SI: Re-Imagining a More Trustworthy Social Media Future

SM+S
social media + society

Social Media + Society
April-June 2020: 1–20

DOI: 10.1177/2056305120916850
journals.sagepub.com/home/sms
SAGE

Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech

Sefa Ozalp, Matthew L. Williams, Pete Burnap, Han Liu, and Mohamed Mostafa

Abstract

In this article, we conduct a comprehensive study of online antagonistic content related to Jewish identity posted on Twitter between October 2015 and October 2016 by UK-based users. We trained a scalable supervised machine learning classifier to identify antisemitic content to reveal patterns of online antisemitism perpetration at the source. We built statistical models to analyze the inhibiting and enabling factors of the size (number of retweets) and survival (duration of retweets) of information flows in addition to the production of online antagonistic content. Despite observing high temporal variability, we found that only a small proportion (0.7%) of the content was antagonistic. We also found that antagonistic content was less likely to disseminate in size or survive for a longer period. Information flows from antisemitic agents on Twitter gained less traction, while information flows emanating from capable and willing counter-speech actors—that is, Jewish organizations—had a significantly higher size and survival rates. This study is the first to demonstrate that Sampson's classic sociological concept of collective efficacy can be observed on social media (SM). Our findings suggest that when organizations aiming to counter harmful narratives become active on SM platforms, their messages propagate further and achieve greater longevity than antagonistic messages. On SM, counter-speech posted by credible, capable and willing actors can be an effective measure to prevent harmful narratives. Based on our findings, we underline the value of the work by community organizations in reducing the propagation of cyberhate and increasing trust in SM platforms.

Keywords

antisemitism, collective efficacy, hate speech, social media, twitter

Introduction

With the increase in the use of digital communication technologies such as social media networks, online hate speech has become an increasingly prevalent and visible problem which threatens cohesion and trust among online citizens, and hence, their ability to work together to control their environment, what social scientists have term "collective efficacy" (Sampson et al., 1997). Social media acts as a polarization amplifier—it opens up a potential space for the galvanizing of attitudes and emotions, through the spread of negative expression toward minority groups and counter-narratives accelerated by algorithm driven partisan network contagion (Sunstein, 2017). Over the past decade, social media has become a safe harbor for launching campaigns of antisemitism, including harassment and criminal threats directed at members of the Jewish community. In the first 6 months of 2019, Community Security Trust (CST, 2019), a charity that supports the Jewish community, recorded 323 online antisemitic incidents in the United Kingdom, representing 36% of all incidents. This represents an increase of 46% on the same period the year before. Understandably, antisemitism on social media has become a matter of concern in the Jewish community and in broader public debate. Although conventional hate crime recording (i.e., police crime records and Crime Survey of England and Wales), has been improving, both online and offline antisemitic incidents are significantly under-reported, leading to a significant dark figure. Unlike

Cardiff University, UK

Corresponding Author:
Sefa Ozalp, Cardiff University, Cardiff CF10 3WT, UK.
Email: ozalpas@cardiff.ac.uk

previous research that has aimed to outline patterns of online antisemitism (e.g., Anti-Defamation League [ADL], 2018; Finkelstein et al., 2018; Woolley & Joseff, 2019), this article illustrates a scalable methodology that can identify future antisemitic communications and reveal patterns of online antisemitic perpetration at source. Furthermore, this article addresses the "collective efficacy" phenomenon on social media in the case of controlling antisemitic communications.

In this article, we present an analysis of the production and propagation of online antagonistic content targeting Jewish people posted on Twitter between October 2015 and October 2016 in the United Kingdom. We collect data from Twitter's streaming API using keywords which explicitly make reference to Jewish people and/or to Jewish identity and locate 2.7 million tweets from UK-based users. Drawing on emerging computational criminology methods, we train a machine learning algorithm to classify antisemitic content on Twitter with high accuracy and at scale (Burnap & Williams, 2015, 2016; Williams & Burnap, 2016). After illustrating significant variability in the frequency of antagonistic tweets related to Jewish identity, we identify three "spikes" in antagonistic content, the highest of which follows the suspension of MPs from the Labour party over antisemitism allegations. We then examine these three spikes by building statistical models around 15-day study windows. We model Twitter information flows (retweets) and explore (1) the inhibiting and enabling factors of online antisemitism, (2) the propagation of antisemitic content in terms of size (number of retweets) and survival (duration of retweets), and (3) the types of actors (e.g., Jewish organizations, antisemitic actors, media agents, MPs) that gain significant information flow traction. This article contributes to academic literature in the following three distinct ways: it introduces a supervised machine learning model capable of identifying future antisemitic incidents, it reveals patterns of online antisemitism perpetration at source, and for the first time, it introduces collective efficacy as a useful concept for interpreting the countering of online hate in a social media context.

Literature Review

Hate Crimes, Social Media, Cyberhate

Hate crimes have the potential to damage the fabric of trust between communities within society by undermining social cohesion. Current literature underlines the importance of the social polarization behind the mechanics of hate crime victimization. Gerstenfeld (2017) argues that the motivation behind hate crimes is not necessarily the hate directed toward the individual victim but rather the victim's perceived "outgroup" status. Complementing this view, Perry (2001, p. 5) explains that hate crimes aim to polarize communities by sending "messages" to the wider community of the "others" that they must "conform to the standards" set by the privileged majority. From a broad societal point of view, fluctuations with regard to polarization can be observed through hate crime statistics. Studying hate crime figures from conventional quantitative data sources such as police crime records and self-report studies (e.g., victimization and crime surveys) may prove to be beneficial in order to understand the patterns of divisive tensions within a society, provided that biases attached to these data sources are carefully considered when drawing conclusions.

Data from conventional sources suggest that hate crime is on the rise in England and Wales. According to the most recent records, hate crimes recorded by the police in England and Wales have increased by 17%, from 80,393 (2016/2017) to 94,098 (2017/2018) (Home Office, 2018). The upward trend in police-recorded hate crime has been seen since 2012/2013. Figures have more than doubled (123%) in England and Wales with an increase from 42,255 (2012/2013) to 94,098 (2017/2018). Although these figures are important barometer of societal tensions between groups, criminologists have long argued that the statistics produced by police are insufficient to paint a complete picture to understand both general and hate crime patterns. Existing criminological literature illustrates a number of limitations of police recorded crime data such as non-uniform recording practices across police forces, improvements, and changes in police recording practices over time, and changes in legislation and classification of offense types (Maguire, 2007; Tilley & Tseloni, 2016). In relation to hate crimes, these data sources are incomplete as at least half of all hate victims do not report their victimization (Williams & Tregidga, 2014). A recent Home Office (2018) report recognizes some of the shortcomings of police recorded hate crime figures, suggesting that the increases in recent years are "largely driven by improvements in police recording" (p. 7).

Another useful conventional data source to understand hate crime victimization figures is the Crime Survey of England and Wales (CSEW). Surveying a nationally representative sample of roughly 35,000 households each year, the CSEW is regarded "as a gold-standard survey of its kind" (Flatley, 2014, p. 199). Recent estimates from the CSEW show that racial and religious aggravated hate crimes increased by 4.5%, from 112,000 per year (2013–2015, two-year average) to 117,000 per year (2015–2017, two-year average). Combined estimates suggests that there were 184,000 hate incidents per year from 2015/2016 to 2017/2018 (Home Office, 2018, p. 7). Despite the robust nature of CSEW statistics, they are limited by their reliance upon victim interviews. Some victims of hate incidents might not be willing to report hate crime in victimization surveys. For instance, the wording (i.e., using the term "hate crime") of questions in surveys can be problematic. Williams and Tregidga (2014, p. 948) found that while some survey respondents may find the word "hate" too restrictive, others may be confused by the word "crime," hesitating "whether their experiences constituted acts serious enough to be classified as crimes". Correspondingly, some victims prefer not

to report the prejudiced incidents they experienced either to the police or in surveys, leading to dark figures in hate crime victimization rates.

Given the shortcomings of conventional police hate crime and victimization data, it is important to supplement these with other sources to paint a more complete picture. New data sources, such as internet searches and social media communications, lend themselves well to the analysis of public sentiment trends. Recent computational and social science advances in machine learning and statistical modeling allow researchers to utilize new "big data" sources to address a variety of social research questions, such as tracking the spread of influenza (Ginsberg et al., 2009) or to build psychological constructs of nations linked to GDP (Noguchi et al., 2014). Furthermore, 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., 2017). Besides conventional hate crime statistics, othering and divisive sentiment trends within society can also manifest in subtler forms, such as prejudiced online communications. Referred to as cyberhate, these divisive and prejudiced online communications have been present since the dawn of the public internet in the 1990s (Wall & Williams, 2007; Williams, 2006). Similar to offline hate crimes, the motivation of cyberhate perpetrators is rarely the hate of individual victims, but the community of "others" in which they represent (Douglas et al., 2005). Previous cyberhate literature illustrates that perpetrators target victims because of their perceived belonging to groups with protected characteristics such as sexual orientation (McKenna & Bargh, 1998), race (Leets, 2001), and religion (Williams & Burnap, 2016). By analyzing prejudiced online communications, we can identify the ebb and flow of societal tensions through the monitoring of subtler "hate incidents," many of which would not reach the criminal threshold used by law enforcement agencies, and therefore, would not be included in conventional hate crime statistics. Therefore, current researchers and practitioners should take advantage of the affordances provided by online communications data and supplement conventional statistics with cyberhate perpetration in order to shed light on "dark" figures of hate crime victimization trends.

Collective Efficacy and Social Media

Social media companies have generally presented themselves as strong advocates of free-speech and have until very recently allowed hate speech to proliferate on their platforms. Online hate speech has become an increasing problem that to date has been largely controlled by online community cooperation, what social scientists term "collective efficacy" (Sampson et al., 1997, p. 918). Sampson (2001) describes collective efficacy as "the linkage of mutual trust and shared willingness and intention to intervene for the common good" (p. 95). On social media platforms, an abundance of cyberhate speech in the absence of capable and willing counter-speech actors can reduce collective efficacy which, in turn, can result in decreased trust in platforms, their users, and online communities. On the contrary, if capable, trustworthy and willing actors on social media platforms can successfully intervene cyberhate perpetrators with counter-speech, we can observe the benefits of online collective efficacy. Current research on online collective efficacy is scarce. In a demographically balanced survey of Americans, Costello et al. (2017) explored the presence of online collective efficacy and found that 21.3% of respondents reported that they observe others telling perpetrators of cyberhate to stop, and 21% indicated that they witnessed others defending victims of cyberhate. However, their logistic regression model failed to demonstrate a statistically significant association—neither positive nor negative—between either form of collective efficacy and being targeted by cyberhate.[1] Therefore, unlike the long-proven negative correlation between the perception of collective efficacy in offline communities and offline crime rates (Mazerolle et al., 2010; Sampson et al., 1997), the effectiveness of collective efficacy on social media platforms is yet to be proven in the literature.

Data from social media platforms can be utilized to explore the effectiveness of collective efficacy on online communities. Social media communications can be amplified and redistributed through platform-specific dissemination mechanisms such as retweeting (Boyd et al., 2010). This unique conversational aspect of online communications enables researchers to study online information propagation networks and information flows. Unlike traditional methods, such as surveys or interviews, through studying information flows through retweets, researchers can "identify what information or sentiment is being endorsed and propagated by users, and which users have the most or least influence in the spread of such messages" (Williams & Burnap, 2016, p. 215). By comparing the retweet rates of trustworthy and capable users engaged in counter-speech practices to rates of retweets of biased and prejudiced users engaged in spreading divisive messages and cyberhate, arguably researchers can measure a proxy of collective efficacy in online communities.

Related Work: Offline and Online Antisemitism

In this article, we focus our attention solely on the growing problem of online antisemitism in the United Kingdom, which is an important policy and community safety issue. In a survey conducted by the European Union Agency for Fundamental Rights (FRA) among individuals who consider themselves Jewish in eight European countries, including the United Kingdom, 75% ($n = 5{,}847$) stated that they consider online antisemitism as a problem (European Union Agency for Fundamental Rights, 2013, p. 12). In addition, 75% of the respondents who were exposed to negative statements toward Jews ($n = 5{,}385$), cited the internet as the medium that

exposed them to negative sentiments (European Union Agency for Fundamental Rights, 2013, p. 25).[2] Of those who were exposed to antisemitic harassment ($n = 1{,}941$), which can be both online and offline, only 23% stated that the incident was reported to the police, to another organization or both, while 76% stated that the event was not reported at all (European Union Agency for Fundamental Rights, 2013, p. 49). Given the staggering rates of non-reported antisemitic victimization and the growing concerns about online antisemitism, FRA suggested that the "EU Member States should consider establishing specialised police units that monitor and investigate hate crime on the internet and put in place measures to encourage users to report any antisemitic content they detect to the police" (European Union Agency for Fundamental Rights, 2013, p. 12). To our knowledge, there are no specialised units dedicated to tracking online antisemitism at the source in any EU states to date.

Previous research on detecting and analyzing online antisemitic incidents at the source is of particular interest to this study. Analyzing over 100M posts from multiple social media platform hosting "fringe" communities, 4 chan's Politically Incorrect board (/pol/) and Reddit's The_Donald subreddit and Gab, Finkelstein et al. (2018) argued that online antisemitism and racist online communications increased considerably following divisive offline political events such as the 2016 US election.[3] By training word2vec models, they devised a text-based methodology which predicts similar words that are likely to appear together in the same context. Although useful for exploring keyword-based discussions of online fringe communities, the unsupervised nature of this methodology limits its practicality for classifying future individual antagonistic instances. Barring this article, most of the existing research on the detection of online antisemitism are commissioned or conducted by Jewish civil society organizations such as the Anti-Defamation League (ADL) and Community Security Trust (CST).[4] From January 2017 to January 2018, ADL collected more than 18 million tweets using keywords referring to Jews and Jewish identity (ADL, 2018). By randomly sampling 1,000 tweets per week that matched with a complex Boolean query and manually annotating $n = 55{,}000$, ADL predicted 4.2 million tweets (23.5% of all tweets collected) were antisemitic within the study period. In another mixed-methods study on Twitter, Woolley and Joseff (2019) explored antisemitism among 5.8 million tweets containing political hashtags during the 2018 US midterm election campaign. Human annotation of 99,075 filtered tweets revealed that 54.1% contained antisemitic conspiracy theories and 46.45% contained derogatory terms.[5] Although these three studies are important to understanding trends in online antisemitic sentiment, none detail the accuracy of the content classification results or provide a discussion of common information retrieval metrics such as precision, recall, and F-measure. Finally, none of these studies suggests a methodology to accurately identify future antisemitic incidents without human annotation.

Given the limitations of current research, new research on the automated detection of antisemitic cyberhate and the statistical dynamics of its propagation is needed. Instead of relying on conventional "terrestrial" data sources, this article reveals patterns of online antisemitic perpetration at source. Although there are multiple social media platforms where antisemitism can be traced, we exclusively draw on Twitter data due to the ease of access and the ability to explore information propagation networks through the retweeting mechanism. Following a human annotation phase, we trained a supervised machine learning classifier that is capable of classifying antisemitic content at scale. Informed by the collective efficacy theory, our hypotheses address the enablers and inhibitors of antisemitic content within UK-based Twitter communications.

Hypotheses

H1: Offline events and discussions concerning Jews will act as "trigger events" and be observed as spikes in online communications related to Jewish identity.

The event-specific increase in hate crimes is an established phenomenon in the literature. For instance, Hanes and Machin (2014) observed significant increases in hate crimes reported to the police in the United Kingdom following 9/11 and 7/7 terror attacks. Similarly, in the aftermath of Woolwich terror attack in 2013, Williams and Burnap (2016) observed a sudden spike and a rapid de-escalation in the frequency of racial and religious cyberhate speech within the first 48 hours of the attack. Findings from these studies indicate that galvanizing "trigger" events such as socially divisive political events and terror attacks motivate prejudiced incidents against outgroups and lead to an increase in incidents targeting minorities, which is reflected in hate crime statistics. Informed by previous research, we hypothesize that offline events that trigger debate around Jewish identity will migrate to social media.

H2: Pre-identified antisemitic Twitter users will be positively correlated with the production of antagonistic content about Jews.

The second hypothesis tests whether Twitter users flagged by Jewish civil society organizations as antisemitic due to their previous online behavior are predictive of cyberhate production.

H3: Trustworthy and capable actors will be positively associated with larger *size of information flows*.

H4: Trustworthy and capable actors will be positively associated with longer *survival of information flows*.

H3 and H4 operationalize collective efficacy theory on Twitter. Sampson (2001) underlines the importance of "mutual

trust and shared willingness" for the capable and trustworthy actors who are willing to "intervene for the common good." On Twitter, by comparing the retweet rates of trustworthy and capable users to rates of retweets of biased and prejudiced users engaged in spreading divisive messages and cyberhate, researchers can measure a proxy of the collective efficacy phenomenon. These hypotheses test whether trustworthy and capable agents, such as Jewish community organizations, verified accounts, official police accounts, and Members of Parliament (MPs), are associated with larger size and longer survival of information flows, lending evidence for the effectiveness of collective efficacy within the particular community of interest on Twitter.

H5: Antagonistic content about Jews will not propagate further in *size*, within the study period.

H6: Antagonistic content about Jews will not *survive* over time, within the study period.

H5 and H6 extend the previous research on computational criminology by exploring the propagation dimension of the antagonistic speech targeting Jews and Jewish identity (Williams & Burnap, 2016). Within the study period, if cyberhate does not propagate in *size* and if it does not *survive* over time, we can tentatively infer an association between collective efficacy and a reduction in the impact of the information flows containing antagonistic sentiments within the particular community of interest on Twitter. Informed by previous research, we assume that antisemitic tweets will be negatively associated with both size and survival of information flows.

Methodology

Data Collection and Preprocessing

The data used in this study 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. The following keywords were used for data collection: "jew, jewish, jews, antisemitic, antisemitic, antisemitism, antisemitism, anti semitic, anti semitism, bonehill, stamford hill, golders green, neo nazis, neo nazi, neonazi, neo-nazis, nazi, nazis."[6] These keywords are a combination of generic terms and terms relating to a far-right demonstration directed at the Jewish community in Golders Green in north London, reflecting events in the United Kingdom at the time of the data collection. 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 image-based hate speech—such as memes—that would not be captured by these keywords (Finkelstein et al., 2018; Woolley & Joseff, 2019). This caveat should be borne in mind when assessing the overall quantity of antagonistic content measured by this research and the generalizability of the findings. The data used for this analysis include tweets posted between 16 October 2015 and 21 October 2016 and were collected in real-time, ensuring all tweets matching with the keywords are collected. The raw dataset for the complete study period contained 31,282,472 tweets.[7] The dataset was imported into the R environment (R Core Team, 2018), which is an open-source statistical programming language, for preprocessing and exploratory data analysis (EDA).

The first aim of preprocessing was to infer the location of users from tweet metadata and extract UK-based tweets from the whole dataset. Unless Twitter users explicitly opt-in to share their geo-locations each time they post a tweet, latitude and longitude coordinates are not provided in the metadata. The majority of Twitter users in this dataset (>99%) opted out of sharing these exact geo-data. Three different approaches were adopted to infer Twitter communications from the United Kingdom, using the metadata of each tweet. First, we derived a list of UK-based place names (referenced as the UK pattern henceforth). Using pattern matching techniques, the UK pattern was identified within the account description of the users. Second, the UK pattern was identified within the user reported locations field (shown under profile pictures). Finally, 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.[8] The number of tweets identified as emanating from the United Kingdom in this study is therefore 8.5%. 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% of these accounts were located within the United Kingdom (Statista, 2019).

The second aim of preprocessing was to classify user types that were of interest for analysis. Using conventional data science methods and tweet metadata and collaborating with organizations with subject expertise, six user types of interest were identified, that is, Media Agents, MPs, Celebrity Agents, Police Agents, Jewish Organisation 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). In total, 181,363 tweets were identified as emanating from Media Agents.[9] Drawing on previous work by Bejda (2015), we used a list of the most followed celebrities on Twitter and by matching them with the users in the dataset, identifying 80 tweets from Celebrity Agents. To identify MP Agents, we used a list of Twitter handles of 590 MPs who served between the 2015 and 2017 general elections, identifying 2,950 tweets.[10] 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 unit). In total, 162 tweets were identified as emanating from Police Agents. To identify Jewish Organization agents, we pattern matched user descriptions against the terms "Jew," "Jewish", and "Jewry" and identified all Jewish organizations followed by @CST_UK. A resulting 102 Jewish organization agents were found in the UK dataset, generating 11,599 tweets in the study period. To identify known antisemitic agents, we used a pre-defined list of 24 accounts which was supplied by CST. In total, 13,240 tweets were identified as posted by these agents.[11] All other users that did not fall into any of these agent types were classified as "other" agents.

Tweet Classification

We devised a supervised machine learning methodology to classify antagonistic content related to Jews in the Twitter dataset. 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 (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.[12]

Given this complexity, a two-stage process to attaining gold standard, human annotation was performed to create a training dataset for the machine learning classifier (see Appendix A for a detailed discussion). The training dataset included 853 human-validated tweets, where 388 instances were annotated as antagonistic toward Jews and Jewish identity and 465 were annotated as non-antagonistic. This human-annotated dataset was used as the gold standard to train the machine learning classifier. We experimented with multiple supervised learning techniques when building the classifier. Both 10-fold cross-validation and 70/30 split validation results suggested that overall, the most efficient machine learning technique for classifying antagonistic content in this dataset was Support Vector Machines combined with a Bag of Words approach (see Appendix B for a detailed discussion of other algorithms we experimented with and Appendix C for the computational cost of the study). In total, this method identified 9,008 original tweets as antagonistic, representing 0.7% of the 1,232,744 original tweets in the UK dataset. This is commensurate with the volume of antagonistic tweets related to Muslim identity following terror attacks in the United Kingdom (0.9%; see Williams & Burnap, 2016). Upon inspection of the classification results, we are confident 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.

Exploratory Data Analysis and Descriptive Statistics

In the first stage of EDA, we visualized the UK dataset and the antagonistic sub-dataset to identify periods of interest to the next stage of analysis. The periods of interest were then isolated for statistical modeling to identify the enablers and inhibitors of the production of antagonistic content, and the factors that predict information flow *size* and *survival*. Figure 1 presents a daily aggregated time-series line graph of overall tweet frequency (black line) and antagonistic tweets (red 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 28 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 observation indicates offline events probably trigger online discussions that contain the keywords used in the collection, confirming both H1 and previous research (Hanes & Machin, 2014; Williams & Burnap, 2016). The Figure 1 also compares the volume of antagonistic tweets to all tweets using the same scale, illustrating their relatively 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 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 are in line with general discussions about Jews on Twitter. The EDA enabled us to visualize the temporal patterns of both antagonistic and non-antagonistic tweets in the UK dataset. As the primary aim of the analysis was to predict the enablers and inhibitors of the production of antagonistic content and of the propagation of information flows through statistical modeling, we selected three events of interest around the highest three peaks in Figure 2: Event 1 includes all tweets posted between 27 April 2016 and 13 May 2016, Event 2 includes all tweets posted between 15 June 2016 and 1 July 2016, and Event 3 includes all tweets posted between 12 August 2016 and 28 August 2016.

Information Propagation Models

Dependent Measures. There are two dependent measures in information propagation modeling: *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 within the study period). In

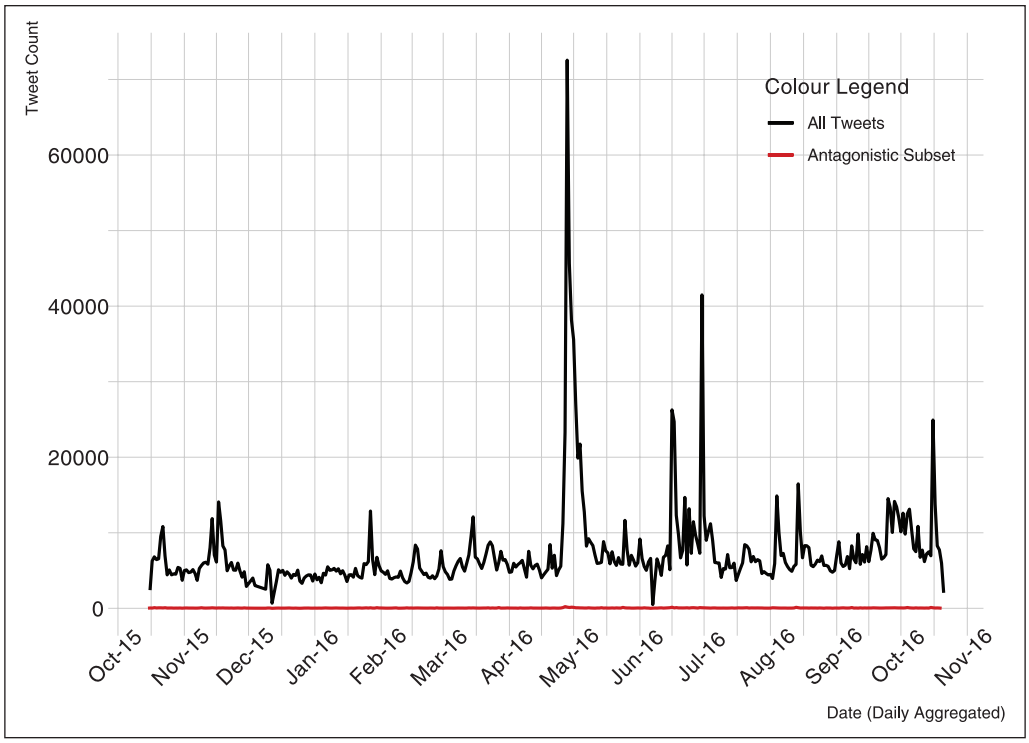

Figure 1. Tweet frequency (12 months).

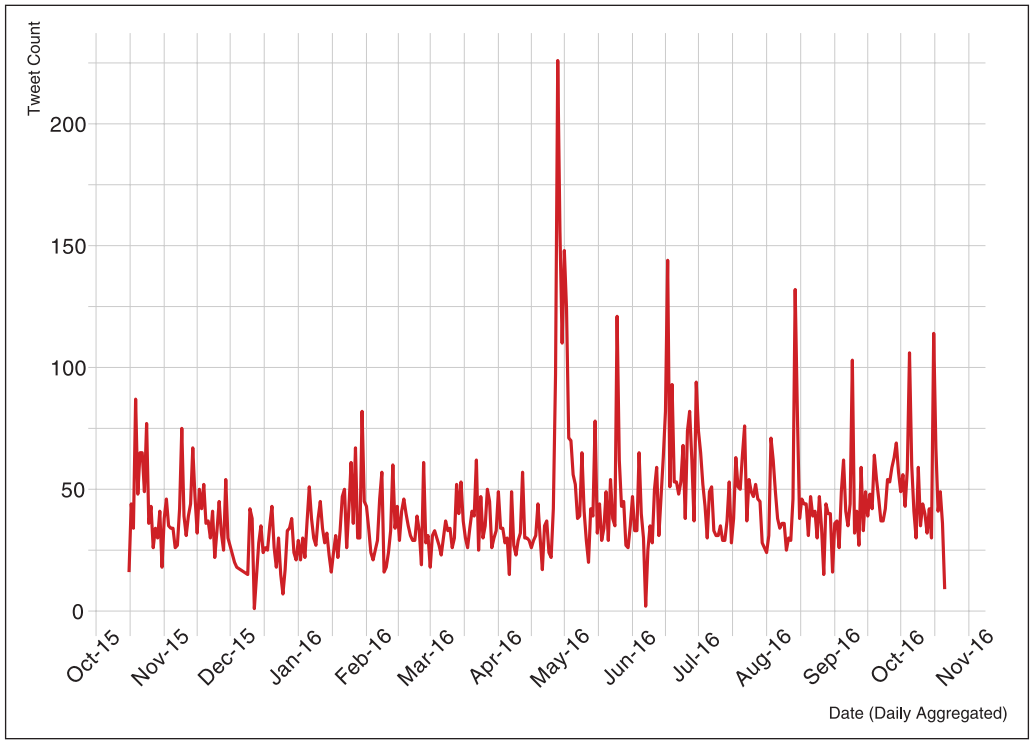

Figure 2. Antagonistic tweet frequency (12 months).

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 the 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).

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: 0–1,349,131	2,604.136	28,456.9
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	Categorized 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

Independent Measures. 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. The following text content factors 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, which is the outcome of our machine learning classifier. 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. The presence of police agents and celebrities were either extremely small or non- existent across all three events of interest. Therefore, police and celebrity agents were re-classified under other agents and five agent types were included: media, MPs, Jewish organizations, known antisemitic accounts, and other agents. 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. Tables 1 to 3 present descriptive statistics for each event we selected to model in this study.

Antagonistic Content Models

Dependent and Independent Measures. For predicting the production of content, which was antagonistic toward Jewish identity, we used the results of our machine learning classifier for the original text as the dependent variable. We converted classification results into a binary numeric format where "1" represents antagonistic content and "0" represents non-antagonistic content. For independent variables, we used the same independent variables (i.e., content factors, social factors, and control factors), as described in the information propagation models.

Methods of Estimation

Information Propagation Size Model. To predict the size of information flows, we use zero-inflated negative binomial (ZINB) regression.[13] We fit ZINB regression models as the *size* measure is best described as a count of retweets, where zeroes were present (i.e., some tweets were not retweeted during the study period). Zero-inflated count variables represent types of events that are largely not experienced by the majority

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

Variables	Coding	Mean	Std. Dev
Dependent variables			
Size (retweets)	Range: 0–1,550	0.3546384	8.780562
Survival (seconds)	Range: 0–1,407,814	2,170.419	31,109.54
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	Categorized 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

of the sample. In this case, it is 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 nonnormal distribution of the errors. We opted to use ZINB regression over zero-inflated Poisson (ZIP) regression because the dependent variable was overdispersed.

Information Propagation Survival Model. To predict the survival of the information flows, we used Cox's proportional hazards regression (1972). Our interest here was to model the factors that pose hazards to the survival of information flows, that is, duration of a retweet (in seconds) within the study period. Therefore, positive relationships indicate an increased hazard to survival of information flows.

Cyberhate Model. Since this variable was best described as binary (0 = non-antagonistic; 1 = antagonistic), we estimated the production of antagonistic content by using generalized ordered logit regression, which allows for the identification of predictive factors.

Results

Cyberhate Model

Results of cyberhate models for each event are presented in Table 4. Across all events, accounts identified as antisemitic by CST were most likely to produce antagonistic content related to Jews, lending strong evidence in support of H2. This is unsurprising given the nature of these accounts and their posting history. This finding also lends strong evidence in support of the semantic 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

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: 0–1,395,227	2,073.954	26,842.71
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	Categorized 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

public interest and authentic," were significantly less likely to be associated with antagonistic content, compared to non-verified accounts. Many of these accounts belong to celebrities, public figures, politicians, news organizations, charities, corporations, and government departments. Media Agents and, unsurprisingly, Jewish organizations 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 (Williams & Burnap, 2016), the inclusion of hashtags in tweets was negatively associated with the production of antagonistic content across the three events. This may suggest users publishing antisemitic content do not aim to increase the discoverability of their messages outside their follower networks.

Information Propagation Size Model

Table 5 presents the results of the *size* models. Sample size in each event only indicates 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.[14] Of particular note is the negative relationship between antagonistic content and the size of retweets. In all three events, antagonistic content did not propagate in terms of size (IRR: 0.285, 0.510, and 0.441, respectively), providing strong support for H5 and confirming 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.

Across all three events, content posted from Twitter verified accounts was most likely to be retweeted in volume, an

Table 4. Generalized Ordered Logit Regression Predicting Production of Antagonistic Content for Each Event.

	Event 1		Event 2		Event 3	
	Odds ratio	Std. err.	Odds ratio	Std. err.	Odds ratio	Std. err.
Content factors						
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	–4,790.623		–3,885.211		–2,360.954	
Chi-square	235.87		171.79		112.73	
Sig	0		0		0	
Pseudo R^2	0.024		0.0216		0.0233	
N	155,566		78,352		55,039	

Significance codes: *: 0.05; **: 0.01;***: 0.001.

unsurprising finding given the types of users behind these accounts. In all but one of the events (Event 3), MPs were highly likely to be retweeted. This pattern is repeated in relation to Jewish organizations, providing strong support for H3. Across all three events, Media Agents were positively associated with larger size of information flows, supporting previous research that indicates "old media" greatly influence the flow of information on "new media" platforms (Williams & Burnap, 2016).

Information Propagation Survival Model

Table 6 presents the results of the information flow survival models for the three events. Positive estimates in the Cox regression models are interpreted as increased hazards to survival and, therefore, a reduction in the duration of information flows on Twitter. 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. Supporting H6, this finding corroborates previous that shows online antisemitic hate speech does not propagate in terms of size or survival (Williams & Burnap, 2016). Figures 3, 5 and 7 visualize the survival estimates of antagonistic content in the 15-day analysis windows of each event. They show that these antisemitic information flows survived between 1 and 3 days. This sharp de-escalation once again lends evidence to H6 and resonates with research that shows offline hate crime following trigger events has a "half- life" (King & Sutton, 2013; Legewie, 2013). It seems likely that this offline pattern is replicated in relation to online antagonistic content concerning Jews.

Table 5. Zero-Inflated Negative Binomial Regression Predicting Counts of Retweets (Size Models).

	Event 1		Event 2		Event 3	
	IRR	Std. err.	IRR	Std. err.	IRR	Std. err.
Content factors						
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
Anti-semitic 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	2.987***	0.045	2.763***	0.067	2.396***	0.086
Model fit						
Log likelihood	–62,519.62		–28,983.15		–17,194	
Chi-square	2,882.14		1,905.2		1,038.26	
Sig.	$p = .00$		$p = .00$		$p = .00$	
LRT for alpha = 0	$p = .00$		$p = .00$		$p = .00$	
Vuong	$z = 35.89$, $p = .000$		$z = 22.32$, $p = .000$		$z = 14.88$, $p = .000$	
Zero obs	144,140		72,931		51,922	
Non-zero obs	12,358		5,501		3,376	
Total obs	156,498		78,432		55,298	

IRR: Incidence-Rate Ratio; LRT: Likelihood Ratio Test; MPs: Members of Parliament.
Significance codes: *: 0.05; **: 0.01;***: 0.001.

Figures 4, 6, and 8 visualize the survival estimates for different agent types. Unexpectedly, antisemitic agents emerged as having the fourth and fifth highest negative hazard ratios in Event 1 and Event 3. This indicates that information flows emanating from some of these agents during these events were likely to outlast those emanating from other agents at some points in the 15-day analysis windows. These figures show that, while information flows from antisemitic agents can last between 3 and 7 days, these are in a minority, as many of them die out rapidly (indicated by the steep decline in the red lines). Conversely, many more information flows emanating from Jewish organizations survive between 3 and 7 days in all events (indicated by a less steep decline in the green lines). This finding is novel and shows information flows from antisemitic agents gain less traction in terms of duration than flows produced by organizations challenging these negative narratives on social media. Furthermore, information flows emanating from Jewish organizations emerge as having the lowest hazard to survival across all events, strongly supporting H4.

General Media Agents emerged as having positive hazard ratios for all three events, with many information flows dying out evenly over the study window (the yellow line). As indicated in previous research, this is likely to be a result of

Table 6. Cox Regression Predicting Hazards to Tweet Survival (Survival Models).

	Event 1		Event 2		Event 3	
	Haz. ratio	Std. err.	Haz. ratio	Std. err.	Haz. ratio	Std. err.
Content factors						
Antagonistic speech	1.379	0.282	1.111	0.2	1.662	0.463
Sentiment	1.036**	0.013	1.005	0.019	0.974	0.024
URL	0.941**	0.02	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	0.1	1.346	0.246	1.427	0.487
Jewish org and media	0.699***	0.064	0.704**	0.079	0.767	0.129
Anti-semitic agents	0.887	0.068	1.237	0.171	0.85	0.153
Media agents	1.049	0.029	1.014	0.044	1.113	0.064
Ref: Other agents						
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***	0.05	0.771***	0.05
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	1.11	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	0.07
Ref: Wednesday (mid-week)						
Model fit						
Log likelihood	–10,1741.56		–41,331.34		–23,455.95	
Chi-square	1,396.32		484.26		278	
Sig.	0		0		0	
N	12,183		5,465		3,319	

Significance codes: *: 0.05; **: 0.01;***: 0.001.

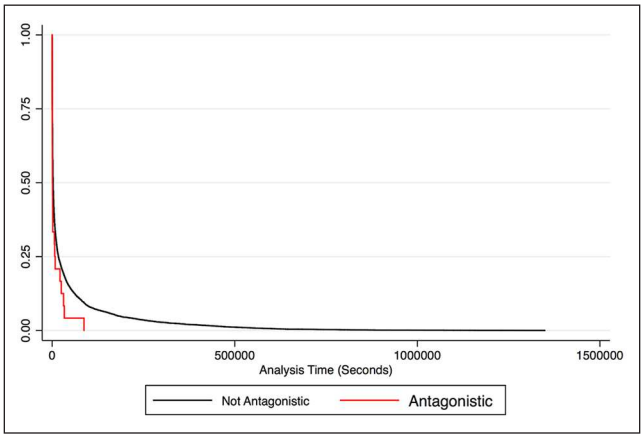

Figure 3. Kaplan-Meier survival estimates for antagonistic tweets in event 1.

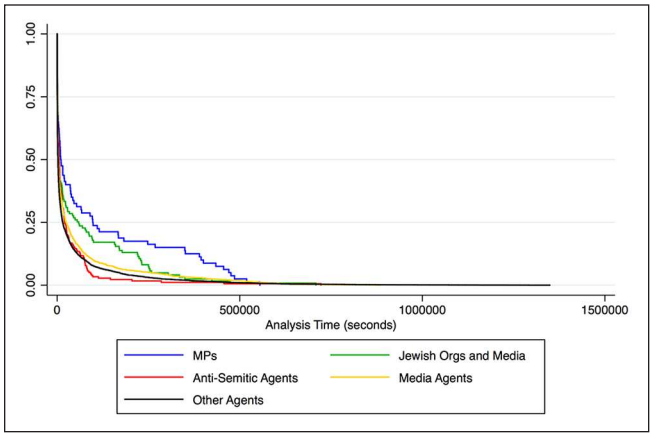

Figure 4. Kaplan-Meier survival estimates for agent type in event 1.

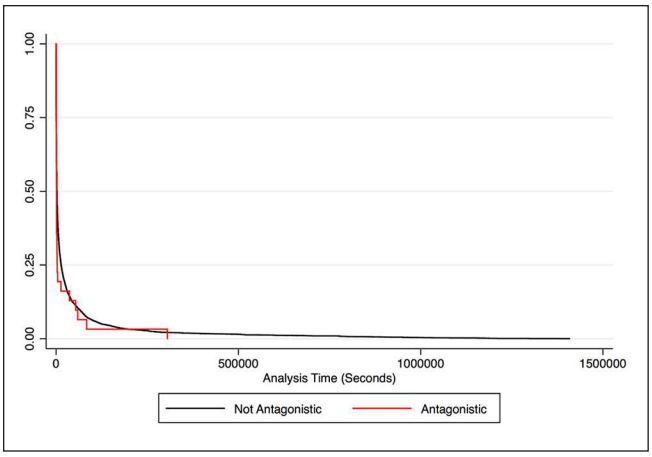

Figure 5. Kaplan-Meier survival estimates for antagonistic tweets in event 2.

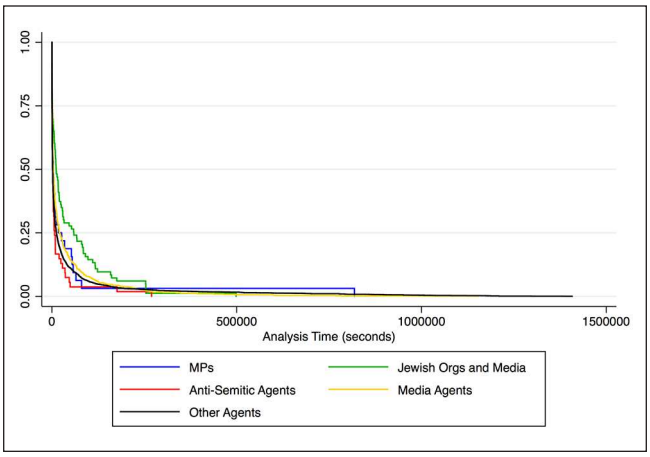

Figure 6. Kaplan-Meier survival estimates for agent type in event 2.

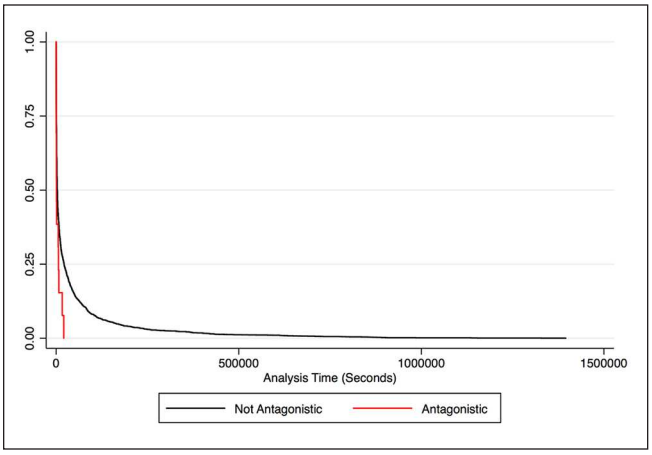

Figure 7. Kaplan-Meier survival estimates for antagonistic tweets in event 3.

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).

Discussion

In this article, we demonstrated how using social media (meta)data when coupled with computational criminology methods (i.e., pattern matching, supervised machine learning classification detecting cyberhate, information propagation modeling) can contribute to conventional hate crime recording practices and extend our understanding of online trends of antisemitism. Our analysis showed significant variability in the frequency of antagonistic tweets related to Jews over the 12-month study period. Supporting H1, we demonstrated offline events, such as the antisemitism row in the Labour party, can trigger online discussion around Jewish identity and antisemitic sentiments. The analysis also revealed the frequency of antagonistic content was on average 32% 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, 2017). Similar to previous research related to anti-Muslim sentiment on Twitter (Williams & Burnap, 2016), we found that only 0.7% of tweets referring to Jews and Jewish identity were classified as antagonistic. Although this finding contradicts with previous higher antisemitism rates of global tweets (ADL, 2018), it suggests that only a small proportion of the content relating to Jews on UK-based Twitter are antagonistic, confirming previous research (Williams & Burnap, 2016).

Across all three events subjected to statistical modeling, our logit model predicting the presence of hate speech suggests that accounts identified as antisemitic by CST were most likely to produce antagonistic content, while verified and media accounts were least likely. These findings lend strong support for the H2 and provide evidence in support of the accuracy of the machine learning classifier built for this study. H5 and H6 also demonstrated that antisemitic content was less 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 (Burnap et al., 2014; Williams & Burnap, 2016). Non-propagation in terms of size means that antagonistic content was not retweeted (shared by other Twitter users) to a great extent and most of the time none at all. This is an encouraging finding which indicates that the majority of Twitter users do not endorse these types of posts through the act of retweeting. Non-propagation of hate within the Twitter community might be interpreted as a demonstration of collective efficacy on Twitter. However, we would like to remain conservative with this claim as there may be other confounding factors. Research shows that where antagonistic content is retweeted, it is contained within online "echo chambers" of like-minded

individuals and if the size of this community is likely to affect the volume of information propagation.

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 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.

We also reported some positive results, particularly with regard to the representation of collective efficacy on social media (Sampson, 2001; Sampson et al., 1997). In support of H3 and H4, this study revealed that information flows emanating from Jewish organizations gained significant traction during two of the three events, as evidenced by the combined positive size and survival findings. We found that information flows from antisemitic agents on Twitter gain less traction in terms of duration than information flows produced by organizations challenging these negative narratives lending tentative support to the effectiveness of "collective efficacy" on social media. This suggests that when organizations which aim to counter harmful narratives such as antagonistic speech become active on social media platforms, their messages propagate further and achieve higher longevity than antagonistic messages. This is a positive finding that underlines the importance of the work of organizations that aim to protect communities and increase collective efficacy on social media.

Conclusion

Police crime and CSEW figures indicate that hate crimes have increased significantly in the past few years in the wake of the vote over the UK's future in the EU and recent terror attacks. Despite being useful, conventional hate crime recording practices are limited by their reliance on victims or witnesses reporting incident. Correspondingly, the FRA survey shows more than three-quarters of antisemitic harassment are never reported, leading to a dark figure in hate crime records. There is a clear policy and community safety need to devise new methodologies to detect and analyze online antisemitic incidents, as highlighted by the FRA. Given the sheer size of social media communications at any given hour, manually sifting through millions of posts every month to detect cyberhate would be extremely laborious, if even possible. Computational approaches without human input, such as unsupervised learning and clustering, are limited when detecting future instances of cyberhate. Instead of relying on "terrestrial" data or reports from the public on antisemitic victimization, this study used a relatively novel online source, Twitter, to mine big social media data to reveal patterns of perpetration at the source using a supervised machine learning classifier. By doing so, this study has demonstrated how a unique blend of computational and social science techniques can be harnessed to transform and analyze these new forms of data to gain insight into the growing problem of online antisemitism in the United Kingdom.

Findings from this study should be a source of some optimism. A key finding of this study is that information flows emanating from Jewish organizations, capable and willing counter-speech actors, had a significantly higher size and survival of retweets. While antisemitism is present on Twitter and can cause severe offense when it is not removed, it is challenged by positive content, which is present in greater amounts, lasts longer, and spreads further than hate content. Measuring the production of cyberhate, and the size and survival of information flows, this study is the first to evidence the classic sociological notion that collective efficacy can be observed on social media. Our findings suggest that counter-speech posted by credible organizations can be an effective measure to prevent harmful narratives, such as online antisemitism. Based on our findings, we underline the value of the work of charities and organizations that aim to protect communities, such as ADL and CST. The presence of such organizations on social media is key to increasing trust in digital communications and platforms and reducing the propagation of cyberhate.

We end this article with suggestions for future research. The online pattern of antagonistic content related to Jews, as identified by text-based classification methods, can act as a proxy for the ebb and flow of antisemitism in the United Kingdom. However, it should be noted that we did not capture tweets that expressed antisemitic conspiracy theories (or allusions to such theories) or antisemitic images posted without accompanying the antisemitic text. Future research investigating the production and propagation of image-based cyberhate and antisemitic theories can further improve our understanding of online antisemitism. Furthermore, the quantitative nature of our collective efficacy observation prevents us from understanding which type of actions from willing and credible actors helps reduce cyberhate perpetration. Future research should look at whether publishing counter-hate speech and counterclaims reduce cyberhate on social media platforms and if so, which types of counter-messages are more effective to reduce the negative effects of hate speech.

Acknowledgements

This research was undertaken using the supercomputing facilities at Cardiff University operated by Advanced Research Computing at Cardiff (ARCCA) on behalf of the Cardiff Supercomputing Facility and the HPC Wales and Supercomputing Wales (SCW) projects. The authors acknowledge the support of the latter, which is part-funded by the European Regional Development Fund (ERDF) through the Welsh Government.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclose receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Community Security Trust.

ORCID iD

Sefa Ozalp https://orcid.org/0000-0002-4104-1541

Notes

1. The same study reports that a smaller number of respondents indicated that they frequently tell the cyberhate perpetrators to stop (10.2%) or frequently defend the victims (13.7%). The authors also report a positive association between cyberhate victimization and those who personally confronted cyberhate perpetrators themselves (termed as self-help in this article). This group was 1.73 times more likely to be targeted by the cyberhate, unlike those who witnessed the presence of collective efficacy.
2. The internet was followed by "in a social situation" with 51%, "among the general public" with 47% and "at political events" with 42% as the most common places where respondents encountered antisemitism.
3. Although understanding the differences and similarities between these three platforms, especially their user base, platform-specific regulations, and platform information propagation mechanics is important, limited space precludes a lengthier discussion here. See Finkelstein et al. (2018) for a discussion.
4. Both ADL and CST aim to record, aggregate, and report antisemitic incidents and help victims of antisemitic abuse. While the former focus their efforts to the United States, the latter focus on the United Kingdom.
5. The authors report that "age-old anti-Semitic tropes and conspiracies are flourishing, particularly among Twitter users that identify as Republicans and/or supporters of President Trump" (p. 10). The most common conspiracy was related to "Soros" keyword with 85% of tweets containing this keyword being antisemitic. The authors also observed that majority of users spreading antisemitic sentiments were human operated (i.e., they were not bots) and of those, only 4.7% were suspended by 15 March 2019.
6. To prevent missing keywords relevant to Jewish identity, we consulted and agreed these keywords with CST.
7. Twitter's streaming API has a policy of allowing users to collect 1% of worldwide daily Twitter communications. The volume of data collected for this study did not breach Twitter's daily limits at any point. We ensured this by looking at the total number of tweets collected during the study period and we did not observe the total number of tweets plateau at any given hour, suggesting that our dataset has never been throttled by the 1% API limit. Therefore, it is unlikely there are any missing data based on rate limiting for this study.
8. This figure is calculated by counting retweets and original tweets separately. We report this figure to be consistent with the original larger dataset the UK subsample was filtered from. The number of original tweets (excluding retweets) is 1,232,744.
9. 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 (which can even contain a negative sentiment, for example, "I don't trust reporters"), leading to false positives. To check the accuracy of the identification, a random sample of 100 users was manually inspected for false positives. Only 13 false positives were identified, meaning 87% of true positives of Media Agents were identified correctly.
10. This list was compiled by a web service (i.e., http://www.mpsontwitter.co.uk), which tracks Twitter accounts of MPs. Please note that only 590 out of 650 had active Twitter accounts.
11. Note that not all of this content was identified as antagonistic in the analysis.
12. CST (2017, p. 27) notes 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'.
13. Given the nature of the social media data (i.e., data are not sampled from a larger population), rather than statistical significance, effect sizes should be interpreted primarily across all models.
14. 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.

References

Anti-Defamation League. (2018). *Quantifying hate: A year of anti-semitism on Twitter*. https://www.adl.org/resources/reports/quantifying-hate-a-year-of-anti-semitism-on-twitter

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

Boyd, D., Golder, S., & Lotan, G. (2010, January). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the Annual Hawaii International Conference on System Sciences* (pp. 1–10). IEEE. https://doi.org/10.1109/HICSS.2010.412

Burnap, P., Rana, O., Williams, M., Housley, W., Edwards, A., Morgan, J., Sloan, L., & Conejero, J. (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. https://doi.org/10.1080/17445760.2014.902057

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

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*, Article 11. http://doi.org/10.1140/epjds/s13688-016-0072-6

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R., & Voss, A. (2014). Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, *4*(1), 2544–2558. https://doi.org/10.1007/s13278-014-0206-4

Community Security Trust. (2017). *Antisemitic incidents report 2016*.

Community Security Trust. (2019). *Antisemitic incidents report January: June 2019*.

Costello, M., Hawdon, J., & Ratliff, T. N. (2017). Confronting online extremism: The effect of self-help, collective efficacy, and guardianship on being a target for hate speech. *Social Science Computer Review*, *35*(5), 587–605. https://doi.org/10.1177/0894439316666272

Douglas, K. M., Mcgarty, C., Bliuc, A.-M., & Lala, G. (2005). Understanding cyberhate: Social competition and social creativity in online white supremacist groups. *Social Science Computer Review*, *23*(1), 68–76.

European Union Agency for Fundamental Rights (Ed.). (2013). *Discrimination and hate crime against Jews in EU member states: Experiences and perceptions of antisemitism*.

Finkelstein, J., Zannettou, S., Bradlyn, B., & Blackburn, J. (2018). A quantitative approach to understanding online antisemitism. *arxiv*: 1809.01644 [cs]. http://arxiv.org/abs/1809.01644

Flatley, J. (2014). British Crime Survey. In G. Bruinsma & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 194–203). Springer. https://doi.org/10.1007/978-1-4614-5690-2_449

Gerstenfeld, P. B. (2017). *Hate crimes: Causes, controls, and controversies*. SAGE.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Hanes, E., & Machin, S. (2014). Hate crime in the wake of terror attacks: Evidence from 7/7 and 9/11. *Journal of Contemporary Criminal Justice*, *30*(3), 247–267. https://doi.org/10.1177/1043986214536665

Home Office. (2018). *Hate crime, England and Wales, 2017/18 statistical bulletin 20/18*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/748598/hate-crime-1718-hosb2018.pdf

King, R. D., & Sutton, G. M. (2013). High times for hate crimes: Explaining the temporal clustering of hate-motivated offending: High times for hate crimes. *Criminology*, *51*(4), 871–894. https://doi.org/10.1111/1745-9125.12022

Leets, L. (2001). Explaining perceptions of racist speech. *Communication Research*, *28*(5), 676–706. https://doi.org/10.1177/009365001028005005

Legewie, J. (2013). Terrorist events and attitudes toward immigrants: A natural experiment. *American Journal of Sociology*, *118*(5), 1199–1245. https://doi.org/10.1086/669605

Maguire, M. (2007). Crime data and statistics. In M. Maguire, R. Morgan, & R Reiner. (Eds.), *The Oxford handbook of criminology* (4th ed., pp. 240–301). Oxford University Press.

Mazerolle, L., Wickes, R., & McBroom, J. (2010). Community variations in violence: The role of social ties and collective efficacy in comparative context. *Journal of Research in Crime and Delinquency*, *47*(1), 3–30. https://doi.org/10.1177/0022427809348898

McKenna, K. Y. A., & Bargh, J. A. (1998). Coming out in the age of the Internet: Identity "demarginalization" through virtual group participation. *Journal of Personality and Social Psychology*, *75*(3), 681–694. https://doi.org/10.1037/0022-3514.75.3.681

Noguchi, T., Stewart, N., Olivola, C. Y., Moat, H. S., & Preis, T. (2014). Characterizing the time-perspective of nations with search engine query data. *PLOS ONE*, *9*(4), Article e95209. https://doi.org/10.1371/journal.pone.0095209

Perry, B. (2001). *In the name of hate: Understanding hate crimes*. Routledge. https://www.taylorfrancis.com/books/9780203905135

R Core Team (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Sampson, R. J. (2001). Crime and public safety: Insights from community-level perspectives on social capital. In S. Saegert, J. Phillip Thompson, & M. R. Warren (Eds.), *Social capital and poor communities* (pp. 89–114). Russels Sage.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, *277*(5328), 918–924. https://doi.org/10.1126/science.277.5328.918

Statista. (2019). Twitter: number of active users 2010–2019. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Tilley, N., & Tseloni, A. (2016). Choosing and using statistical sources in criminology: What can the crime survey for England and Wales tell us? *Legal Information Management*, *16*(2), 78–90. https://doi.org/10.1017/S1472669616000219

Wall, D. S., & Williams, M. (2007). Policing diversity in the digital age: Maintaining order in virtual communities. *Criminology & Criminal Justice*, *7*(4), 391–415. https://doi.org/10.1177/1748895807082064

Williams, M. L. (2006). *Virtually criminal: Crime, deviance and regulation online*. Routledge. http://orca.cf.ac.uk/3120/

Williams, M. L., & 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*(2), 211–238. https://doi.org/10.1093/bjc/azv059

Williams, M. L., Burnap, P., & Sloan, L. (2017). Crime sensing with big data: The affordances and limitations of using open source communications to estimate crime patterns. *British Journal of Criminology*, *57*(2), 320–340. https://doi.org/10.1093/bjc/azw031

Williams, M. L., & Tregidga, J. (2014). Hate crime victimization in Wales: Psychological and physical impacts across seven hate crime victim types. *British Journal of Criminology*, *54*(5), 946–967. https://doi.org/10.1093/bjc/azu043

Woolley, S., & Joseff, K. (2019). *Jewish Americans computational propaganda in the United States: Trends in anti-semitic harassment and political disinformation on social media*. Institute for the Future. http://www.iftf.org/fileadmin/user_upload/downloads/ourwork/IFTF_JewishAmerican_comp.prop_W_05.07.19.pdf

Yang, J., & Counts, S. (2010, May). Predicting the speed, scale, and range of information diffusion in Twitter. In Proceedings of the Fourth International Conference on Weblogs and Social Media (p. 4). Association for the Advancement of Artificial Intelligence.
Zarrella, D. (2009). The Science of Retweets. https://cdn2.hubspot.net/hub/53/file-13207809-pdf/docs/science-of-retweets-201003.pdf

Author Biographies

Sefa Ozalp is the Lead Data Science Researcher at HateLab, Cardiff University. His research interests include online antagonism, data mining, applications of social data science, machine learning, big data and high-performance computing, quantitative criminology, and computational social science.

Matthew L. Williams is the Director of HateLab and the Social Data Science Lab, and Professor of Criminology in the School of Social Sciences at Cardiff University. His main areas of research activity are Hate Crime, Hate Speech and Extremism Online, Computational Social Science, Cybercrime/Human Factors in Cybersecurity.

Pete Burnap is Professor of Data Science & Cybersecurity at Cardiff University. He is Director of Cardiff's Social Data Science Lab, as well as the Academic Centre of Excellence in Cyber Security Research (ACE-CSR). He has been involved in grants in worth in excess of £14 m, leading large awards from EPSRC, ESRC, and industry on the topics of social data science and cybersecurity analytics—the fusion of AI, cybersecurity, and risk. He is co-director of the Data Innovation Accelerator (DIA)—a £3.75 m investment in upskilling SMEs in South Wales to develop innovative AI-driven products and services. He sits on the UK Government's AI Council, advising on the implementation of the industrial strategy in AI and the Data Economy.

Han Liu is currently a Research Associate in Data Science in the School of Computer Science and Informatics at Cardiff University. His research interests include data mining, machine learning, neural networks, rule-based systems, intelligent systems, expert systems, fuzzy systems, big data analytics, granular computing and applications in natural language processing and computer vision.

Mohamed Mostafa (PhD, MBCS) is a Senior Lecturer of Data Science at the Cardiff School of Technologies. His research interests include data science, machine learning, social network analysis, and human-computer interaction.

Appendix A

Creating a Training Sample for the Supervised Learning Classifier

During the process of machine classification experimentation, it became evident that human annotators struggled with classifying antisemitic "hate speech" with a high degree of accuracy. In the first stage, we sampled 4,000 tweets from the UK dataset and used the online Figure 8 service to source human annotators to perform annotation tasks on each tweet to determine, in their view, whether it was antagonistic in relation to Jews. 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 offense, 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 labeled, resulting in insufficient data to train the machine learning classifier. We requested four annotators per tweet and removed instances from the training data if the inter-annotator agreement score dropped below 75%. In the second stage, human annotations from the Figure 8 dataset were checked against the text sample of offensive online communications provided by CST and adjustments were made where misclassifications were identified.

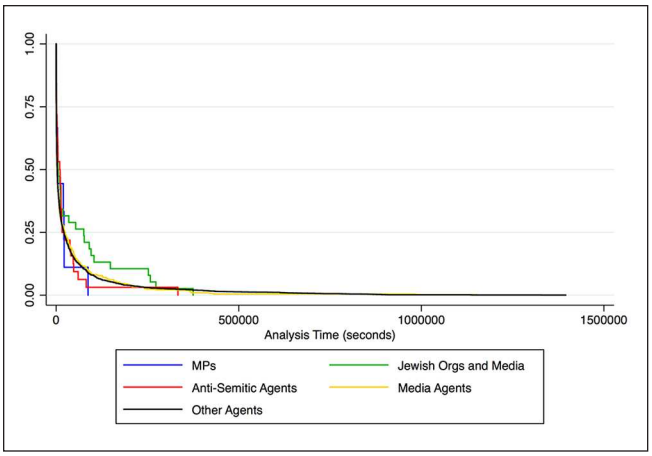

Figure 8. Kaplan-Meier survival estimates for agent type in event 3.

In the next stage, we conducted a manual "sanity checking" using a sample of antagonistic text from Twitter, Facebook, emails, and other forms of online communications provided by CST. 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. However, many of them were deemed sufficiently offensive to warrant requests to social media providers to delete content for infringing platform Terms of Service. Informed by these antagonistic text examples, we checked the consistency of the human annotation on the Figure 8 subsample with at least 75% agreement score ($n = 1{,}322$). Where deviations were evident, we made adjustments to Figure 8 coder annotations. For example, tweets coded as antagonistic toward Jews in the Figure 8 dataset, that were clearly only anti-Israel in nature, were recoded as not-antisemitic. In total, 29% of the Figure 8 subset was adjusted in this way. Although multiple processing steps meant that the number of tweets used for the training dataset was less than the initial 4,000 tweets sampled for human annotation, we are confident that the training sample was semantically representative of the classification problem at hand.

Table 7. Classification Results for Antisemitism Hate Speech—10 Fold Cross Validation.

Feature extraction		DT			NB			SVM			Fuzzy		
		P	R	F	P	R	F	P	R	F	P	R	F
BOW	No	0.656	0.738	0.694	0.562	0.953	0.707	0.665	0.776	0.716	0.607	0.791	0.687
	Yes	0.630	0.536	0.579	0.662	0.111	0.190	0.665	0.531	0.590	0.638	0.418	0.506
	Overall	0.644	0.647	0.642	0.607	0.574	0.474	0.665	0.666	0.659	0.621	0.623	0.606
NG	No	0.583	0.761	0.660	0.545	1.000	0.706	0.621	0.798	0.699	0.591	0.821	0.687
	Yes	0.549	0.348	0.426	0.000	0.000	0.000	0.633	0.418	0.503	0.599	0.320	0.417
	Overall	0.568	0.575	0.555	0.300	0.550	0.388	0.626	0.627	0.611	0.595	0.596	0.566
TD	No	0.549	0.974	0.702	0.545	1.000	0.706	0.545	0.968	0.698	0.551	0.957	0.699
	Yes	0.571	0.041	0.077	0.000	0.000	0.000	0.464	0.034	0.062	0.556	0.064	0.115
	Overall	0.559	0.554	0.421	0.300	0.550	0.388	0.509	0.548	0.412	0.553	0.555	0.436

BOW: Bag of Words; DT: Decision Trees; F: F-Measure; NB: Naïve Bayes; NG: N-Grams; P: Precision; R: Recall; SVM: Support Vector Machine; TD: Typed Dependencies.

Table 8. Classification Results for Antisemitism Hate Speech—70/30 Split.

FE	DT			NB			SVM			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F
BOW	0.680	0.610	0.640	0.770	0.090	0.160	1.000	1.000	1.000	0.530	0.490	0.510
NG	0.000	0.000	0.000	1.000	0.330	0.500	0.330	0.670	0.440	1.000	0.400	0.570
TD	0.750	0.050	0.100	0.000	0.000	0.000	0.900	0.080	0.140	0.580	0.130	0.210

BOW: Bag of Words; DT: Decision Trees; F: F-Measure; FE: Feature extraction; NB: Naïve Bayes; NG: N-Grams; P: Precision; R: Recall; SVM: Support Vector Machine; TD: Typed Dependencies.

Appendix B

Supervised Machine Learning Classifier for Online Antisemitism on Twitter

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 in Table 7 using standard text classification measures of the following: 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 normalized 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 normalized by the sum of true positives and false negatives); and F-Measure (F), a harmonized 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). Initially, we used a 10-fold cross-validation approach to test the supervised machine learning method. This functions by splitting the dataset into 10 equal randomly shuffled subsets and iteratively using 9 folds to train the classifier and 1 fold to test it. After 10 iterations the results are averaged. It is particularly useful with small labeled datasets as was the case in this instance.

Table 1 shows that SVM + BOW performed 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% of the manually coded dataset, and classifying the remaining 30% 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 generalized 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 cause further confusion—exemplifying the difficulty in using highly frequent words extracted from the short informal text as features, with such a small "gold standard" dataset.

Appendix C

Computational Cost of This Study

Unlike traditional quantitative datasets, such as survey or panel data, "big data" collected from social media platforms are not structured to answer social research questions. We conducted a series of preprocessing steps to filter the dataset to UK-based tweets in accordance with the research questions, extract the metrics that would be used to build statistical models and automatically identify antagonistic tweets by training a machine learning classifier. It is important to note that in computational social science research, where the big data are captured from the "wild," it is common practice to spend more time on preprocessing the data. The process of cleaning and structuring raw datasets according to research questions is also called *data wrangling* and it usually takes 80% of the time spent on data analysis.

It is, therefore, worth mentioning the computational data processing cost of the conducted in this study in order to provide perspective for the reader. Performance wise, our data processing pipeline was relatively slow. Once we were confident with our query inferring UK-based location from metadata, we extracted roughly 2.7M tweets from UK-based users. We tested the performance of the data processing on a small sample on a single core on a local desktop computer. Our initial performance tests indicated that it took 4.5 seconds to extract various metrics for statistical modeling (agent type, antagonistic content classification, etc.) for each tweet on average. This was not feasible, given the size of the dataset. Roughly, this would take a single desktop computer running on 1 physical core 140 days. To shorten the processing time, we refactored our code in a parallelised fashion and run the process on High-Performance Computing (HPC) servers of the Supercomputing Wales. This allowed our "Big Data" processing job to be split and run concurrently over multiple cores and multiple nodes, allowing us to complete the whole data processing (i.e., classification and extraction of independent variables) just under 24 hours. We are grateful to Supercomputing Wales for enabling this research to run in such a short time.