims. A number of these commented of the murder of Jo Cox:

@[XXX] Mair isn't mentally unstable loner He is a terrorist If a Muslim said his name was "death to infidels" you be all over it #JoCox

Can you imagine the uproar if this terrorist had yelled "Allah Akbar" instead of "Britain First" when shooting AND stabbing Jo Cox?!

Many posts were also found denying that Islam and violent attacks were synonymous:

Media never portrays Muslim countries being bombed and attacked EVERY DAY because they want to portray us as the terrorists #PrayForPakistan

@[XXX] how is it clear? Because I don't believe that ALL Muslims are terrorists? I think that proves I know more than you do about Islam

'News sharing and comment': Researchers also found a number of Tweets misclassified as hateful, which commented on or shared news concerning terrorist attacks without invoking Islam itself as the cause:

I completely condemn a brutal terrorist attack in #Istanbul. Terrorism is pure evil. No justification. Condolences to families & friends of victims

Yday Turkish air strike backing jihadi fighters hit 2 internationalists Ypg volunteers [link to article]

Australian judge to jury in trial of jihadi: Islam is not on trial

Data Collection

Method52 was used to directly collect Tweets from Twitter's Stream and Search 'Application Programming Interfaces' (or APIs). They allow all Tweets to be collected that contain one of a number of specified keywords. The keywords used in the various collections used in this research are detailed in the annex.

Data Analysis

Method52 allows researchers to train algorithms to split apart ('to classify') Tweets into categories, according to the meaning of the Tweet, and on the basis of the text they contain. To do this, it uses a technology called natural language processing. Natural language processing is a branch of artificial intelligence research, and combines approaches developed in the fields of computer science, applied mathematics, and linguistics. An analyst 'marks up' which category he or she considers a tweet to fall into, and this 'teaches' the algorithm to spot patterns in the language use associated with each category chosen. The algorithm looks for statistical correlations between the language used and the categories assigned to determine the extent to which words and bigrams (pairs of consecutive words) are indicative of the pre-defined categories.

A series of algorithms were built to respond to the different challenges that this dataset posed in order to identify the anti-Islamic subset within the larger body of data. Each was designed to remove Tweets which were not derogatory and anti-Islamic from the dataset:

- A large number of Tweets contained the word ‘Paki’.[6] A classifier was used to separate derogatory uses of this word from non-derogatory uses.

- A large number of Tweets also contained the word ‘terrorist’. Of course, many Tweets containing this word were in no way derogatory or anti-Islamic. Two classifiers were built to analyse tweets containing these words:
    - First, a classifier was trained to separate Tweets referring to Islamist terrorism from other forms of terrorism.
    - Second, of the Tweets referring to Islamist terrorism, a classifier was built to distinguish views broadly attacking Muslim communities in the context of terrorism, from those broadly defending Muslim communities.

- A classifier was trained to separate all other Tweets in the dataset into those that were derogatory and anti-Islamic from those which were not.

- Last, the Tweets that, based on the above, (a) used the term ‘Paki’ in a derogatory way, (b) that used the term ‘terrorist’ to broadly attack Muslims or Muslim communities, (c) that used the other possible slur terms in the collection in a way that was anti-Islamic were combined. These were then filtered to include only Tweets sent from the UK. This resulted in the final total of derogatory, anti-Islamic Tweets.

CAVEATS

- **There is no suggestion of criminality.** The definitions and categories of ‘hateful content’ used throughout this report are sociological constructs. They are not intended to match or relate to any definition, threshold or framework within criminal law, and there is no suggestion of any criminality through the research.

- **The algorithms used are not perfect:** throughout the report, some of the data will be mis-classified. The technology used to analyse Tweets is inherently probabilistic, and none of the algorithms trained and used to produce the findings for this paper were 100% accurate. In order to get a measure of the overall accuracy of this process, we examined a random sample of 200 Tweets classified as hateful within the year’s collection, and manually annotated them into categories. This process found that 143 of the 200 Tweets were found to be expressing hatred, with 57 Tweets classified as non-hateful. This suggests that our algorithmic classification has a precision of roughly 72%

- **Some data will be missed:** Acquiring Tweets on the basis of the keywords that they contain presents two possible problems. First, the initial dataset may contain Tweets that are irrelevant to the thing being studied. Secondly, it may miss Tweets that are relevant to the thing being studied. Researchers worked to construct as comprehensive a list of keywords as possible (these are detailed in the report, below), however it is likely some were missed, and the numbers presented in this report are likely a subset of the total.

- **Twitter is not a representative window into British society:** Twitter is not evenly used by all parts of British society. It tends to be used by groups that are younger, more socio-economically privileged and more urban. Additionally, the poorest, most marginalised and most vulnerable groups of society are least represented on Twitter; an issue especially important when studying the prevalence of xenophobia, Islamophobia and the reporting of hate incidents.

[6] N.B. whilst this word refers to an ethnic rather than religious group, it was found that it was often used interchangeably to refer to Muslim communities

- Overall, this research is intended to be an indicative, first-take of the reaction on Twitter to these important events. It is not presented as either exhaustive or definitive; and it is very much hoped that it will stimulate further research on this vital topic in the future.

Annex - Data Collection Keywords

Words/Hashtags used to collect Tweets that could be derogatory towards Muslims

- Jihad
- Jihadi
- Sand Flea
- Terrorist
- hijab
- Camel Fucker
- Carpet Pilot
- Clitless
- Derka Derka
- Diaper-Head
- Diaper Head
- Dune Coon
- Dune Nigger
- Durka-durka
- Jig-Abdul
- Muzzie
- Q-Tip Head
- Rab
- Racoon
- Rag-head
- Rug Pilot
- Rug-Rider
- Sand Monkey
- Sand Moolie
- Sand Nigger
- Sand Rat
- Slurpee Nigger
- Towel-head
- Muslim Paedos
- Muslim pigs
- Muslim scum
- Muslim terrorists • Muzrats
- muzzies
- Paki
- Pakis
- Pisslam
- raghead
- ragheads
- Towel head
- FuckMuslims
- WhiteGenocide
- Pegida
- EDL
- BNP
- Rapefugee
- Rapeugee

- mudshark
- kuffar
- kafEir