CST

PROTECTING OUR
JEWISH COMMUNITY

ANTISEMITIC CONTENT ON TWITTER

Published by Community Security Trust.
Registered charity in England and Wales (1042391) and Scotland (SC043612).

CONTENTS

LIST OF FIGURES

FOREWORD

In little more than a decade, and much less in some cases, social media platforms like Twitter, Facebook, YouTube and Snapchat have taken up a central role in discussions about the transmission or communication of antisemitism in modern life and, crucially, how to tackle it.

Between a quarter and a fifth of the antisemitic hate incidents reported to CST in an average year now occur on social media, and most of these are on Twitter – because it is the most public platform, where anybody can view the opinions of anyone else. Antisemitic conspiracy theories, insults and tropes that previously were limited to far right newspapers or muttered comments in the pub can now go around the world and be viewed by thousands of strangers in an instant.

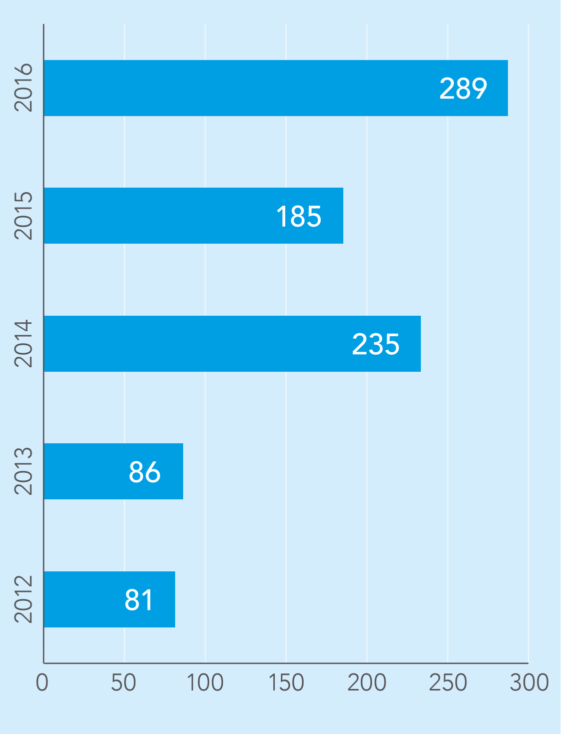
This freedom of expression, and the ability to connect like-minded people around the world, brings huge benefits to most users of social media; but it is exploited by those who wish to harass, threaten or insult Jews, and to encourage others to do the same. As CST wrote in its Antisemitic Incidents Report 2016, online antisemitism involves *"transnational networks of online activists, some of whom are involved in extremist politics"*, facilitated by social media and united by their antisemitism.

The scale of this problem and its influence on how antisemitism operates required new research. This report does not provide an overall count of the total amount of antisemitism on Twitter which would be impossible to measure accurately due to the wide range of ways in which online antisemitism is expressed. Many antisemitic tweets employ coded conspiracy theories or subtle tropes and images that do not make explicit mention of the word 'Jew' and would be hard to pick up in this kind of research. Instead, this report provides insights into the broad patterns in online communications related to Jews in the UK in a specific time period, and the enablers and inhibiters of the spread of antisemitic content on Twitter: what factors help predict its increase, what accounts are associated with the production of antisemitic content, when it spreads, and – crucially – what kind of voices gain most traction on Twitter in speaking out against it.

This research project was led by Professor Matthew Williams and Dr. Pete Burnap of the Social Data Science Lab at Cardiff University, and draws upon the methods developed in previous funded research they have done on anti-Muslim and Brexit related hatred on social media. Professor Williams and Dr. Burnap are the authors of this report and we are grateful to them for their work on this project.

This report, and the research behind it, would not have been possible without funding from an anonymous donor. We are grateful for their support for this work.

Social media incidents recorded by CST



EXECUTIVE SUMMARY

- This report presents an analysis of the production and propagation of **online antagonistic content related to Jews** posted on Twitter between October 2015 and October 2016 in the UK.

- The unprecedented uptake of social media over the past decade has created a significant online forum for the mass production and sharing of opinion, and hence a rich source of information on public sentiment towards topics and events. This study demonstrates how a unique blend of computational and social science techniques can be harnessed to transform and analyse these new forms of data to gain **insight into the growing problem of online antisemitism in the UK**.

- Hate crimes reported to the police in England and Wales have increased by 29 per cent, from 62,518 (2015/16) to 80,393 (2016/17). The most recent estimates from the Crime Survey of England and Wales (CSEW) show that racial and religious aggravated hate crimes increased by 4.5 per cent, from 112,000 per year (13-15 two-year average) to 117,000 per year (15-17 two-year average). However, despite the robust nature of CSEW statistics, they are limited by their reliance upon victim interviews. Instead of relying on 'terrestrial' data or reports from the public on antisemitic victimisation, **this study used a relatively novel online source, Twitter, to mine big social media data** to reveal patterns of online perpetration at the source.

- Over **31 million** tweets related to Jews and antisemitism were collected globally from Twitter in the 12-month study window. Approximately **2.7 million** of the tweets could be located within the UK, and these formed the dataset for analysis.

- Machine learning, a particular approach to artificial intelligence, was adopted to automatically classify online antagonistic content related to Jews with a high degree of accuracy. In total, 9,008 original tweets were classified as antagonistic (15,575 including retweets), representing 0.7 per cent of the UK dataset. **This volume is similar to the frequency of anti-Muslim content measured on Twitter in previous research**. The classifier was trained using a dataset containing comments related to Jews, antisemitism and explicit terms. The research did not capture tweets that, for example, expressed antisemitic conspiracy theories (or allusions to such theories) or antisemitic images posted without accompanying antisemitic text. Nonetheless, it shows that **antagonistic content related to Jews represents only a small proportion of the overall content relating to Jews on Twitter**.

- A timeline of tweets (antagonistic and non-antagonistic) was produced over the 12-month period. This analysis showed significant variability in the frequency of antagonistic tweets related to Jews over the 12-month study period. **Three spikes in antagonistic content were identified as events related to allegations of antisemitism in the Labour Party** during the period when tweets were collected for this research. In particular, the highest peak in antagonistic tweets towards or about Jews came in late April and early May 2016, following the suspensions of Naz Shah MP and Ken Livingstone from the Labour Party for alleged antisemitism. This indicates that **offline events in mainstream politics and public life can trigger online antisemitism**.

- The three events that showed temporary peaks in antagonistic content towards Jews on Twitter were subject to statistical modelling to reveal the enabling and inhibiting factors related to the production of antagonistic content, and the propagation of information flows. Across all events, accounts identified as antisemitic by CST were most likely to produce antagonistic content, while verified and media accounts were least likely, lending strong evidence in support of the **accuracy of the machine learning classifier** built for the study.

- Statistical information propagation modelling revealed that antagonistic content towards Jews was least likely to be retweeted and to survive for a long period across all events, supporting previous research on the 'half-life' of hate speech on social media. Overall, antagonistic content towards Jews was unlikely to propagate in terms of volume and survived between one and three days on Twitter. In two of the three events subjected to this modelling, a minority of information flows stemming from antisemitic agents identified by CST were likely to survive between three and seven days in the 15-day study windows. Conversely, many more information flows emanating from Jewish organisations and media survived for the same duration of three to seven days. This shows that **information flows from antisemitic agents on Twitter gain less traction in terms of duration than flows produced by organisations challenging these negative narratives**.

- The analysis also revealed that the frequency of antagonistic content was on average 32 per cent higher in the second half of 2016 compared to the first half of the year. CST found a similar sustained increase in antisemitic incidents reported both on and offline in the same period, as detailed in their Antisemitic Incidents Report 2016. This suggests that, despite the relatively short 'half-life' of antagonistic content towards Jews, **once this temporary increase in online hate speech receded it left behind a new, higher baseline of online hate**. The same phenomenon was found by similar research into other forms of online hate following the EU referendum in June 2016 and the Woolwich terror attack in May 2013.

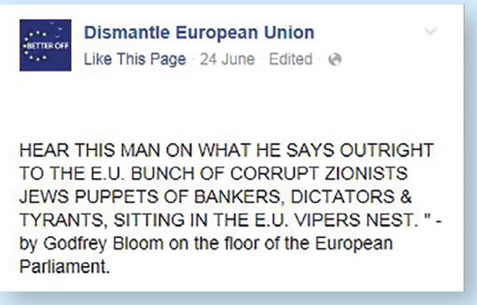
- The small (in terms of retweeting) but sustained (in terms of survivability) information flows of a minority of antisemitic agents indicate that there is limited endorsement of these Twitter narratives. Where there is support, and where antagonistic content is retweeted, it emanates from **a core group who seek out each other's messages over time: an online 'echo chamber' of like-minded individuals** who encourage and amplify their antisemitic narratives. This suggests that contagion of antagonistic information flows appears to be contained within groups of like-minded users and, while it may be viewed by others, it is unlikely to be accepted and disseminated widely beyond such groups.

- Information flows emanating from Jewish organisations and media gained significant traction during two of the three events, evidenced by the combined positive size and survival findings. Overall, this report confirms previous research that **positive content and mainstream information sources have greater longevity and propagate further on Twitter than antagonistic content**, which does not propagate widely in terms of size or survival.

INTRODUCTION

This report presents an analysis of the production and propagation of online antagonistic content related to Jews posted on Twitter between October 2015 and October 2016 in the United Kingdom. During this period, several events unfolded related to claims of antisemitism in the Labour Party, resulting in discernible online social reactions. This was also a period of heightened offline hate crime following the EU referendum vote. The need for the research emanated from the growing use of social media to target minority groups with threatening and grossly offensive communications. While there is a rich body of research on antisemitism in the United Kingdom and overseas, few studies focus on the Internet as a means of the delivery of hate, and the organisation of hateful groups around events and message content. The analysis of the online dataset produced for this report is based on the novel application of computational and social science techniques for studying the production and propagation of 'information flows' on social media.

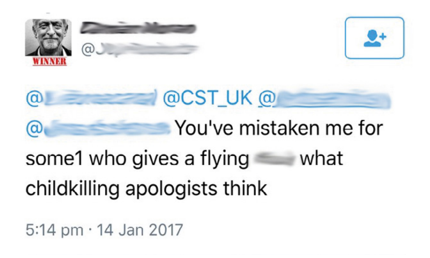
Tweets reported to CST after the Brexit referendum and from a Labour supporter



Dismantle European Union

Like This Page · 24 June · Edited ·

HEAR THIS MAN ON WHAT HE SAYS OUTRIGHT TO THE E.U. BUNCH OF CORRUPT ZIONISTS JEWS PUPPETS OF BANKERS, DICTATORS & TYRANTS, SITTING IN THE E.U. VIPERS NEST. " - by Godfrey Bloom on the floor of the European Parliament.



[redacted]

@[redacted]

@[redacted] @CST_UK @[redacted] @[redacted] You've mistaken me for some1 who gives a flying [redacted] what childkilling apologists think

5:14 pm · 14 Jan 2017

The tweet examples included throughout the report do not stem from the research conducted for the report. They have been taken from the CST database of reported content and have been deemed printable for one or more of the following reasons: the tweet has already been printed in previous CST reports; the tweet is from an organisational account where a private individual is not identifiable; the tweeter provided their informed consent for the tweet to be reproduced outside of Twitter; the tweet has been recorded by CST as an antisemitic incident.

The study was guided by a set of research questions co-produced by Community Security Trust and the research team:

1. Can tweets related to Jews and antisemitism be collected from Twitter and located within the UK over a 12-month period?

2. Can artificial intelligence techniques, such as machine learning, be adapted to automatically classify online antagonistic content related to Jews with accuracy?

3. Utilising machine learning, can a timeline of antagonistic tweets be produced over the 12-month period, to identify patterns related to offline events?

4. Can the enablers and inhibiters of the production of antagonistic content be identified via statistical modelling techniques?

5. Can spikes in communications via Twitter be isolated to facilitate statistical information propagation modelling?

6. Does antagonistic content about Jews propagate in size (number of retweets) and survival (duration of retweets)?

7. What types of Twitter Actors (e.g. Jewish organisations, media, MPs) gain the most information flow traction?

PATTERNS OF HATE CRIME AND SPEECH

Hate crimes reported to the police in England and Wales have increased by 29 per cent, from 62,518 (2015/16) to 80,393 (2016/17). The most recent estimates from the Crime Survey of England and Wales (CSEW), the definitive source of information on criminal victimisation, show that racial and religious aggravated hate crimes increased by 4.5 per cent, from 112,000 per year (13-15 two-year average) to 117,000 per year (15-17 two-year average).[1] This follows a pattern of decline between 2008 and 2015, in line with a general decline in all forms of crime. Overall the number of police referrals to the Crown Prosecution Service (CPS) increased by 0.7 per cent in 2016/17, though this follows a 9.6 per cent drop in police referrals from 2014/15 to 2015/16. CPS data show a total of 14,480 prosecutions for hate crimes in 2016-17 in England and Wales compared to 15,442 the previous year; a 6.2 per cent drop. This is not surprising given we know the bulk of the rise in race and religious aggravated crimes was due to increased volume in criminal damage and public order offences, such as causing fear, alarm or distress (58.6 and 47.8 per cent change respectively between July 2015 and July 2016). It is less likely that the police or victims can identify the perpetrators in relation to these two types of offences, meaning apprehension and prosecution are low.

Evidence suggests that hate crimes tend to have a deeper and more lasting impact on victims than non-hate crimes. Across all CSEW impact measures but one (annoyance), victims of hate crimes report much higher levels of harm (including, in order of magnitude: anger; shock; fear; loss of confidence; difficultly sleeping; anxiety/panic attacks; depression and crying). The 2012/13 to 2014/15 CSEW showed that overall 11 per cent of adults were *"very"* worried about being subject to a physical attack because of their skin colour, ethnic origin or religion. The survey also illustrates that rates of under-reporting vary significantly between different strands of hate crime: recent figures suggest that one in two racist hate crimes are reported to the police, falling to one in four for homophobic hate crimes, one in ten for religiously motivated hate crimes, and one in 19 for disability hate crimes (Corcoran and Smith, 2016). Reporting of hate crime has increased since 2012, which is thought to be due to improved victim confidence in coming forward and police practice.[2] However, hate crime still remains significantly under-reported, and victims of hate crime are less satisfied with the response they receive from the criminal justice system than victims of other crimes (Corcoran et al., 2015).

Despite the robust nature of CSEW statistics, like police recorded hate crime and subsequent CPS prosecutions, they are limited by their reliance upon victim reporting and memory of incidents. In response to this, governments are investing resources into alternative techniques for identifying changes in patterns of racial and religious antagonism between individuals and communities.[3] Researchers are turning to online sources, such as social media, that provide a digital evidence trail left behind by perpetrators. Using a blend of computational and social science techniques, perpetration can be identified as it happens, and stored in databases for detailed analysis. This means analysts do not need to rely upon victim and witness reports, or police recording practices, to identify peaks and troughs in the volume of hate speech online. While not all racial and religious antagonistic content on social media meets the criminal threshold set out by the CPS, some of it does; and more content is deemed sufficiently offensive to warrant requests to social media providers to delete content for infringing platform terms of service.[4] Indeed, many racially and religiously

[1] https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/adhocs/007743csewestimatesofnumberofraceandreligionrelatedhatecrimeinenglandandwales12monthsaveragesyearendingmarch2014toyearendingmarch2017

[2] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/543679/Action_Against_Hate_-_UK_Government_s_Plan_to_Tackle_Hate_Crime_2016.pdf

[3] https://external.ojp.usdoj.gov/selector/awardDetail?awardNumber=2016-MU-MU-0009&fiscalYear=2016&applicationNumber=2016-90958-CA-IJ&programOffice=NIJ&po=NIJ ; and http://gtr.rcuk.ac.uk/projects?ref=ES%2FP010695%2F1

[4] CST has 'trusted flagger' status with Twitter, Facebook and YouTube and reports antisemitic content on behalf of complainants. See also https://www.adl.org/adl-cyber-safety-action-guide

antagonistic posts can cause fear, alarm and distress, especially when they are grossly offensive or threatening in nature. The largest independent study of hate crime in the UK showed that victims were most impacted by so-called 'low level' incidents, such as name calling in public and threats, due to their frequent and incessant nature, often resulting in anxiety, stress and depression (Williams and Tregidga, 2013).

Previous research that analyses online sources, such as Twitter, has shown an increase in the production of online hate speech around events such as the EU referendum vote in June 2016 and the Woolwich terror attack in May 2013. These increases are temporary (24-48 hours) indicating a 'half-life of hate' (Williams & Burnap, 2016). However, previous research also found that once this temporary increase in online hate speech receded, it left behind a new, higher baseline of online hate (for example, after the EU referendum vote). This report on the production and propagation of antagonistic posts related to Jews shows the same increase.

ONLINE HATE SPEECH

Despite online hate speech being evident from the beginning of the domestic Internet,[5] it has only recently become identified as a social problem that requires addressing. The prominence of the problem is linked to the recognition that online spaces, such as social media platforms, now represent new public spaces where key aspects of civil society are played out (Mossberger et al., 2008). Reflecting this, the Crown Prosecution Service has issued guidance to police establishing online networks as 'public spaces' allowing for prosecution to be brought under the Public Order Act as well as the Malicious Communications Act (Crown Prosecution Service, 2015). In 2015 the sending of menacing messages via the Internet became punishable by up to two years imprisonment (Malicious Communications Act 1998 as amended by the Criminal Justice and Courts Bill 2015).

Defining online hate speech is complex given cultural and linguistic variations. However, legal scholars have focussed on the expressive value of language in their attempts to classify hateful speech. Greenawalt (1989) states that any analysis has to consider the extent to which language has expressive value. He considers four criteria that might justify making such expressions criminal: i) that they might provoke a response of violence; ii) that they may deeply wound those at whom the speech is directed; iii) that such speech causes offence to those that hear it; and iv) that slurs and epithets have a degrading effect on social relationships within any one community.

Several of these conditions are encapsulated within legal provisions. In England and Wales, hate crime is prosecuted under a range of legislation including the Crime and Disorder Act 1998, the Criminal Justice Act 2003, the Malicious Communications Act 1998, the Communications Act 2003, the Protection from Harassment Act 1997, the Offences Against the Person Act 1861 and the incitement provisions of Part III of the Public Order Act 1986. Hateful social media posts (other than those which amount to specific offences in their own right, such as making threats to kill, blackmail, stalking etc.) will be considered to be criminal if:

- Their content is grossly offensive;

[5] Stormfront existed in bulletin board format in the early 1990s before being reformed as a website.

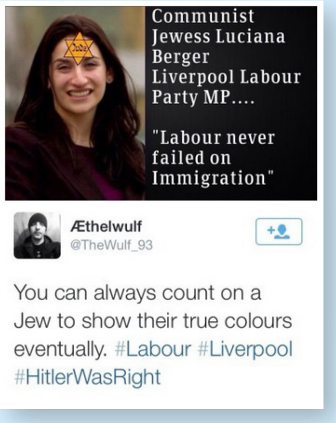
- Their content is threatening or abusive and is intended to or likely to stir up racial hatred;

- Their content is threatening and is intended to stir up hatred on the grounds of religion or sexual orientation.

When considering cases involving offensive communications, prosecutors operate a high threshold at the evidential stage and consider whether a prosecution is in the public interest based on the nature of the communication and the impact upon the targeted victim. They must also be satisfied that the communication is not protected under the free speech principle of the European Convention on Human Rights (Article 10), that provides the freedom to cause offence.

Despite these provisions, for over a decade much of the hate speech that has manifested online (pre-social media) met with little criminal justice response in the UK. Further afield, in countries like the US, it continues largely unchallenged by law enforcement due to freedom of speech protections. Levin (2002) studied how US right-wing groups promoted their goals on the web largely unchallenged by law enforcement, concluding that the online medium has been useful to hatemongers because it is economic, far reaching and protected by the First Amendment. Perry and Olsson (2009) found that the web created a new common space that fostered a 'collective identity' for previously fractured hate groups, strengthening their domestic presence in countries such as the US, Germany and Sweden. They warn a 'global racist subculture' could emerge if online hate is left unchallenged. Eichhorn (2001) focussed on how the online environment opens up the possibility for a more immediate and radical recontextualisation of hate speech, while also highlighting its affordances for more effective modes of response, such as vigilantism and counter-speech. Leets (2001) in a study of the impacts of hate related web-pages found that respondents perceived the content of these sites as having an indirect but insidious threat, while Oksanen et al. (2014) showed how 67 per cent of 15- to 18-year olds in their study had been exposed to hate material on Facebook and YouTube, with 21 per cent becoming victims of such material. Keipi et al. (2017) found that the use of social media across the US, Germany, Finland and the UK is associated with an increased risk of encountering hate speech and harassment online. They concluded that anti–hate speech laws may provide a source of security against exposure. These final studies evidence how the rise of social media platforms has been

Antisemitic tweet sent to Luciana Berger by Garron Helm in 2014

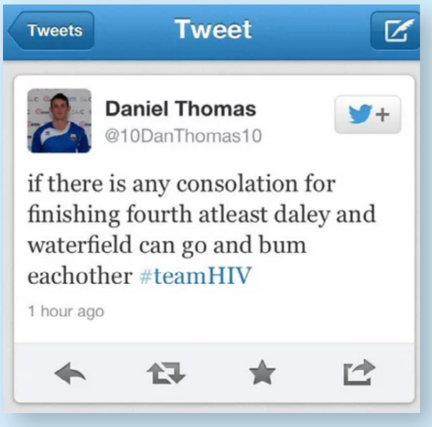

Communist Jewess Luciana Berger Liverpool Labour Party MP....

"Labour never failed on Immigration"

Æthelwulf
@TheWulf_93

You can always count on a Jew to show their true colours eventually. #Labour #Liverpool #HitlerWasRight

Homophobic tweet sent to Olympic diver Tom Daley



Daniel Thomas
@10DanThomas10

if there is any consolation for finishing fourth atleast daley and waterfield can go and bum eachother #teamHIV

1 hour ago

accompanied by an exponential increase in online hate speech.

Williams and Burnap (2016) argued that following trigger events, it is often social media users who are first to publish a reaction. For example, in 2012, Liam Stacey was sentenced to 56 days in prison for posting racist comments on Twitter after a footballer's cardiac arrest, and Daniel Thomas was arrested after a homophobic message was sent to Olympic diver Tom Daley. In 2014, Isabella Sorley, John Nimmo and Peter Nunn were jailed for abusing feminist campaigner Caroline Criado-Perez and MP Stella Creasy, and Declan McCuish was jailed for a year for tweeting racist comments about two Rangers Football Club players. Serial offender Nimmo was jailed again in 2017 for sending antisemitic and threatening online messages to Luciana Berger MP, who had previously been targeted with antisemitic abuse by Garron Helm on Twitter in 2014. Helm received a short custodial sentence and fine. While these extreme negative cases met with a firm criminal justice response, tens of thousands of other users posting less extreme views in relation to these events went unpunished.

Williams and Burnap (2016) developed an automated online hate speech classification tool to identify hate speech originating from individual Twitter users following the Woolwich, UK terrorist attack in 2013. They found that those identifying with right-wing political groups were most likely to produce hateful content on Twitter following the attack. Like offline hate, online hate speech was shown to spike and rapidly decline within the first 48 hours of the attack, indicating a 'half-life'. They concluded that social media acts as a force-amplifier for hate as it can open up a potential space for the rapid galvanising and spread of hostile beliefs, via the spread of rumours through online contagion.

EXPLANATIONS OF ONLINE HATE SPEECH

Recent studies of online radicalisation confirm the importance of contacts and peers via the Internet (von Behr et al., 2013). Similar observations have been made in relation to computer hacking (Holt, 2007), digital piracy (Holt and Copes, 2010), and sexual deviance (Quinn and Forsythe, 2013). One mechanism of peer influence is 'techniques of neutralisation' (Sykes and Matza, 1957), the idea that people often acquire from their peers rationalisations for deviant actions. Features of the Internet itself may contribute to at least temporary relaxation of social norms enabling delinquent acts, such as the production of online hate speech. Suler (2004) and Williams (2006) observed an online disinhibition effect among Internet users, suggesting that affordances of the online environment (reduction in social cues, anonymity etc.) serve to relax certain inhibitions.

A combination of these factors is likely to explain why someone might send hateful and grossly offensive social media communications. Individuals may begin by getting involved in 'gateway' cyber deviance, such as bullying and mild cyber-stalking, where they learn that the rewards outweigh the potential chance of being caught and punished. Individuals then learn from peers in online social networks who to target and how. Hate speech then becomes 'normalised' and labelled as 'desirable' on social networking sites amongst certain peer networks.

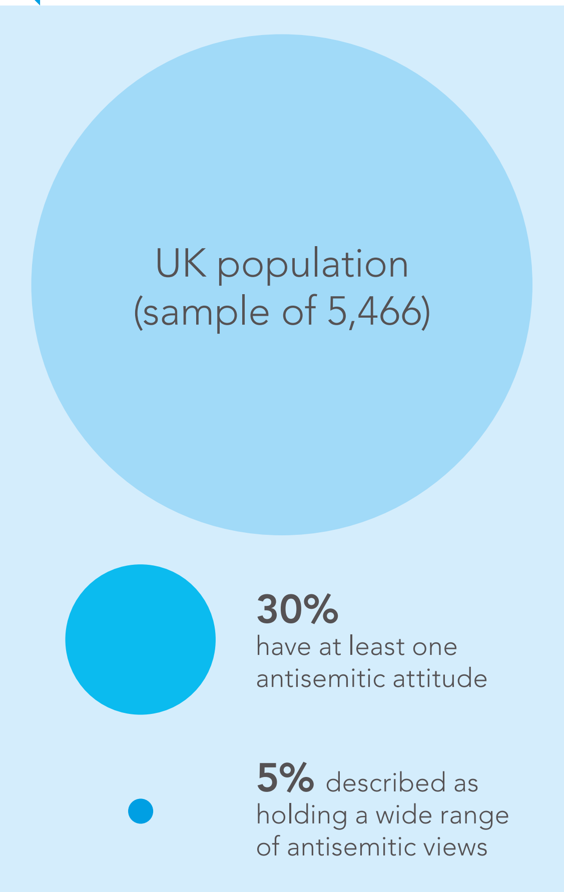
ANTISEMITISM AND ONLINE INCIDENTS

The Anti-Defamation League commissioned First International Resources to research attitudes and opinions towards Jews in over 100 countries around the world. In 2015 this survey showed that around eight per cent of the population in the UK held potentially prejudicial views towards Jewish people.[6] CST and the Institute for Jewish Policy Research surveyed 5,466 people in the UK between 2016 and 2017. They found that approximately five per cent of the UK adult population could be described as holding a wide range of negative attitudes towards Jews, and 30 per cent hold at least one antisemitic attitude (Staetsky, 2017).[7]

In 2016 CST recorded 1,346 antisemitic incidents (an average of approximately 112 per month), the highest on record in the UK (CST, 2016a). This rise was not attributable to a single 'trigger' event, but rather a series of events that saw a sustained antisemitic sentiment on the streets and on social media. This series of events included the high-profile allegations of antisemitism in the Labour Party and the increase in hate speech and crime in the run up to and in the aftermath of the EU referendum. Over the past decade social media has become a safe space for launching campaigns of antisemitic hate speech, including harassment and criminal threats directed at members of the Jewish community in the UK. Of the CST's total recorded antisemitic incidents in 2016, 21 per cent (289) were committed via social media. For comparison, CST recorded 185 incidents in 2015 that involved the use of social media, which was 19 per cent of the overall incident total that year. Patterns of anti-Muslim incidents are similar, with Tell MAMA reporting a large proportion of cases involving a social media dimension (Tell MAMA, 2015). CST does not proactively 'trawl' social media platforms to look for incidents of this type, and only records social media incidents that have been proactively reported to them by a member of the public, where the offender is based in the UK or the incident involves the direct antisemitic targeting of a UK-based victim.

CST (2016b: 27) states it *"works closely with several platforms, particularly with Facebook and Twitter, to improve their removal of antisemitic material. In November 2016, following work with CST and several other groups, Twitter launched new policy guidelines to reduce hateful conduct, including antisemitism, from its platform… Twitter's new policy means that users can no longer direct hate against a generalised religious or ethnic group. This led to the suspension of several accounts that CST had long complained of."*

Institute for Jewish Political Research and CST survey on antisemitic attitudes



[6] http://global100.adl.org/#map

[7] The report states: *"A majority of people who agreed with just one negative statement about Jews in this survey also agreed with one or more positive statements about Jews, suggesting that the existence of one anti-Semitic or stereotypical belief in a person's thinking need not indicate a broader, deeper prejudice towards Jews."*

METHODOLOGY

This study uses data collected exclusively from Twitter to identify and trace the propagation of antagonistic content related to Jews. Currently, some of the social interactions produced on social media platforms are free to collect for research purposes. In particular, data from Twitter is free of charge up to a limit of between 3-5 million interactions per day. Access to 'big data' sources of this type provides unprecedented opportunities for researchers to gain insights into the social world in near-real-time, often at low cost as compared to conventional methods, such as social surveys and interviews. Recent computational and social science advances in machine learning and statistical modelling have been made, allowing researchers to utilise these data to address a variety of research questions. For example, transactional data generated during Internet searches has been used to track the spread of flu in the US (Ginsberg et al., 2009) and to build psychological constructs of nations linked to GDP (Noguchi et al., 2014). Twitter posts have been used to investigate the spread of hate speech following terrorist attacks (Williams & Burnap, 2016) and to estimate offline crime patterns (Williams et al., 2016). The next section outlines the array of 'big data' techniques used to collect, manage, transform and analyse the Twitter data used in this study.

Data Collection, Management and Transformation

The data were collected using the COSMOS platform (Burnap et al., 2015), a free software tool that allows researchers to connect directly to Twitter's streaming Application Programming Interface (API) to collect real-time social media posts by specifying keywords. Twitter's streaming API has a policy of allowing users to collect one per cent of worldwide daily Twitter communications. The volume of data collected for this project did not breach Twitter's daily limits at any point; therefore, it is unlikely there are any missing data based on rate limiting. The following keywords were agreed with CST and used for data collection:

> jew, jewish, jews, antisemitic, anti-semitic, antisemitism, anti-semitism, anti semitic, anti semitism, bonehill, stamford hill, golders green, neo nazis, neo nazi, neo-nazi, neo-nazis, nazi, nazis.

These keywords partly reflected events in the UK at the time the research project was designed. They are a combination of generic terms (Jew, Jewish, antisemitic etc.), and terms relating to a far right demonstration directed at the Jewish community in Golders Green in north London, that was planned for summer 2015. This list was not intended to be a comprehensive set of keywords relating to all aspects of antisemitic hate speech. In particular, much antisemitic hate speech comes in the form of conspiracy theories (or allusions to such theories) and images that would not be captured by these keywords. This caveat should be borne in mind when assessing the overall quantity of antagonistic content measured by this research. The data used for this analysis include tweets posted between 16/10/2015 and 21/10/2016 and were collected in real-time (this ensures all tweets are collected).[8] The raw dataset for the complete study period contained 31,282,472 tweets. The dataset was imported into the open source statistics package R for pre-processing and exploratory data analysis. The first aim of pre-processing was to identify UK-based tweets.[9] Three different approaches were adopted using the metadata of each tweet to identify those posted from the UK. First, a list of keywords was identified (e.g. place names) that signalled that the user was UK-based (referenced as the UK pattern henceforth). Using pattern matching techniques, the UK pattern was identified

[8] 8. An alternative was to purchase tweets from the previous year from Twitter, however the data would not include those tweets deleted by users or the platform for breaching Terms of Service. As the content of interest in this study is likely to be deleted, this retrospective data collection process was not an option. It is important to note the dataset collected for the study likely includes deleted tweets, but none of this content is reproduced in the report. All data are presented at aggregate level to preserve the privacy of users.

[9] 9. Unless Twitter users explicitly opt-in to share their locations each time they post a tweet, latitude and longitude coordinates are not provided in the dataset. Unlike other social media platforms, such as Foursquare and Swarm, the majority of Twitter users (>99 per cent) opt out of sharing these exact geo-data.

within account descriptions. Second, the UK pattern was identified with the user reported locations field (that allows users to report their locations under their profile pictures). Lastly, London and Edinburgh were selected from Twitter time-zone user selections (the only two UK-based time zones Twitter provides). In total 2,677,058 tweets were identified as emanating from UK-based users.[10]

The second aim of pre-processing was to classify user types that were of interest for analysis. Using conventional data science methods, such as pattern matching and web scraping in addition to manual inspection, six different user types were identified: Media Agents; Members of Parliament (MPs); Celebrity Agents; Police Agents; Jewish Organisation and Media Agents; and known Antisemitic Agents. To identify Media Agents pattern matching was used against a list of keywords that the media frequently employ in their account descriptions (the media pattern).[11] In total, 181,363 tweets were identified as emanating from Media Agents in the UK dataset. A pre-defined list was used to identify Celebrity Agents.[12] In total, 80 tweets were identified as posted by celebrities in the UK dataset. To identify MP Agents, a web resource was used which tracks the Twitter accounts of current members.[13] A total of 2,950 tweets were identified as emanating from MPs in the UK dataset. To identify Police Agents, a list of force area Twitter accounts was used in combination with identifying lower level accounts (e.g. at basic command level) by using pattern matching on the use of '999' in the user description.[14] All police accounts followed by @CST_UK were also included. In total 162 tweets were identified as emanating from Police Agents in the UK dataset. To identify Jewish Organisation and Media Agents we pattern matched user descriptions against the terms 'Jew', 'Jewish' and 'Jewry' and identified all organisations followed by @CST_UK. A resulting 102 Jewish Organisation and Media Agents were found in the UK dataset, generating 11,599 tweets in the study period. To identify known Antisemitic Agents a pre-defined list was supplied by CST.[15] In total 13,240 tweets were identified as posted by these Agents (note that not all of this content was identified as antagonistic in the analysis – see classification results in Appendix 2). All other users that did not fall into any of these Agent type categories were classified as 'other' Agents in the analysis.

Classifying Antagonistic Content Related to Jewish Identity

Machine learning was used to classify antagonistic content related to Jews in the Twitter dataset. During the process of machine classification experimentation, it became evident that it was not possible to automatically classify antisemitic 'hate speech' with a high degree of accuracy. Work on identifying hate speech has shown variable success rates with accurate classification across multiple protected characteristics. In particular, machine learning has been found to be most accurate at classifying anti-Muslim hate speech and least accurate at classifying anti-Gay-male hate speech (see Burnap & Williams, 2015). Building a classifier to identify antisemitic hate speech proved particularly problematic due to the high degree of disagreement between human coders on what they considered as hateful. Much of the confusion stemmed from a conflation of antisemitic and anti-Israel content on Twitter. CST (2016a:27) note that:

> *"Clearly it would not be acceptable to define all anti-Israel activity as antisemitic; but it cannot be ignored that contemporary antisemitism can occur in the context of, or be accompanied by, extreme feelings over the Israel/Palestine conflict. Discourse relating to the conflict is used by antisemitic incident offenders to abuse Jews; and anti-Israel discourse can sometimes repeat, or echo, antisemitic language and imagery. Drawing out these distinctions, and deciding on where the dividing lines lie, is one of the most difficult areas of CST's work in recording and analysing hate crime."*

[10] The number of tweets classified as emanating from the UK in this study is therefore 8.5 per cent. This figure is in line with general global usage patterns: in Q2 2016 there were circa 313 million active Twitter accounts, and approximately 6.4 per cent of these accounts were located within the UK (Statista, 2016).

[11] False positives were anticipated due to commonly used keywords in the media pattern. For example, any Twitter user can add the keyword 'reporter' to their description (e.g. 'I don't trust reporters') leading to false positives in classification. To check the accuracy of the classification, a random sample of 100 users was manually inspected for false positives. Only 13 false positives were identified, meaning 87 per cent of media users were classified correctly.

[12] For celebrities, drawing on previous work by Bejda (2015), we used a list of most followed celebrities on Twitter and pattern matched them with the complete dataset.

[13] See: http://www.mpsontwitter.co.uk/. The website administrator provided a list of Twitter handles of 590 MPs who served between the 2015 and 2017 general elections.

[14] Police accounts include the following warning in their user description: 'Do not report crime here. Call 999 in emergency/101 for non-emergency'.

[15] A list of 24 accounts was supplied by CST.

Given this complexity, a two-stage process to attaining gold standard human annotation was performed for training the machine learning classifier. In the first stage, a sample of tweets from the UK dataset was taken for the crowdsourced human annotation task on the online CrowdFlower service.[16] Human coders were asked to identify tweets containing *"Antagonistic content related to Jewish identity"* with a Yes/No response. Given the complexity of criminal law relating to online hate, and the high threshold used by prosecutors, the term 'hate speech' was not used to avoid coder confusion between tweets that may constitute a criminal offence, and those that may be offensive, but not reach the criminal threshold. Using the term 'hate speech' may have also resulted in too few tweets being labelled, resulting in insufficient data to train the machine learning classifier.

Results from the Crowdflower human coding task were reviewed and instances where agreement dropped below 75 per cent were dropped from the training data.

A sample of text from Twitter, Facebook and other forms of online communication was provided by CST.[17] These texts were either reported to CST by the public or identified by CST staff, and were deemed to contain antisemitic words and phrases. Not all of the text examples met the criminal threshold set out by the CPS for hate speech on social media.[18] However, many of them were deemed sufficiently offensive to warrant requests to social media providers to delete content for infringing platform Terms of Service.[19] These text examples were used to inform the second stage of human annotation on the Crowdflower subsample (75 per cent agreement), performed by the research team. In this stage researchers made adjustments to Crowdflower coder annotations, guided by the CST sample. For example, tweets coded as antagonistic towards Jews in the Crowdflower dataset, that were clearly only anti-Israel in nature, were recoded. In total, approx. 30 per cent of the Crowdflower dataset was adjusted in this way. The resulting dataset included tweets that target Jews and the Jewish community with words and phrases that intended to antagonise; and which incorporated racial and religious slurs and offensive statements, positive references to Nazism, or Holocaust denial. Many of these tweets would probably not reach the threshold for a criminal offence in England and Wales, but may contravene Twitter's Terms of Service. This final gold standard human annotated dataset was used to train the machine learning classifier (see Appendix 2).

Upon inspection of the results it was determined that the classifier was able to distinguish between antagonistic content related to Jews, and non-antagonistic posts that contained a combination of the keywords used to generate the dataset over the 12-month period of the study. Validation results suggested that overall the most efficient machine learning technique for classifying antagonistic content was Support Vector Machines combined with a Bag of Words approach. In total, this method identified 9,008 original tweets as antagonistic, representing 0.7 per cent of the 1,232,744 original tweets in the UK dataset.[20] This is commensurate with the volume of antagonistic tweets related to Muslim identity following terror attacks in the UK (0.9 per cent; see Williams and Burnap, 2016).

It was estimated that deriving the various metrics for statistical modelling (agent type, antagonistic content etc.) over the whole UK dataset (2.7M tweets) would take a single desktop computer 140 days. To cut down the processing time, High Performance Computing was used, allowing the job to be split and run concurrently over multiple cores.

[16] https://www.crowdflower.com

[17] These data were not used to directly train the machine classifier for several reasons: i) Some of the texts in the sample were derived from sources that were not native to the study dataset; ii) Not all texts were subject to agreement across four annotators; and iii) There was an insufficient number of texts to train the machine learning classifier.

[18] See page 9.

[19] See https://www.adl.org/adl-cyber-safety-action-guide

[20] Original tweets do not include retweets. Including retweets, there were 15,575 tweets classified as antagonistic out of a total of 2,677,058 tweets in the dataset (or 0.6 per cent)

FINDINGS

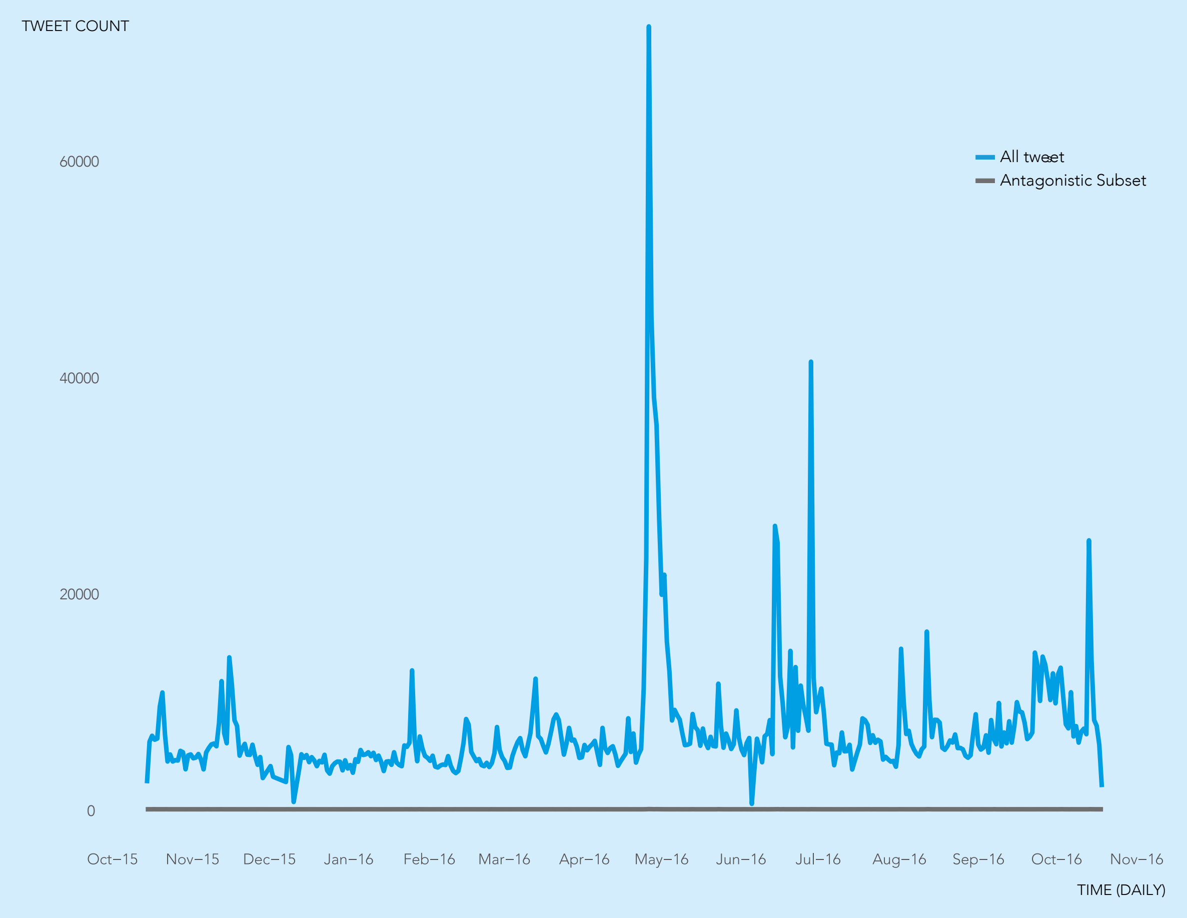
The following sections detail the various stages of analysis required for mining big data obtained from social media sources. The first stage was exploratory data analysis, that involved visualising the whole UK dataset and the antagonistic sub-dataset to identify periods of interest to the next stage of analysis. These periods of interest were then isolated for statistical modelling to identify the enablers and inhibiters of the production of antagonistic content, and the factors that predict information flow size and survival.

Exploratory Data Analysis

Figure 1 presents a time-series line graph of overall tweet frequency (cyan line) and antagonistic tweets (black line) based on the UK dataset. The volume of tweets containing the keywords used for the collection varies considerably over time. For instance, the highest peak in the complete study period for all tweets is around 28th April 2016, the day that Ken Livingstone was suspended from the Labour Party, and the day after Naz Shah MP was suspended, both for alleged antisemitic comments. This indicates offline events probably trigger online discussion that contains the keywords used in the collection, confirming previous research (Williams and Burnap, 2016). The Figure also compares the volume of antagonistic tweets to all tweets using the same scale, illustrating their relative low frequency over the study period.

Figure 2 presents a line graph of antagonistic content related to Jews in the UK dataset. Even though the frequency pattern of antagonistic tweets is not identical to the

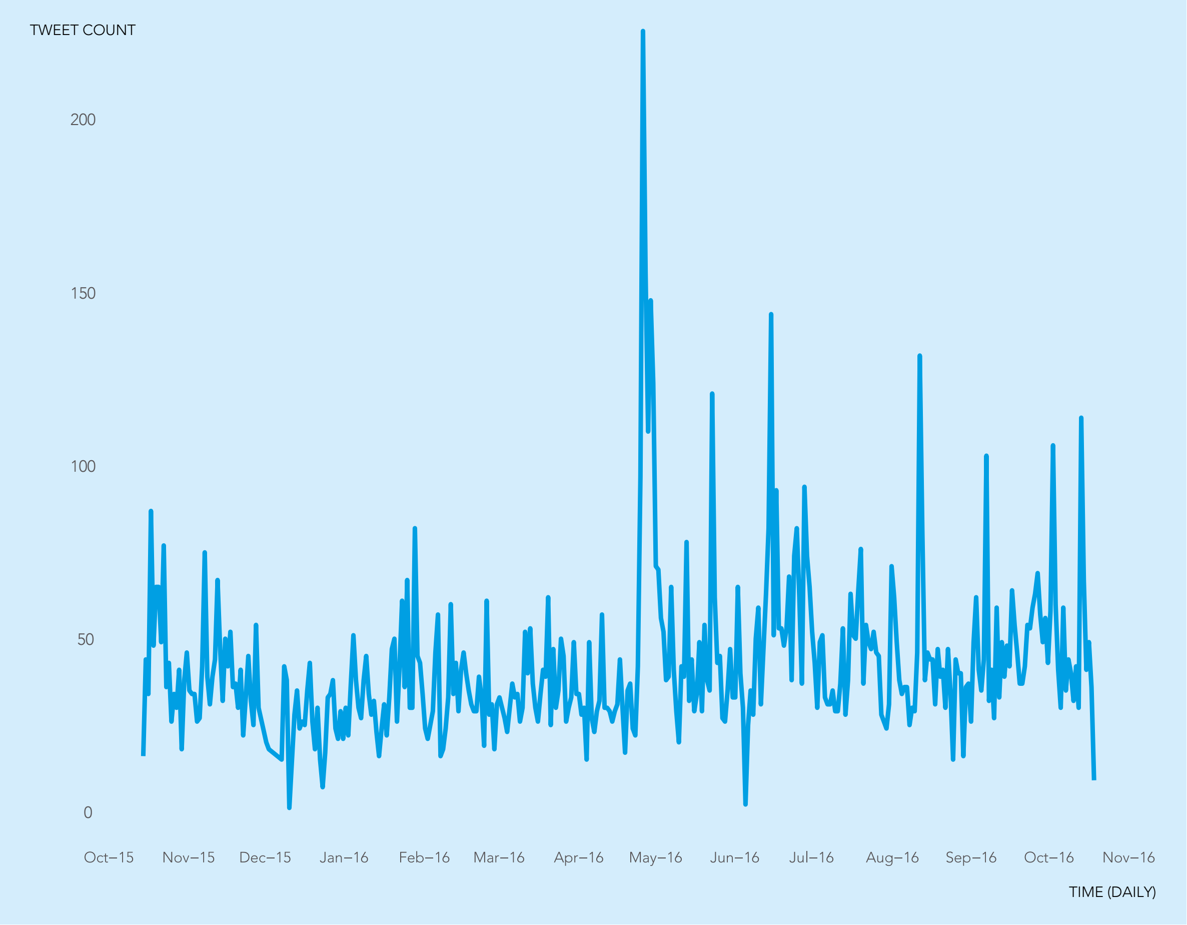
Figure 1: Tweet Frequency (12 months)



pattern of all tweets presented in Figure 1, there are similarities. For example, the highest peak in antagonistic tweets is late April/early May 2016, following the Shah/ Livingstone events. The second highest peak in antagonistic content is mid-June 2016, which is also in line with the peak in mid-June in Figure 1, indicating antagonistic content peaks and falls in line with general discussion about Jews on Twitter. It is of interest to note that the overall frequency of antagonistic content on Twitter is higher in the second half of the data collection window compared to the first (an average of 1,380 antagonistic tweets per month post-April 2016 compared to 1,042 antagonistic tweets per month pre-April 2016). This matches previous research findings that, when temporary increases in online hate speech have receded, they can leave behind a new, higher baseline of online hate. This 32 per cent sustained increase in antagonistic content also correlates with an increase in online and offline antisemitic incidents reported to CST in the same period, with the highest recorded number in May 2016 (CST, 2016a).

As the primary aim of the analysis was to model information propagation in the study period, we selected three events of interest around the highest three peaks in Figure 2: Event-1 includes all tweets posted between 27th April 2016 and 13th May 2016; Event 2 includes all tweets posted between 15th June 2016 and 1st July 2016; and Event 3 includes all tweets posted between 12th August 2016 and 28th August 2016. Subsets of data for each event were created and used in statistical modelling to predict the enablers and inhibiters of the production of antagonistic content and of the propagation of information flows (see Tables 1-3, Appendix 1 for descriptive statistics for each event).

Figure 2: Antagonistic Tweet Frequency (12 months)



Predicting Antagonistic Content

The production of antagonistic content was estimated using generalised ordered logit regression, that allows for the identification of predictive factors. Results for each event are presented in Figures 3-5 (and Table 4, Appendix 1). Across all events, accounts identified as antisemitic by CST were most likely to produce antagonistic content related to Jews. This is unsurprising given the nature of these accounts and their posting history. This finding also lends strong evidence in support of the accuracy of the machine learning classifier built for this study. The only other variables that increased the likelihood of the production of antagonistic content were the control factors of day of week and time of day.

All remaining factors in the analysis decreased the likelihood of the production of antagonistic content. Social factors, such as type of tweeting agent, account verification status, and retweet count, were all negatively associated with the production of antagonistic content. Across all events, verified accounts, those that Twitter deem are 'of public interest and authentic', were significantly less likely to produce antagonistic content, compared to non-verified accounts. Many of these accounts belong to celebrities, public figures, politicians, news organisations, charities, corporations, and government departments. Media Agents and (unsurprisingly) Jewish organisations and media were also significantly less likely to produce antagonistic content. These negative associations add further evidence in support of the accuracy of the machine learning classifier.

Similar to previous research on the spread of online hate speech, tweets containing links to other content (URLs) were less likely to contain antagonistic content. URLs are possibly less common in antagonistic tweets given linked content (most often popular media sources) is less likely to support antisemitic opinion. Contrary to previous research, the inclusion of hashtags in tweets was negatively associated with the production of antagonistic content across the three events (Williams & Burnap, 2016).

Figure 3:
Predicting Antagonistic Content: Event 1

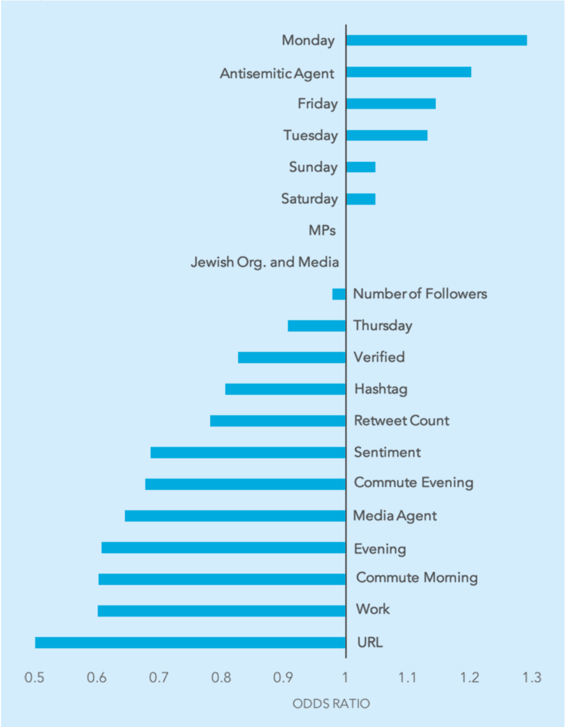


Figure 4:
Predicting Antagonistic Content: Event 2

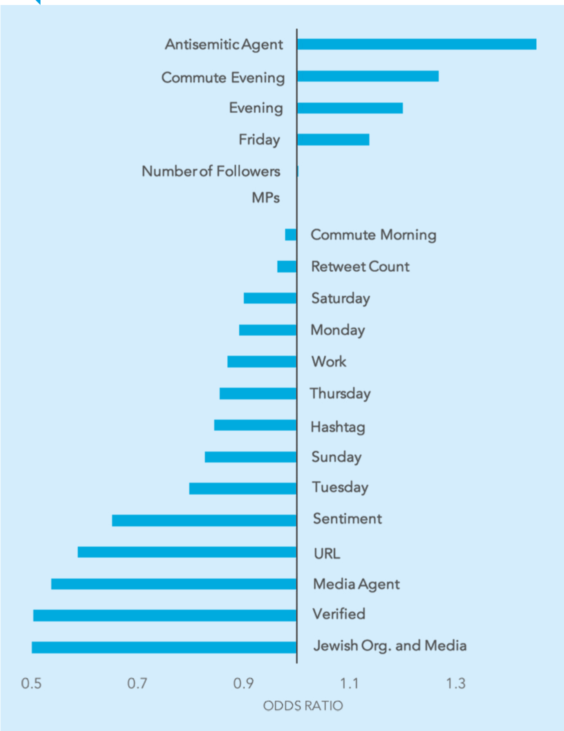
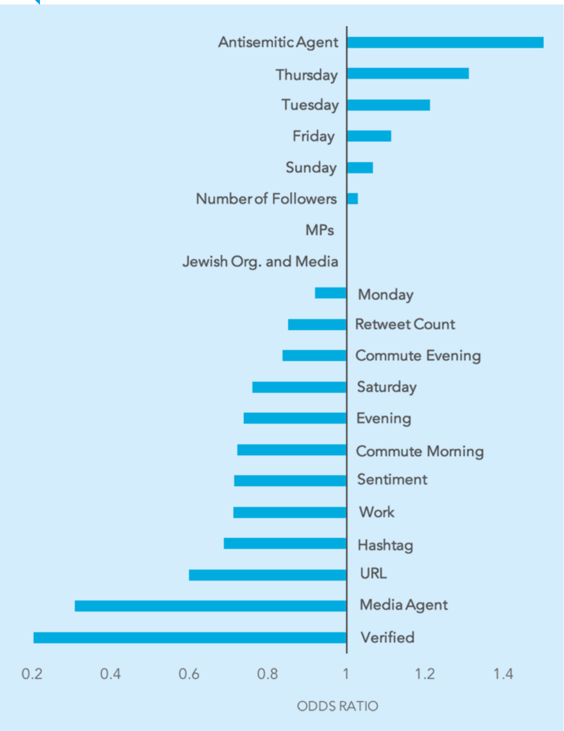


Figure 5:
Predicting Antagonistic Content: Event 3



Information Propagation Modelling

Information propagation modelling can be used to identify the inhibiting and enabling factors of the spread of Twitter messages in a given time frame. A focus on these factors can assist decision making with respect to communication strategy around events. For example, if it is identified that certain types of accounts are important to the spread of information (e.g. Jewish organisations and media accounts) then it would be possible to seek endorsement of messages by these agents, via retweeting. There are two dependent measures in information propagation modelling: Size of information flows (measured by counting the number of retweets) and Survival of information flows (measured by counting the seconds between the first and last retweet). In terms of size, the number of retweets is a measure of the volume of public interest and endorsement of the information, while survival (or duration) is a measure of persistence of interest over time. These measures are established in the literature on online social networks and information propagation (Burnap et al., 2014; Yang & Counts, 2010).

Three sets of variables were entered as independent predictors of information flow size and survival in the models: Content factors, Social factors and Control factors. Content factors relate to the text of the tweet. In the models the following text content features were included: sentiment (binary negative/positive); URLs pointing to an external source (such as a news item); hashtags which create an interest-based micro-network; and antagonistic content. Social factors relate to the characteristics of user accounts. In the models the following user social features were included: number of followers; verified status; and agent type. There were five agent types included: media (newspapers, TV news); MPs; Jewish organisations and Jewish media; known antisemitic accounts (supplied by CST); and other agents (all other Twitter users not included in these categories).[21] Multiple control factors were included that have been shown to influence the flow of information in social media networks (Zarrella, 2009). These include time of day and day of week. Descriptive statistics of each event are presented in Tables 1 to 3 (see Appendix 1).

Two modelling techniques were used to predict the size and survival of information flows: Zero Inflated Negative Binomial (ZINB) regression and Cox's Proportional Hazards regression. ZINB regression was used to model the Size measure as this is best described as a count of retweets. Count variables represent types of events that are largely not experienced by the majority of the sample (in this case retweets where the majority of tweets are not retweeted with a minority being retweeted). Linear regression models are not appropriate for count variables given the nonlinear distribution of the data. Cox's proportional hazards regression was used to model the survival measure. Our interest here was to model the factors that pose hazards to the survival of information flows. Therefore, positive relationships (bars on the right hand side of the Figures) indicate an increased hazard to survival.

[21] Here, it is important to note that presence of police agents and celebrities were either extremely small or non-existent in the three events. Therefore, police and celebrity agents were re-classified under other agents.

Size of Information Flows

Figures 6-8 present the results of the Size models (and Table 5, Appendix 1). Each event only includes original tweets, with the number of retweets entered as the dependent variable. Incidence-Rate Ratios (IRRs) are used to indicate the magnitude of the effect on retweets.[22] Of particular note is the negative relationship between antagonistic content and the size of retweets. In all events antagonistic content did not propagate in terms of size, reflecting previous work on anti-Muslim online hate speech (Williams & Burnap, 2016). Correspondingly, the content posted by antisemitic agents identified by CST did not propagate to a significant extent across the three events. This double negative pattern provides further confidence in the accuracy of the machine learning classifier for antagonistic content related to Jewish identity.

It is important to note that while this content did not propagate, it was produced and published by a minority of Twitter users during the events under study.

Non-propagation in terms of size means that antagonistic content was not retweeted (shared by other Twitter users) to a great extent (and sometimes not at all). This is an encouraging finding, and it indicates that the majority of Twitter users do not endorse these types of posts via the act of retweeting. Research shows that where antagonistic content is retweeted, it is contained within online 'echo chambers' of like-minded individuals.

Across all three events, content posted from Twitter verified accounts was most likely to be retweeted in volume, an unsurprising finding given the types of users behind these accounts (celebrities, public figures, MPs, government departments, media outlets etc.). In all but one of the events (Event 3) MPs were highly likely to be retweeted. This pattern is repeated in relation to Jewish organisations and media. Again, given the nature of the events, the attention on politicians and Jewish media content is not unexpected. Across all three events, Media Agents were positively associated with larger information

Figure 6:
Predicting Information Flow Size: Event 1

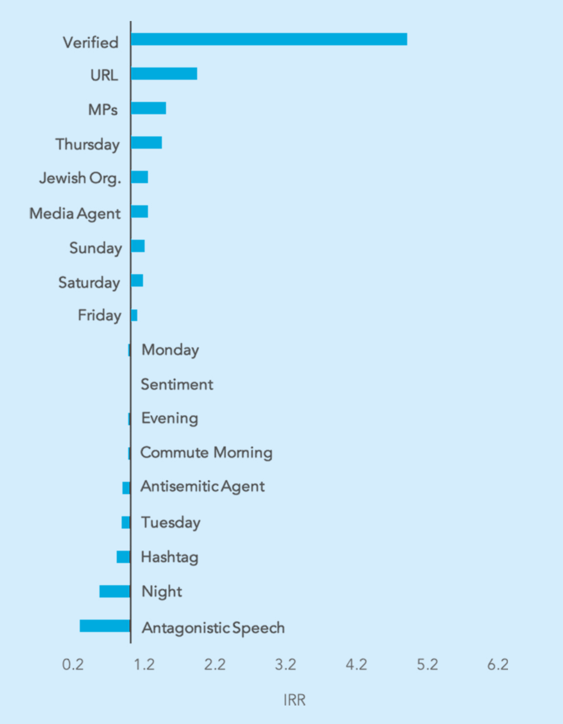
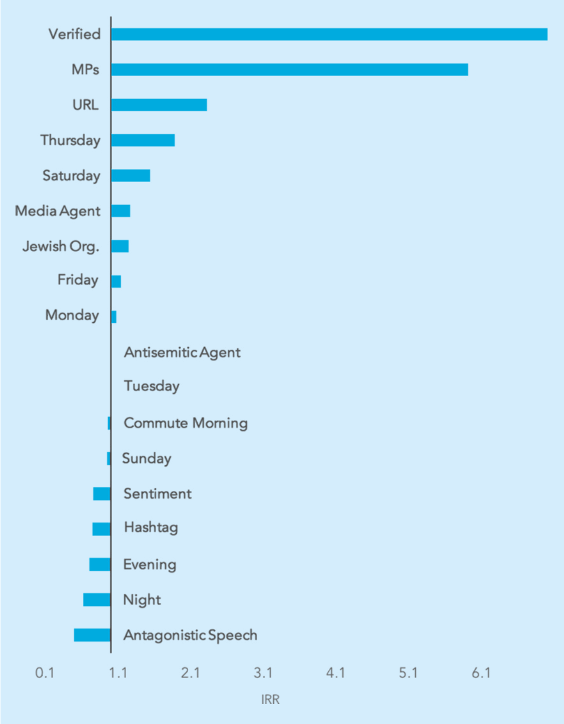


Figure 7:
Predicting Information Flow Size: Event 2



[22] An IRR is a univariate transformation of the estimated coefficient for the ZINB model. It is a relative difference measure used to compare the incidence rates of events (retweets) occurring at any given point in time. A score above 1 indicates an increased incidence rate ratio and below 1 a reduced incidence rate ratio for retweets.

Figure 8:
Predicting Information Flow Size: Event 3

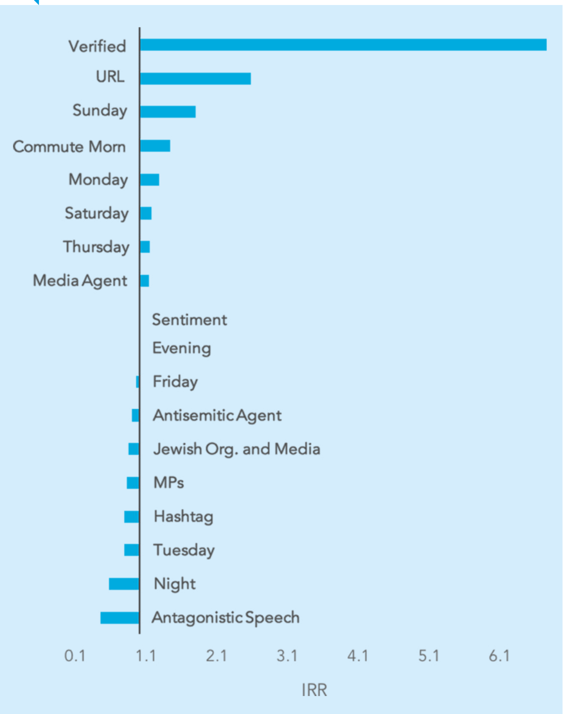
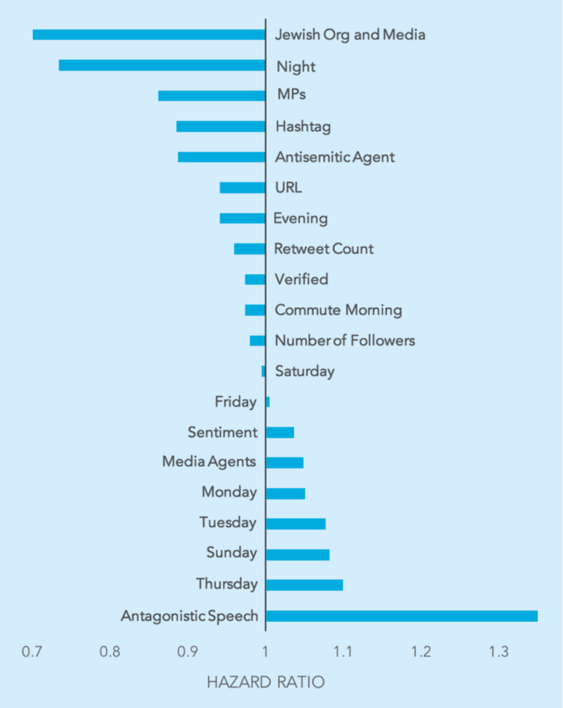


Figure 9:
Predicting Information Flow Survival: Event 1



flows, supporting previous research that indicates 'old media' greatly influence the flow of information on 'new media' platforms (Williams & Burnap, 2016).

Survival of Information Flows

Figures 9, 11 and 13 present the results of the information flow survival models over the three events (and Table 6, Appendix 1). Positive estimates in the Cox regression models (bars to the right of the vertical axes in the Figures) are interpreted as increased hazards to survival and therefore a reduction in the duration of information flows. In all events antagonistic content is negatively associated with long-lasting information flows. In two of the events it emerges as having the highest positive hazard ratio. This finding corroborates previous research, that shows online hate speech does not propagate in terms of size or survival (Williams & Burnap, 2016).

Figure 10: Antagonistic Content Survival Estimates (Events 1-3)

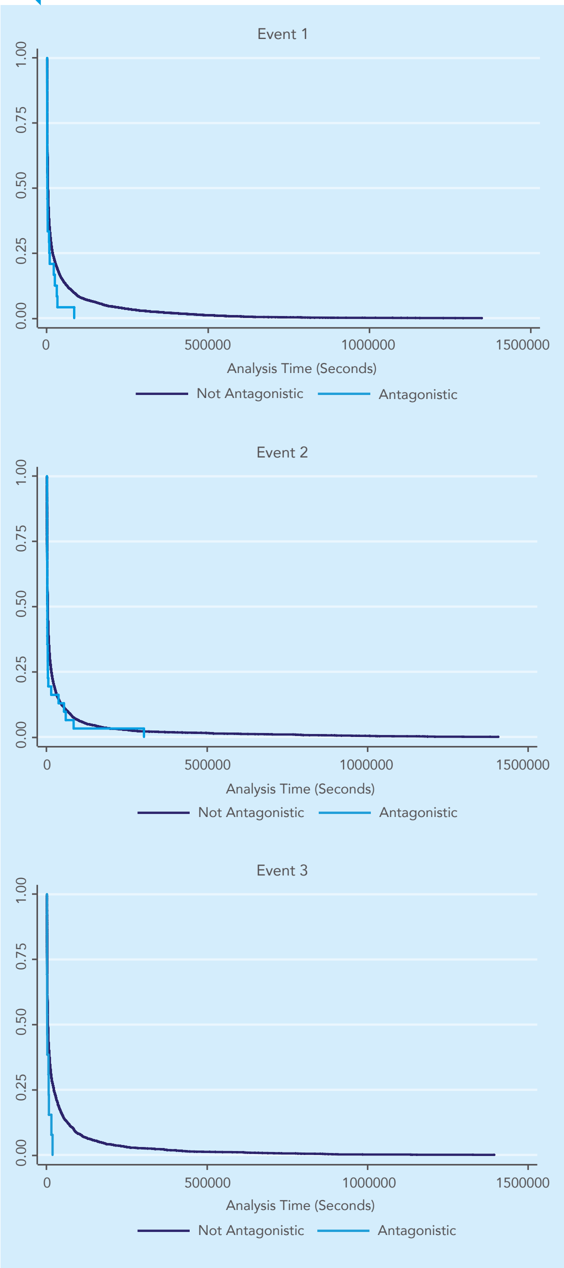
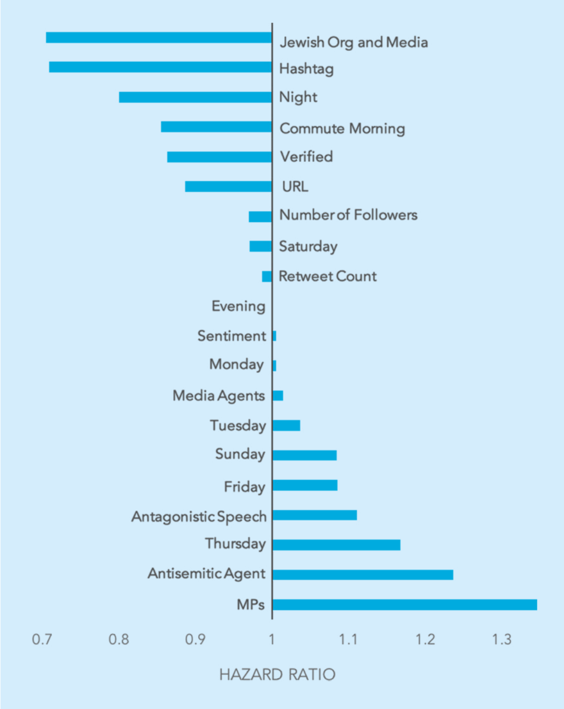


Figure 10 visualises the survival estimates of antagonistic content in the 15-day analysis windows of each event. They show that these information flows survived between one to three days. This sharp de-escalation resonates with research that shows offline hate crime following trigger events has a 'half-life'. It seems likely that this offline pattern is replicated in relation to online antagonistic content concerning Jews.

Figures 9, 11 and 13 show that Antisemitic Agents emerged as having the fourth and fifth highest negative hazard ratios in two of the events. This indicates that information flows emanating from some of these agents during these events were likely to outlast those emanating from Media Agents and more general agents at some points in the 15-day analysis windows. Figure 12 visualises the survival estimates of Agent Type and shows that while information flows from Antisemitic Agents can last between three and seven days, these are in a minority, as many of them die out rapidly (indicated by the steep decline in the cyan lines). Conversely, many more information flows emanating from Jewish organisations and media survive between three and seven days in all events (indicated by a less steep decline in the orange lines). This finding is novel, and shows information flows from Antisemitic Agents gain less traction in terms of duration than flows produced by organisations challenging these negative narratives on social media.

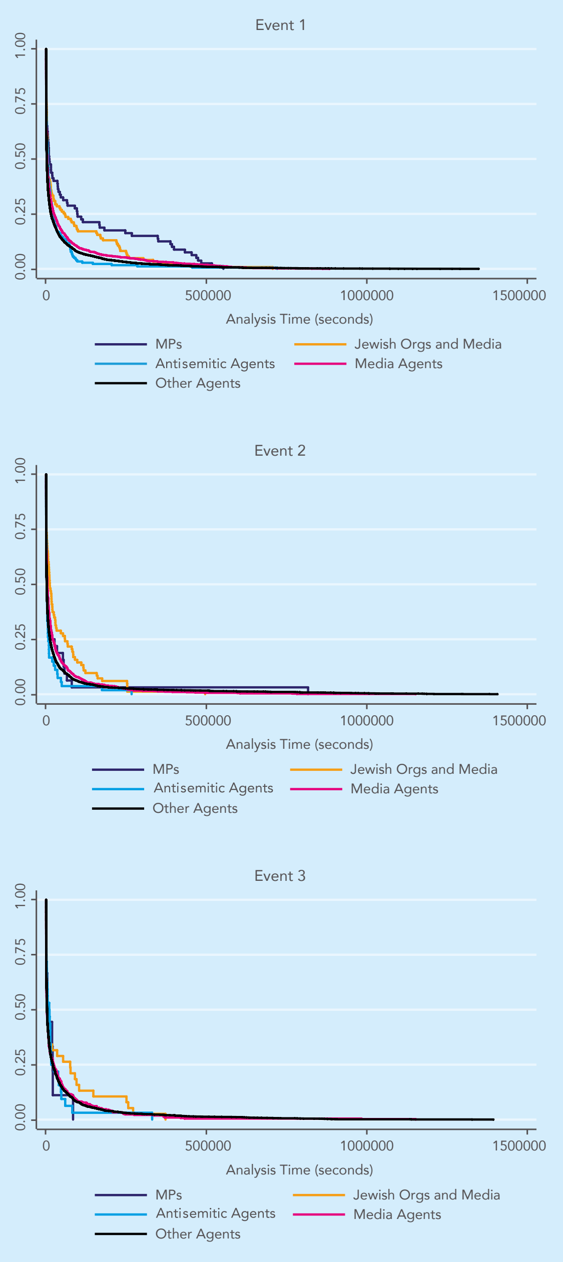
The small (in terms of retweeting) but sustained (in terms of survivability) information

Figure 11:
Predicting Information Flow Survival: Event 2



flows of a minority of Antisemitic Agents indicate that there is limited endorsement of these Twitter narratives. Yet, where there is support it emanates from a core group who seek out each other's messages over time: an 'echo chamber' of like-minded individuals who encourage and amplify each other. This suggests that contagion of antagonistic information flows appears to be contained and, while it may be viewed by others, it is unlikely to be accepted and disseminated widely by other users beyond such groups.

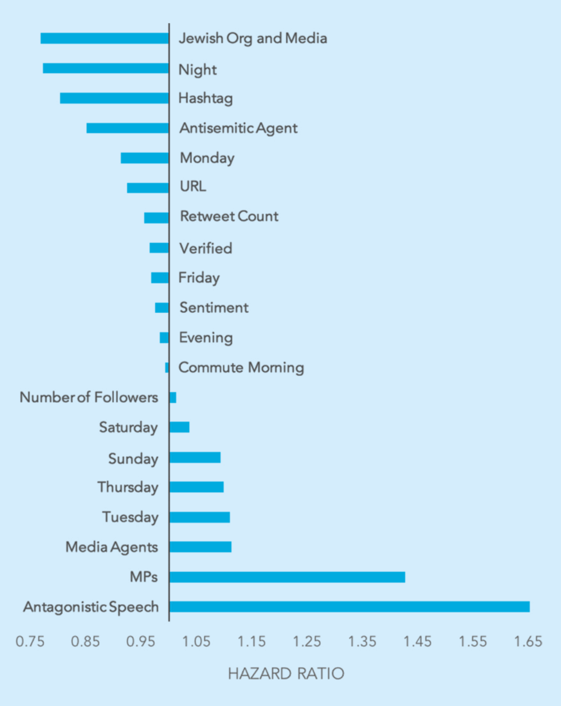
Figure 12: Agent Information Flow Survival Estimates (Events 1-3)



The combined positive size and survival findings relating to Jewish organisations and media show that information flows from these agents gained significant traction during two of the three events.

General Media Agents emerged as having positive hazard ratios for all three events, with many information flows dying out evenly over the study window (see pink line). As indicated in previous research, this is likely to be a result of frequent news turnover, where new stories replace old ones on a daily basis. These new stories create new information flows that replace the old (Williams & Burnap, 2016).

Figure 13:
Predicting Information Flow Survival: Event 3



CONCLUSION

Hate crimes have increased significantly in the past few years in the wake of successive cases of alleged antisemitism in mainstream politics; the vote over the UK's future in the EU; and recent terror attacks. The police record a 29 per cent increase (2015/16 to 2016/17), while the Crime Survey for England and Wales records a 4.5 per cent increase (an additional 5,000 race and religious hate crimes per year on average between 2013-15 and 2015-17). Similarly, CST antisemitic incident figures show heightened incident totals over the same period. Despite the robust nature of CSEW statistics, they are limited by their reliance upon victim interviews, and CST and police statistics rely on crime and incident reports primarily from victims or witnesses. Instead of relying on 'terrestrial' data or reports from the public on antisemitic victimisation, this study used a relatively novel online source, Twitter, to mine big social media data to reveal patterns of perpetration at the source.

Antisemitism (and hate in general) on social media has become a matter of some concern in the Jewish community and in broader public debate. While not all racial and religious antagonistic content on social media meets the criminal threshold set out by the Crown Prosecution Service, some of it is deemed sufficiently offensive or hateful to warrant requests to social media providers to delete content for infringing platform terms of service. Social media acts as a force-amplifier, as it can open up a potential space for the rapid galvanising of prejudiced beliefs, via the spread of negative expression towards minority groups through online contagion. Previous research that analyses online sources, such as Twitter, has shown an increase in the production of online hate speech around events such as the referendum vote and the Woolwich terror attack.

Over the past decade social media has become a safe harbour for launching campaigns of antisemitism, including harassment and criminal threats directed at members of the Jewish community in the UK. Of the total number of antisemitic incidents recorded by CST in 2016, 21 per cent were committed via social media. The online pattern of antagonistic content related to Jews found in this study can act as a proxy for the ebb and flow of negative expressions targeting Jews in the UK.

This analysis showed significant variability in the frequency of antagonistic tweets related to Jews over the 12-month study period. Three spikes in antagonistic content were identified as events related to allegations of antisemitism in the Labour Party. The analysis also revealed the frequency of antagonistic content was on average 32 per cent higher in the second half of 2016. CST found a similar sustained increase in incidents reported both on and offline in the same period (CST, 2016a).

These three events were subject to statistical modelling to reveal the enabling and inhibiting factors related to the production of antagonistic content, and the propagation of information flows. Across all events, accounts identified as antisemitic by CST were most likely to produce antagonistic content, while verified and media accounts were least likely, lending strong evidence in support of the accuracy of the machine learning classifier built for this study.

Information flow propagation models revealed that antagonistic content was least likely to be retweeted in volume and to survive for long periods across all events, supporting previous research on the 'half-life' of hate speech on social media. While information flows emanating from antisemitic agents were unlikely to propagate in terms of volume, in two of the three events, a minority were likely to last between three and seven days in the 15-day study windows.

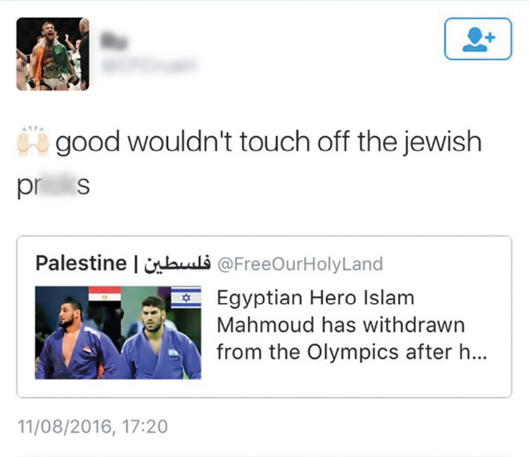
The small (in terms of retweeting) but sustained (in terms of survivability) information flows of a minority of antisemitic agents indicate that there is limited endorsement of these Twitter narratives. Yet, where there is support it emanates from a core group who seek out each other's messages over time: an 'echo chamber' of like-minded individuals. Therefore, contagion of antagonistic information flows appears to be contained and unlikely to be disseminated widely by users beyond such groups, although it can of course be viewed by others.

The study also revealed that information flows emanating from Jewish organisations and media gained significant traction during two of the three events, as evidenced by the combined positive size and survival findings.

These findings should be a source of some optimism. While antisemitism is present on Twitter and can cause severe offence when it is not removed, it is outweighed by positive content, which is present in greater amounts, lasts longer and spreads further than antisemitic content.

The unprecedented uptake of social media over the past decade has created a significant online forum for the mass production and sharing of opinion, and hence a rich source of information on public sentiment towards topics and events. This study has demonstrated how a unique blend of computational and social science techniques can be harnessed to transform and analyse these new forms of data to gain insight into the growing problem of online antisemitism in the UK.

Several antisemitic tweets reported to CST between October 2015 and October 2016

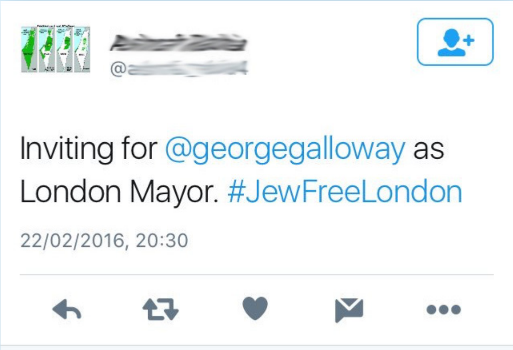


[redacted]
[redacted]

🙌 good wouldn't touch off the jewish pr[redacted]s

Palestine | فلسطين @FreeOurHolyLand
Egyptian Hero Islam Mahmoud has withdrawn from the Olympics after h...

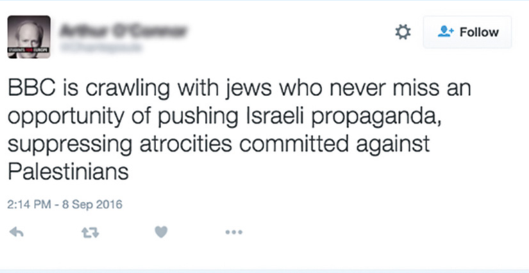
11/08/2016, 17:20



[redacted]
@[redacted]

Inviting for @georgegalloway as London Mayor. #JewFreeLondon

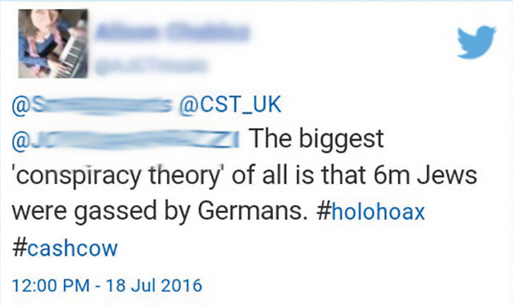
22/02/2016, 20:30



[redacted]
[redacted]

BBC is crawling with jews who never miss an opportunity of pushing Israeli propaganda, suppressing atrocities committed against Palestinians

2:14 PM - 8 Sep 2016



[redacted]
[redacted]

@S[redacted] @CST_UK @J[redacted] The biggest 'conspiracy theory' of all is that 6m Jews were gassed by Germans. #holohoax #cashcow

12:00 PM - 18 Jul 2016

REFERENCES

Bejda, M. (2015). Top 1000 Celebrity Accounts, <https://gist.github.com/mbejda/9c3353780270e7298763>

Burnap, P. and Williams, M. L. (2015). 'Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making', Policy & Internet, 4: 223-42.

Burnap, P., Williams, M. L., Sloan, L., Rana, O. et al. (2014). 'Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack', Social Network Analysis and Mining, 4: 206, 1-14.

Burnap, P., Williams, M. L., Rana, A., Avis, N., Morgan, J. and Sloan, L. et al. (2015). 'COSMOS: Towards an Integrated and Scalable Service for Analysing Social Media on Demand', International Journal of Parallel, Emergent and Distributed Systems, 30:2, 80-100.

Corcoran, H., Lader, D., Smith, K. (2015). Hate Crimes, England and Wales, 2014/15, Statistical Bulletin 05/15, London: Home Office.

Corcoran, H and Smith, K. (2016). Hate Crime, England and Wales, 2015/16: Statistical Bulletin 11/16, London: Home Office.

CST (2016a). Antisemitic Incidents January to June 2016, CST: London.

CST (2016b). Annual Review, CST: London

Eichhorn, K. (2001). 'Re-in/citing Linguistic Injuries: Speech Acts, Cyberhate, and the Spatial and Temporal Character of Networked Environments', Computers and Composition, 18: 293-304.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). 'Detecting influenza epidemics using search engine query data', Nature 457 (7232).

Greenawalt, K. (1989). Speech Crime & the Uses of Language, Oxford: Oxford University Press.

Holt T. (2007). 'Subcultural evolution? Examining the influence of on- and off-line experiences on deviant subcultures', Deviant Behavior, 28: 171-198.

Holt T. and Copes H. (2010). 'Transferring Subcultural Knowledge On-Line', Deviant Behavior, 31: 625-654.

Keipi, T., Kaakinen, J., Oksanen, A. and Räsänen, P. (2017). 'Social Tie Strength and Online Victimization: An Analysis of Young People Aged 15–30 Years in Four Nations', Social Media + Society, 1-12.

Leets, L. (2001). 'Responses to Internet Hate Sites: Is Speech Too Free in Cyberspace?', Communication Law and Policy, 6: 287-317.

Levin, B. (2002). 'Cyberhate: A Legal and Historical Analysis of Extremists' Use of Computer Networks in America', American Behavioral Scientist, 45: 958-88.

Mossberger, K., Tolbert, C. J. and McNeal, R. S. (2008). Digital Citizenship: The Internet, Society and Participation, MIT Press.

Noguchi T., Stewart N., Olivola C., Moat H., Preis T. (2014). 'Characteristing the Time-Perspective of Nations with Search Engine Query Data', PLoS One 9:4.

Oksanen, A., Hawdon, J., Holkeri, E., Nasi, M. and Rasanen, P. (2014). 'Exposure to Online Hate among Young Social Media Users', in M. Nicole Warehime, ed., Soul of Society: A Focus on the Lives of Children & Youth, 253-73. Emerald.

Perry, B. and Olsson, P. (2009). 'Cyberhate: The Globalisation of Hate', Information & Communications Technology Law, 18: 185-99.

Quinn, J. and Forsyth, C. (2013). 'Red Light Districts on Blue Screens: A Typology for Understanding the Evolution of Deviant Communities on the Internet', Deviant Behavior, 34: 579-585.

Staetsky, D. L. (2017). Antisemitism in contemporary Great Britain: A study of attitudes towards Jews and Israel, CST/JPR Report.

Statista (2016). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 3rd quarter 2017, Statista.com <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>

Suler J. (2004). 'The online disinhibition effect', Cyberpsychology and Behaviour, 7: 321-326.

Sykes G. and Matza, D. (1957). 'Techniques of Neutralization: A Theory of Delinquency', American Sociological Review, 22: 664-670.

Tell MAMA (2015). The Geography of Anti-Muslim Hatred: Tell MAMA Annual Report 2015, London: Tell MAMA.

von Behr, I., Reding, I., Edwards, C. et al. (2013). Radicalisation in the digital era, London: RAND Europe.

Williams, M. (2006). Virtually Criminal: Crime, Deviance and Regulation Online, London: Routledge.

Williams, M. L. and Burnap, P. (2015). 'Crime Sensing with Big Data: The Affordances and Limitations of using Open Source Communications to Estimate Crime Patterns', British Journal of Criminology, 52:2, 320-340.

Williams, M. L. and 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: 211-38.

Williams, M. L. and Tregidga, J. (2014). 'Hate Crime Victimisation in Wales: Psychological and Physical Impacts across Seven Hate Crime Victim-Types', British Journal of Criminology, 54: 946-67.

Williams, M. L. and Tregidga, J. (2013). All Wales Hate Crime Research Project: Final Report, Race Equality First.

Yang, J. and Counts, S. (2010). Predicting the Speed, Scale, and Range of Information Diffusion in Twitter, International Conference on Weblogs and Social Media (ICWSM).

Zarrella, D. (2009). 'The Science of Retweets', available online at http://danzarrella.com/science-of-retweets.pdf.

APPENDIX 1: TABLES

Table 1: Descriptive Statistics of Event 1 (N=156,498)

VARIABLES	CODING	MEAN	STD. DEV
Dependent Variables			
Size (retweets)	Range: 0-376	0.3056077	3.240511
Survival (seconds)	Range: 0-1349131	2604.136	28456.9
Antagonistic Speech	0 = no; 1 = yes	0.0049841	0.0704222
Independent Variables			
Content Factors			
Sentiment	-1 = negative; 0 = neutral; 1 = positive	-0.3825608	0.7140561
URL	0 = no; 1 = yes	0.5482818	0.497665
Hashtag	0 = no; 1 = yes	0.1995105	0.3996337
Social Factors			
MPs	0 = no; 1 = yes	0.0017828	0.0421853
Jewish Org. and Media	0 = no; 1 = yes	0.0041726	0.0644608
Antisemitic Agent	0 = no; 1 = yes	0.0061918	0.0784441
Media Agent	0 = no; 1 = yes	0.1069407	0.3090388
Other Agent	0 = no; 1 = yes	0.8809122	0.3238928
Verified	0 = no; 1 = yes	0.0319748	0.1759337
Number of Followers	Categorised into 1-10th percentiles	5.499367	2.872617
Control Factors			
Work	0 = no; 1 = yes	0.4250214	0.4943478
Commute Morning	0 = no; 1 = yes	0.1128002	0.3163494
Commute Evening	0 = no; 1 = yes	0.1349091	0.3416275
Evening	0 = no; 1 = yes	0.2530831	0.4347795
Night	0 = no; 1 = yes	0.0741863	0.2620746
Sunday	0 = no; 1 = yes	0.1185766	0.3232907
Monday	0 = no; 1 = yes	0.0959948	0.2945851
Tuesday	0 = no; 1 = yes	0.0818094	0.2740749
Wednesday	0 = no; 1 = yes	0.1466536	0.3537614
Thursday	0 = no; 1 = yes	0.252016	0.4341718
Friday	0 = no; 1 = yes	0.180456	0.384568
Saturday	0 = no; 1 = yes	0.1244936	0.3301449

Table 2: Descriptive Statistics of Event 2 (N=78,432)

VARIABLES	CODING	MEAN	STD. DEV
Dependent Variables			
Size (retweets)	Range: 0-1550	0.3546384	8.780562
Survival (seconds)	Range: 0-1407814	2170.419	31109.54
Antagonistic Speech	0 = no; 1 = yes	0.0088484	0.0936496
Independent Variables			
Content Factors			
Sentiment	-1 = negative; 0 = neutral; 1 = positive	-0.4498929	0.7063633
URL	0 = no; 1 = yes	0.5154784	0.4997635
Hashtag	0 = no; 1 = yes	0.1851413	0.3884146
Social Factors			
MPs	0 = no; 1 = yes	0.00102	0.0319212
Jewish Org. and Media	0 = no; 1 = yes	0.0050362	0.0707878
Antisemitic Agent	0 = no; 1 = yes	0.0035827	0.0597489
Media Agent	0 = no; 1 = yes	0.0864061	0.2809645
Other Agent	0 = no; 1 = yes	0.903955	0.2946548
Verified	0 = no; 1 = yes	0.227841	0.1492154
Number of Followers	Categorised into 1-10th percentiles	5.498317	2.873169
Control Factors			
Work	0 = no; 1 = yes	0.4225061	0.4939613
Commute Morning	0 = no; 1 = yes	0.1166361	0.3209882
Commute Evening	0 = no; 1 = yes	0.1086546	0.3112074
Evening	0 = no; 1 = yes	0.25561	0.4362063
Night	0 = no; 1 = yes	0.0965932	0.295405
Sunday	0 = no; 1 = yes	0.1052887	0.306927
Monday	0 = no; 1 = yes	0.088089	0.2834262
Tuesday	0 = no; 1 = yes	0.0919013	0.2888884
Wednesday	0 = no; 1 = yes	0.1459863	0.3530948
Thursday	0 = no; 1 = yes	0.2582492	0.4376745
Friday	0 = no; 1 = yes	0.218444	0.4131929
Saturday	0 = no; 1 = yes	0.0920415	0.2890864

Table 3: Descriptive Statistics of Event 3 (N=55,298)

VARIABLES	CODING	MEAN	STD. DEV
Dependent Variables			
Size (retweets)	Range: 0-191	0.1843466	2.29645
Survival (seconds)	Range: 0-1395227	2073.954	26842.71
Antagonistic Speech	0 = no; 1 = yes	0.0074144	0.0857877
Independent Variables			
Content Factors			
Sentiment	-1 = negative; 0 = neutral; 1 = positive	-0.4014069	0.7193844
URL	0 = no; 1 = yes	0.5159499	0.4997501
Hashtag	0 = no; 1 = yes	0.1651597	0.371328
Social Factors			
MPs	0 = no; 1 = yes	0.0007414	0.0272195
Jewish Org. and Media	0 = no; 1 = yes	0.0039423	0.0626642
Antisemitic Agent	0 = no; 1 = yes	0.004105	0.0639395
Media Agent	0 = no; 1 = yes	0.0837824	0.2770637
Other Agent	0 = no; 1 = yes	0.9074288	0.2898332
Verified	0 = no; 1 = yes	0.0199284	0.1397555
Number of Followers	Categorised into 1-10th percentiles	5.497649	2.873597
Control Factors			
Work	0 = no; 1 = yes	0.3803754	0.4854835
Commute Morning	0 = no; 1 = yes	0.1157546	0.3199334
Commute Evening	0 = no; 1 = yes	0.1172918	0.3217705
Evening	0 = no; 1 = yes	0.2668451	0.4423147
Night	0 = no; 1 = yes	0.1197331	0.3246521
Sunday	0 = no; 1 = yes	0.19413	0.395533
Monday	0 = no; 1 = yes	0.1291909	0.3354142
Tuesday	0 = no; 1 = yes	0.1176896	0.3222431
Wednesday	0 = no; 1 = yes	0.1253933	0.3311674
Thursday	0 = no; 1 = yes	0.1253572	0.3311264
Friday	0 = no; 1 = yes	0.1611632	0.3676847
Saturday	0 = no; 1 = yes	0.1470758	0.3541847

Table 4: Generalised Ordered Logit Regression Predicting Production of Antagonistic Content

	EVENT 1		EVENT 2		EVENT 3	
Content Factors	Odds Ratio	Std. Err.	Odds Ratio	Std. Err.	Odds Ratio	Std. Err.
Retweet Count	0.781	0.074	0.963	0.039	0.849	0.109
Sentiment	0.685	0.039	0.651	0.042	0.712	0.055
URL	0.493	0.038	0.586	0.048	0.597	0.064
Hashtag	0.806	0.079	0.843	0.092	0.685	0.112
Social Factors						
MPs	1.000	(empty)	1.000	(empty)	1.000	(empty)
Jewish Org. and Media	1.000	(empty)	0.493	0.497	1.000	(empty)
Antisemitic Agent	1.201	0.545	1.451	0.453	1.536	0.904
Media Agent	0.644	0.101	0.537	0.106	0.305	0.104
Ref: Other Agent						
Verified	0.826	0.259	0.502	0.231	0.188	0.189
Number of Followers	0.978	0.013	1.002	0.014	1.028	0.019
Control Factors						
Work	0.600	0.072	0.868	0.120	0.710	0.107
Commute Evening	0.677	0.101	1.267	0.200	0.835	0.156
Commute Morning	0.602	0.088	0.977	0.165	0.720	0.143
Evening	0.607	0.078	1.199	0.168	0.737	0.116
Ref: Night						
Sunday	1.047	0.150	0.826	0.130	1.065	0.193
Monday	1.292	0.186	0.891	0.146	0.918	0.188
Tuesday	1.131	0.179	0.797	0.134	1.212	0.240
Thursday	0.907	0.113	0.854	0.109	1.310	0.249
Friday	1.144	0.146	1.136	0.141	1.113	0.210
Saturday	1.047	0.148	0.900	0.143	0.759	0.157
Ref: Wednesday (mid-week)						
Constant	0.011	0.002	0.010	0.002	0.010	0.002
Model fit						
Log Likelihood	-4790.623		-3885.211		-2360.954	
Chi-square	235.87		171.79		112.73	
Sig	p=0.00		p=0.00		p=0.00	
N	155,566		78,352		55,039	

Table 5: Zero-Inflated Negative Binomial Regression Predicting Information Flow Size (Size Models)

	EVENT 1		EVENT 2		EVENT 3	
Content Factors	IRR	Std. Err.	IRR	Std. Err.	IRR	Std. Err.
Antagonistic Speech	0.285	0.069	0.510	0.124	0.441	0.144
Sentiment	0.996	0.018	0.758	0.022	0.990	0.033
URL	1.942	0.054	2.319	0.100	2.570	0.132
Hashtag	0.806	0.027	0.743	0.039	0.788	0.049
Social Factors						
Verified	5.004	0.269	7.295	0.727	6.750	0.784
MPs	1.500	0.311	5.916	2.525	0.817	0.464
Jewish Org. and Media	1.241	0.193	1.237	0.273	0.841	0.249
Antisemitic Agent	0.890	0.114	1.038	0.268	0.889	0.259
Media Agent	1.242	0.049	1.263	0.085	1.132	0.090
Ref: Other Agent						
Control Factors						
Commute Morning	0.969	0.038	0.949	0.066	1.427	0.108
Evening	0.970	0.031	0.701	0.036	0.984	0.056
Night	0.561	0.032	0.617	0.049	0.561	0.047
Sunday	1.197	0.061	0.947	0.081	1.786	0.155
Monday	0.970	0.053	1.066	0.099	1.276	0.120
Tuesday	0.877	0.050	1.038	0.091	0.778	0.077
Thursday	1.444	0.061	1.869	0.128	1.142	0.110
Friday	1.093	0.051	1.130	0.079	0.947	0.086
Saturday	1.172	0.059	1.537	0.135	1.167	0.108
Ref: Wednesday (mid-week)						
Constant	0.340	0.017	0.266	0.020	0.173	0.017
Binomial model (Inflation/Excess Zeros)						
Number of Followers	-0.442	0.009	-0.419	0.013	-0.343	0.014
Constant	0.011	0.002	0.010	0.002	0.010	0.002
Model fit						
Log Likelihood	-62519.62		-28983.15		-17194	
Chi-square	2882.14		1905.2		1038.26	
Sig	p=0.00		p=0.00		p=0.00	
LRT for alpha= 0	p=0.00		p=0.00		p=0.00	
Vuong	z=35.89, p=0.00		z=22.32, p=0.00		z=14.88, p=0.00	
N	156,498		78,432		55,298	

Table 6: Cox Regression Predicting Hazards to Information Flow Survival (Survival Models)

	EVENT 1		EVENT 2		EVENT 3	
Content Factors	Haz. Ratio	Std. Err.	Haz. Ratio	Std. Err.	Haz. Ratio	Std. Err.
Antagonistic Speech	1.379	0.282	1.111	0.200	1.662	0.463
Sentiment	1.036	0.013	1.005	0.019	0.974	0.024
URL	0.941	0.020	0.886	0.028	0.924	0.039
Hashtag	0.885	0.021	0.709	0.026	0.803	0.039
Retweet Count	0.959	0.002	0.987	0.001	0.955	0.004
Social Factors						
Number of Followers	0.980	0.004	0.970	0.006	1.012	0.007
Verified	0.974	0.032	0.863	0.044	0.965	0.067
MPs	0.862	0.100	1.346	0.246	1.427	0.487
Jewish Org. and Media	0.699	0.064	0.704	0.079	0.767	0.129
Antisemitic Agent	0.887	0.068	1.237	0.171	0.850	0.153
Media Agent	1.049	0.029	1.014	0.044	1.113	0.064
Ref: Other Agent						
Control Factors						
Commute Morning	0.974	0.027	0.855	0.038	0.993	0.055
Evening	0.941	0.021	1.001	0.034	0.983	0.041
Night	0.734	0.033	0.800	0.050	0.771	0.050
Sunday	1.082	0.039	1.084	0.063	1.092	0.068
Monday	1.051	0.039	1.005	0.063	0.913	0.063
Tuesday	1.077	0.043	1.037	0.062	1.110	0.082
Thursday	1.099	0.032	1.168	0.053	1.099	0.079
Friday	1.005	0.032	1.085	0.051	0.968	0.064
Saturday	0.995	0.035	0.971	0.058	1.036	0.070
Ref: Wednesday (mid-week)						
Model fit						
Log Likelihood	-101741.56		-41331.34		-23455.95	
Chi-square	1396.32		484.26		278	
Sig	p=0.00		p=0.00		p=0.00	
N	156,498		78,432		55,298	

APPENDIX 2: MACHINE CLASSIFICATION RESULTS

Four thousand tweets were systematically sampled from the complete dataset and CrowdFlower was used to source human annotators to perform annotation tasks on each tweet to determine, in their view, whether it was antagonistic in relation to Jews. Four annotators per tweet were required and those with agreement scores over 75 per cent (3 out of 4) were selected for the 'gold standard' training dataset for machine learning. The training dataset included 853 human-validated texts, where 388 instances were annotated as antagonistic towards Jews and 465 were annotated as non-antagonistic. Human annotations were checked against the text sample of offensive online communications provided by CST, and adjustments were made where misclassifications were identified.

In preparation for machine classification, the original text was transformed into feature vectors by using three feature extraction (FE) methods: Bag of Words (BOW), N-Grams (NG) and Typed Dependencies (TD). Four machine learning methods were used for training classifiers to identify antagonistic content about Jews: Decision Trees (DT), Naïve Bayes (NB), Support Vector Machine (SVM) and Fuzzy Rules.

The results of the classification experiments are provided using standard text classification measures of: precision (P) (i.e., for class x, how often are tweets classified as x when they should not be (false positives) - a measure of true positives normalised by the sum of true and false positives); recall (R) (i.e., for class x, how often are tweets not classified as x when they should be (false negatives) - a measure of true positives normalised by the sum of true positives and false negatives); and F-Measure (F), a harmonised mean of precision and recall. The results for each measure range between 0 (worst) and 1 (best). We provide results for the hateful class (Yes), non-hateful class (No) and overall (average over Yes/No).

Table 7: Classification Results for Antisemitism Hate Speech on CrowdFlower Annotations – 10 Fold CV

		DT			NB			SVM			FUZZY		
FE		P	R	F	P	R	F	P	R	F	P	R	F
BOW	No	0.66	0.74	0.69	0.56	0.95	0.71	0.67	0.78	0.72	0.61	0.79	0.69
	Yes	0.63	0.54	0.58	0.66	0.11	0.19	0.67	0.53	0.59	0.64	0.41	0.51
Overall		0.64	0.65	0.64	0.61	0.57	0.47	0.67	0.67	**0.66**	0.62	0.62	0.61
NG	No	0.58	0.76	0.66	0.55	1	0.71	0.62	0.79	0.69	0.59	0.82	0.69
	Yes	0.55	0.35	0.43	0	0	0	0.63	0.41	0.50	0.59	0.32	0.41
Overall		0.57	0.56	0.55	0.29	0.55	0.39	0.63	0.63	0.61	0.59	0.59	0.57
TD	No	0.55	0.97	0.70	0.55	1	0.71	0.55	0.97	0.69	0.55	0.96	0.69
	Yes	0.571	0.04	0.08	0	0	0	0.46	0.03	0.06	0.55	0.06	0.12
Overall		0.558	0.55	0.42	0.29	0.55	0.39	0.51	0.55	0.41	0.55	0.56	0.434

Note: P=Precision; R=Recall; F=F-Measure

Initially we used a 10-fold cross validation approach to test the supervised machine learning method. This functions by splitting the dataset into ten equal randomly shuffled subsets and iteratively using nine folds to train the classifier and one fold to test it. After ten iterations the results are averaged. It is particularly useful with small labelled datasets as was the case in this instance.

Table 7 shows that SVM + BOW performs the best. The high performance of SVM + BOW is likely due to the case that the SVM algorithm only needs a small number of instances as support vectors for teaching a classifier (identifying the boundary to separate the two classes in multi-dimensional feature space). As the dataset is small it is likely that features such as words are more effective as they will occur in each class more frequently than bigrams, trigrams and typed dependencies.

We experimented further using a 70/30 split on the data to train and test the supervised machine learning method. This functions by training the classifier with features from 70 per cent of the manually coded dataset, and classifying the remaining 30 per cent as 'unseen' data, based on the features evident in the cases it has encountered. The accuracy of the classification process is then determined. This process was repeated five times using the mean average of all runs to calculate the overall accuracy.

Table 8 shows only the results for the 'Yes' class (hateful language), and that SVM + BOW performs best again – this time with perfect classification, while the performance of the other methods is much lower. Again, the high performance of SVM + BOW is likely due to the SVM algorithm needing only a small number of instances as support vectors for teaching a classifier. With the small sample size, exposing the classifier to more examples of hate speech in the training process improves its ability to learn generalised word use which has led to an exact match between human and machine annotated labels for the hateful class. In other cases, such as decision trees and probabilistic approaches such as the NB method, more data actually causes further confusion – exemplifying the difficulty in using highly frequent words extracted from short informal text as features, with such a small 'gold standard' dataset.

Table 8: Classification Results for Antisemitism Hate Speech on CrowdFlower Annotations – 70/30 Split

	DT			NB			SVM			FUZZY		
FE	P	R	F	P	R	F	P	R	F	P	R	F
BOW	0.68	0.61	0.64	0.77	0.09	0.16	1	1	1	0.53	0.49	0.51
NG	0	0	0	1	0.33	0.5	0.33	0.67	0.44	1	0.4	0.57
TD	0.75	0.05	0.1	0	0	0	0.9	0.08	0.14	0.58	0.13	0.21

Note: P=Precision; R=Recall; F=F-Measure

WEBSITE **www.cst.org.uk**

FACEBOOK **Community Security Trust**

TWITTER **@CST_UK** YOUTUBE **CSTmedia**

☎ NATIONAL EMERGENCY NUMBER (24-hour) **0800 032 3263**
LONDON (Head Office) **020 8457 9999**
MANCHESTER (Northern Regional Office) **0161 792 6666**

Published by Community Security Trust. Registered charity in England and Wales (1042391) and Scotland (SC043612)