Document Title:	Understanding Online Hate Speech as a Motivator and Predictor of Hate Crime
Author(s):	Meagan Cahill, Katya Migacheva, Jirka Taylor, Matthew Williams, Pete Burnap, Amir Javed, Han Liu, Hui Lu, Alex Sutherland
Document Number:	304532
Date Received:	April 2022
Award Number:	2016-MU-MU-0009

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Final Summary Overview

Understanding Online Hate Speech as a Motivator and Predictor of Hate Crime

Authors:

Meagan Cahill, Katya Migacheva, and Jirka Taylor *RAND U.S.*
Matthew Williams, Pete Burnap, Amir Javed, and Han Liu *Cardiff University*
Hui Lu and Alex Sutherland, *RAND Europe*

RAND Social and Economic Well-Being

September 2019
Prepared for National Institute of Justice, Office of Justice Programs, U.S. Department of Justice.

Acknowledgements:

This work was made possible by Award Number 2016-MU-MU-0009, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

Abstract

In the United States, a number of challenges prevent an accurate assessment of the prevalence of hate crimes in different areas of the country. These challenges create huge gaps in our knowledge about hate crime—who is targeted, how, and in what areas—which in turn, hinder appropriate policy efforts and allocation of resources to the prevention of hate crime. In the absence of high-quality hate crime data, online platforms may provide information that can contribute to a more accurate estimate of the risk of hate crimes in certain places and against certain groups of people. Data on social media posts that use hate speech or internet search terms related to hate against specific groups has the potential to enhance and facilitate timely understanding of what is happening offline, outside of traditional monitoring (e.g., police crime reports). The current work assessed the utility of Twitter data to illuminate the prevalence of hate crimes in the United States with the goals of (i) addressing the lack of reliable knowledge about hate crime prevalence in the U.S. by (ii) identifying and analyzing online hate speech and (iii) examining the links between the online hate speech and offline hate crimes.

The project drew on four types of data: recorded hate crime data, social media data, census data, and data on hate crime risk factors. We adopted an ecological framework and Poisson regression models to study the explicit link between hate speech online and hate crimes offline and used risk terrain modeling to further assess our ability to identify places at higher risk of hate crimes offline. The ecological models produced mixed results, with weak correlations found between tweets containing hateful language and specific types of hate crime. The strongest associations were found for religiously motivated crimes, and a counter-intuitive result was found for racially motivated crimes. RTM analyses also produced mixed results, though generally supportive of the findings from the ecological models.

Although the results were inconsistent, they did point to the potential for using online behavior to identify offline risk. More exploration of implicit sentiments expressed online—search terms, for example—may be more appropriate in the current context of social media platforms more strictly enforcing policies against the use of hate speech online.

Table of Contents

Tables

Introduction

In the United States, a number of challenges prevent an accurate assessment of the prevalence of hate crimes in different areas of the country. These challenges include inconsistent laws and statutes defining hate crimes from one jurisdiction to another, limited information recorded about crimes that might fit a definition of a hate crime, and a lack of motivation on the part of public safety agencies—at local, state, and federal levels—to improve reporting processes. These challenges create huge gaps in our knowledge about hate crime—who is targeted, how, and in what areas—which in turn, hinder appropriate policy efforts and allocation of resources to the prevention of hate crime. In the absence of high-quality hate crime data, online platforms may provide information that can contribute to a more accurate estimate of the risk of hate crimes in certain places and against certain groups of people.

Because of its anonymity, accessibility, and global reach, the Internet has become an increasingly popular platform for the expression of hate (Banks, 2010; Costello et al., 2017; Foxman and Wolf, 2013, Saleem et al., 2017). Particularly in the United States, where non-criminal speech is legally protected, hate groups and individuals are able to post their opinions online without fear of legal recourse (Hawdon et al., 2017). Even when online hate speech incidents do not amount to criminal offenses, they can serve as important indicators of intergroup tensions. Online hate speech can create an environment in which offline hate crime can occur (Awan, 2014) and can lead to a variety of harmful outcomes, including radicalization (Foxman and Wolf, 2013, Hassan et al., 2018), violence (Ybarra et al., 2008), increased prejudice (Soral et al., 2017), distrust (Nasi et al., 2015), and, for targets, various forms of emotional distress, such as anxiety and fear (Tynes et al., 2014). Online hate speech data can demonstrate the existence of hate or bias sentiment in an area and signal an increased potential for offline criminal behavior against certain groups (Alden & Parker, 2005). Understanding how and whether hate speech *online* translates into hate crimes *offline*, then, can potentially offer an important tool for diagnosing the risk of on-the-ground hate action.

Data on social media posts that use hate speech or internet search terms related to hate against specific groups has the potential to enhance and facilitate timely understanding of what is happening offline, outside of traditional monitoring (e.g., police crime reports) or in areas where victims may be reluctant to report hate crimes (St. Louis & Zorlu, 2012). New reliable methods for identifying hate-related sentiment both online and offline would allow for more prompt reaction from affected communities, policymakers, public safety agencies, and victim services.

Using Twitter to Measure Social Phenomena

Open and widely accessible social media platforms, such as Twitter, are increasingly used across the globe to publish content. Emerging research has successfully relied on Twitter for detection of important trends in a variety of domains including health (e.g., obesity, influenza outbreaks, Paul and Dredze, 2013; heart disease, Eichstaedt et al., 2015) and societal processes (e.g., 'disruptive events,' Elson, et al., 2012; Alsaedi, Burnap, and Rana, 2015). Other research efforts have also established links between perceptions of neighborhood environmental and social disorder expressed on Twitter and actual crime rates there (Williams, Burnap and Sloan 2016). Compared to expensive community surveys, research using Twitter data is relatively inexpensive, has the power to generate estimates of various phenomena based on postings from millions of people across time and space (Eichstaedt et al., 2015).

To date, limited research exists on the links between cyberhate, posted on social media platforms like Twitter, and offline hate crimes. For example, Williams and Burnap (2015) found that Twitter posts containing references to race, ethnicity, and religion in the immediate aftermath of a terrorist act can predict the spread of online hate speech following the event; they then expanded this work to include other forms of hate speech targeting sexual orientation and disability (Burnap & Williams, 2016). These early studies offer promise for how big data from Twitter can be mined to link speech on Twitter to on-the-ground events.

The current research effort built upon the Williams and Burnap (2015, 2016) work and assessed the utility of Twitter as a source of data to illuminate the prevalence of hate crimes in the United States. The overarching goals of the research were to (i) address the lack of reliable knowledge about hate crime prevalence in the U.S. by (ii) identify and analyze online hate speech and (iii) examine the links between the online hate speech and offline hate crimes. To achieve these goals, the project pursued the following three objectives:

1. Classify online hate speech in terms of (i) which individuals and groups direct what kinds of speech (type and severity) at (ii) which groups and (iii) where the tweets are generated.
2. Estimate the relationship between online hate speech classification and offline hate crime
3. Develop and test an empirical model to identify areas at increased risk of hate crimes.

Project Design and Methods

Data

The project drew on four types of data: recorded hate crime data, social media data, census data, and data on hate crime risk factors. These are explained in more detail below.

Recorded hate crime data served as a dependent measure in our analyses. We obtained data on hate crimes recorded in 2017 and 2018 in L.A. County, compiled by the Los Angeles County Commission on Human Relations (LACCHR). These data represent the most comprehensive data set on hate crimes available in the county. The LACCHR receives hate crime incident reports from 46 law enforcement agencies, 5 community organizations, 36 school districts and 13 higher education institutions, as well as directly from victims. LACCHR staff review the data from all sources to determine whether each reported incident meets the definition of a hate crime as defined by applicable statutes. Staff also check for duplicate reports to ensure incidents are not double-counted. For incidents that occurred in public places, we received the actual location of the incident; for those occurring in private locations, we received mid-block location information. Data from LACCHR were coded into three categories for analyses: i) racially motivated hate crimes; ii) religion motivated hate crimes; iii) and sexual orientation motivated hate crimes. For the purposes of the ecological analysis, the data were then aggregated to census tracts, providing us with count data for each measure by census tract and year.

Social media data were the main independent measure of interest. Using the Twitter streaming Application Programming Interface (API) via COSMOS software (Burnap et al., 2014), we collected all tweets posted between September 2017 and September 2018 and geotagged to L.A. County. These data were used to derive a count of all geocoded tweets; 1,813,862 tweets were geolocated within L.A. County in 2017 and 2018.

Supervised machine learning classifiers were then built to identify hateful tweets targeting three characteristics: race (anti-African-American), religion (anti-Muslim, anti-Jewish) and sexual orientation (anti-lesbian, gay, and bisexual). Recorded hate crimes in L.A. County are most frequently reported to target one of these three characteristics. Three gold standard datasets of human coded annotations were generated to train the machine classifiers based on samples of tweets (see Appendix B for classifier results). The classifiers were then used to identify all hateful tweets in the dataset, including which characteristics the tweet targeted. Finally, we aggregated all geolocated tweets to census tracts, giving us counts of all tweets and hateful tweets by tract. An important caveat to both social media and hate crime data is that neither represents a representative sample of the true population: we captured only tweets from users opting to have their tweets geotagged and offline, we have data only on *reported* hate crimes.

Census data. We also collected the latest 5-year estimates from the American Community Survey for use as controls in analytic models. Relevant variables were selected based on literature that estimated hate crime using ecological factors (e.g. Green, 1998; Espiritu, 2004). These include age, employment status, race and educational attainment.

*Hate crime risk factor data***.** We reviewed the existing research literature to identify particular environmental features that served as risk factors in risk-terrain models (see Table 2 for the full list of 20 variables). Data on these factors were obtained from public sources including the public L.A. County GIS portal.

Analytic methods

Ecological analysis. Using census tracts as the unit of analysis in the models allowed for an 'ecological' appraisal of the explanatory power of hate tweets for estimating police recorded hate motivated offences (Sampson, 2012). As we adopt an ecological framework, using census tracts, not individuals, as our unit of analysis, we cannot state with confidence that area level factors *cause* the outcome. There are likely factors and tract characteristics that contribute to the causal pathways, but that we were unable to observe in this study design. Thus, inferential statistics are not used as the data do not represent a random probability sample and claims of causality in this project would stretch the data beyond their limits. To incorporate the temporal variability of recorded hate crimes and tweets into statistical models, we adopted a random-effects (RE) and fixed-effects (FE) regression framework. RE modelling allows for the inclusion of variables that are time-variant (police and Twitter data) and time-invariant (census measures). A Poisson RE/FE estimation with robust standard errors is recognized as the most reliable option in the presence of over-dispersion (Wooldridge, 1999). Indeed, FE models are the most robust test given they are based solely on within-census tract variation, allowing for the elimination of potential sources of bias by controlling for observed *and unobserved* ecological characteristics (Allison, 2009). In contrast, RE models only take into account the factors that are included as regressors. Both RE and FE estimates were produced for all models to address selection bias introduced by unobserved time-invariant variables.[1]

*Risk-terrain modeling***.** Particular characteristics of an area affect crime risk in the area itself and in the surrounding area. However, the relationship between risk factors and crime is complex; it is likely that different combinations of risk factors will determine a location's overall risk. Risk terrain modeling (RTM) comprises a series of geospatial techniques that attempt to (1) identify geographic features (e.g., bars, certain major roads) that contribute to risk (for example, crime risks) and (2) make predictions about risk in a particular location based on how close it is to risk-inducing features and how densely those features cluster. RTM supports consideration of multiple factors concomitantly as well as helping to understand the mechanisms that enable hot spots to emerge and persist over time. The RTM framework is used to identify the places where risk factors are co-located to produce increased risk or vulnerability. The approach has been widely used for crime analysis.

We used risk-terrain modelling as an exploratory method to identify statistically significant risk factors for hate crimes and their spatial influences within the City of Los Angeles only, where the most data on risk factors was available and where hate crimes were numerous. Analyses were conducted using the Risk Terrain Modelling Diagnostics (RTMDx) software.[2] As

[1] To determine if RE or FE is preferred the Hausman test can be used. However, this has been shown to be inefficient, and we prefer not to rely on it for interpreting our models (see, Troeger, 2008). Therefore, both RE and FE results should be considered together.

[2] Available at: https://rtmdx.net/analysis [last accessed 25 September 2019]

the unit of analysis, we designated cells sized 300x300 meters as we were specifically concerned with the spatial influences at the micro-level. The model was invited to select the optimal of two possible operationalizations for each risk factor: proximity (i.e., distance of an environmental feature) or density (i.e., concentration of features).

Results

Ecological analysis

The Breusch-Pagan Lagrange Multiplier test revealed RE regression was favorable over single level regression. The results of the RE/FE models assess variation over space and time represent a significant improvement over models that do not take into account time as a factor. There were no issues with multi-collinearity in the final models. Coefficients and Incidence Rate Ratios (IRR) are presented to show relationships between variables and the magnitude of effects.

Table 1 presents the results of Models A to C for each type of hate crime to assess the effects of all regressors in stages. Model A includes only the census regressors for the RE estimations. For racially motivated hate crime, proportion of African-Americans in a census tract emerged as the most influential (IRR 1.02), followed by proportion unemployed (IRR 1.01), with both regressors exerting a positive effect, in line with existing research on hate crime (Green, 1998; Espiritu, 2004). Both 'proportion with high school diploma' and 'proportion aged 15 to 24' were negatively associated with racially motivated hate crime. For sexual orientation motivated crime, similar associations emerged, with a larger effect for the unemployment regressor (IRR 1.05).

A different pattern of association emerged for religiously motivated crime. Although the effects of 'proportion with high school diploma' in this model are similar in direction and magnitude to its effects on other hate crimes, the effects of 'proportion aged 15 to 24' and 'unemployed' reverse—they were both more strongly associated with higher levels of religiously motivated crime. The ‘proportion of white residents’ also exerted a positive effect (IRR 1.03). In other words, areas with greater proportions of young, unemployed, and white residents tend to have higher levels of religiously motivated hate crimes. Models B and C were estimated with RE and FE, introducing variables measuring online hate speech and count of geocoded tweets. Model B introduced online hate speech alone. Results for both RE and FE models are mixed. Hate tweets only emerged as positively associated with religiously motivated (RE IRR 1.04, FE IRR 1.13) and sexual orientation motivated crimes (RE IRR 1.001, although the FE estimation indicated a negative relation FE IRR 1.05 – see Model C below for further explanation). For religiously motivated crime, the effect of hate tweets was equal to proportion unemployed.

For racially motivated crimes, hate tweets exhibited a strong negative effect (IRR -1.45). This counter-intuitive pattern may relate to a high number of false positives in the race hate speech machine learning classifier. The n-word is frequently used in both negative and positive connotations on Twitter. It is possible the classifier mislabeled non-derogatory uses of the term by African-American Twitter users as hate speech, partly accounting for this negative effect.

Model C estimates control for total counts of geocoded Tweets, reducing the likelihood that the count of hate tweets is acting as a proxy for population density (Malleson and Andresen, 2015). For racially and religiously motivated crimes, the direction of relationships between hate tweets and crimes does not change by the introduction of this regressor, and the magnitude of the effects actually increase. For sexual orientation motivated crimes, the direction of the relationship between hate tweets and crimes reverses, indicating that this regressor may have been acting partly as proxy for population density in this model. This result, and the positive association of Tweet Count (RE IRR 1.004) with sexual orientation motivated hate crimes, may reflect higher use of Twitter in areas with larger populations of LGBTQ establishments.

Risk terrain modeling

A Relative Risk Score (RRS) was assigned to each place in the study area, ranging from 1 for the lowest risk to 76.4 for the highest risk place. These scores allow for easy comparison among places in the risk terrain map. For instance, a place with an RRS of 10 has an expected rate of events pertaining to your study topic that is 10 times higher than a place with a score of 1. Of the 23 environmental features tested, the risk terrain modelling identified several significant risk factors, with results differing by the motivation underlying reported hate crimes. Hateful tweets, along with other features examined, were identified as a significant risk factor for both racially-motivated and religion-motivated hate crimes. However, sexual orientation-based hateful tweets were not found to be significant risk factors for sexual orientation-motivated hate crimes, which echoes the mixed results obtained from the ecological analysis. Table 3 shows the significant risk factors identified for each type of hate crime, along with their most relevant spatial influence, including the operationalization, spatial influence, and relative risk value (RRV).

Discussion and Implications

These mixed results indicate that with additional methodological development, online data sources may provide useful information that can augment traditional police and victim survey sources on the hate crime problem. However, a number of limitations should be noted. We used only geotagged tweets and only a small number of Twitter users geotag their tweets. It is possible that this subset of users has different tweeting behavior than the average user. In addition, our crime data only captures incidents reported to police or another organization that contributes data to the LACCHR; it is thus likely that the number of hate crimes used in the study is lower than the true number that occur.

Finally, a key challenge in this project was the adoption of a hate speech policy by Twitter that coincided with the start of our project. In December 2015, Twitter explicitly banned 'hateful conduct' on the platform for the first time, introducing a set of rules on what users cannot post; where tweets break these rules, they are flagged and deleted). Then, in August 2016, Twitter introduced the 'Quality Filter' for all users and in February 2017, Twitter introduced the 'Hide Sensitive Content' feature. Both filters—enabled by default for all users—were created to tackle increasing harassment on the site. The filters effectively hide hate speech from users' timelines, reducing retweeting. The filters have reduced hate speech on the platform and many users that frequently tweeted hate speech abandoned Twitter for social media with strong 'free speech' principles (e.g., Gab, Voat and 4/8Chan). Our data harvesting began after the introduction of these new rules, and we collected less hate speech than was anticipated when the project was conceived. In our previous UK-based study that used tweets from 2013 and 2014, prior to Twitter's new policy, similar models produced more conclusive results (Williams et al., 2019).

The findings from the current work have some implications for the link between online hate speech and offline hate crime. First, though the findings were uneven depending on type of hate/crime examined, the work demonstrated that Twitter is potential source of insight. Second, hateful tweets can be used to identify areas at increased risk for hate crimes but are best used with other risk factors, as RTM findings demonstrated. In light of the fact that social media users are self-censoring their posts on more popular platforms in favor of 'free speech' platforms that are relatively inaccessible to researchers, future work should rely less on explicit measures of online hate, such as clearly racist posts. Instead, implicit measures of prejudice, such as Google searches and network connections on Twitter and Facebook, may yield more fruitful results.

References

Alden, H. L., & Parker, K. F. (2005). Gender Role Ideology, Homophobia and Hate Crime: Linking Attitudes to Macro-Level Anti-Gay and Lesbian Hate Crimes. *Deviant Behavior*, 26(4), 321–343.

Allison, D. P (2009) *Fixed Effects Regression Models*, London: Sage.

Alsaedi, N., Burnap, P., Rana, O. (2016) Automatic Summarization of Real World Events Using Twitter. Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016). As of 14 October 2018: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13017/12775

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers & Technology*, 24(3), 233–239.

Bobo, L., and Licari, F. C. (1989) 'Education and Political Tolerance: Testing the Effects of Cognitive Sophistication and Target Group Affect', *Public Opinion Quarterly* 53(3):285–308.

Burnap, P., Rana, O., Williams, M., Housley, W., Edwards, A., Morgan, J, Sloan, L. and Conejero, J. (2014) 'COSMOS: Towards an Integrated and Scalable Service for Analyzing Social Media on Demand', *IJPSDS*, 30(2):80-100.

Burnap, P. and Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5, article number: 11.

Costello, M., Hawdon, J., and 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.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... and Weeg, C. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, *26*(2), 159-169.

Elson, S. B., Yeung, D., Roshan, P., Bohandy, S. R., and Nader, A. (2012). *Using Social Media to Gauge Iranian Public Opinion and Mood After the 2009 Election*. Santa Monica, CA: RAND.

Espiritu, A. (2004) Racial Diversity and Hate Crime Incidents, *The Social Science Journal*, 41(2):197-208.

Foxman, A. H., and Wolf, C. (2013). *Viral hate: Containing its spread on the Internet*. London, England: Macmillan

Green, D. P., Strolovitch, D. Z. and Wong, J. S. (1998) 'Defended neighborhoods, integration and racially motivated crime', *American Journal of Sociology*, 104(2):372–403.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Hassan, G., Brouillette-Alarie, S., Alava, S., Frau-Meigs, D., Lavoie, L., Fetiu, A., ... and Sieckelinck, S. (2018). Exposure to extremist online content could lead to violent radicalization: A systematic review of empirical evidence. *International Journal of Developmental Science*, (Preprint), 1-18.

Hawdon, J., Oksanen, A., and Rasanen, P. (2014). Victims of online hate groups: American youth's exposure to online hate speech. In J. Hawdon, J. Ryan, & M. Lucht (Eds.), *The causes and consequences of group violence: From bullies to terrorists* (pp. 165–182). Lanham, MD: Lexington Book.

Malleson, N. and M.A. Andresen (2015) 'Spatio-temporal crime hotspots and the ambient population', *Crime Science*, 4(10)1-8.

Nasi, M., Rasanen, P., Hawdon, J., Holkeri, E., and Oksanen, A. (2015). Exposure to online hate material and social trust among Finnish youth. *Information, Technology and People*, 28, 607–622.

Paul, M. J., and Dredze, M. (2014). Discovering health topics in social media using topic models. *PloS one*, *9*(8), e103408.

Saleem, H. M., Dillon, K. P., Benesch, S., and Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.

Sampson, R. J. (2012), *Great American City: Chicago and the Enduring Neighborhood Effect*. University of Chicago Press.

Soral, W., Bilewicz, M., and Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior*, 44(2), 136-146.

Tynes B., Rose C., Hiss S., Umana-Taylor A. J., Mitchell K., Williams D. (2014). Virtual environments, online racial discrimination, and adjustment among a diverse, school-based sample of adolescents. *International Journal of Gaming and Computer Mediated Simulations*, 6(3), 1-16.

St Louis, C., and Zorlu, G. (2012). Can Twitter predict disease outbreaks?. *Bmj*, *344*, e2353.

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,* 52(2): 211-238.

Williams, M. L., Burnap, P., and Sloan, L. (2017). Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *The British Journal of Criminology*, *57*(2), 320-340.

Williams, M. L. et al. 2019. Hate in the machine: Anti-black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *British Journal of Criminology* (10.1093/bjc/azz049).

Wooldridge, J. M., (1999) ‘Distribution-Free Estimation of Some Nonlinear Panel Data Models’, *Journal of Econometrics*, 90(1):77–97.

Ybarra, M.L., Diener-West, M., Markow, D., Leaf, P.J., Hamburger, M. and Boxer, P. (2008). “Linkages Between Internet and Other Media Violence With Seriously Violent Behavior by Youth.” Pediatrics 122(5):929–937

Appendix A: Tables

Table 1: Random and Fixed Effects Poisson Estimations

			Racially Motivated Crime			
	Model A		Model B		Model C	
Random Model	*Coef*	IRR	*Coef*	IRR	*Coef*	IRR
High School Diploma	-0.00049	0.99951	-0.00052	0.99948	-0.00048	0.99952
15 to 24 year olds	-0.00002	0.99998	-0.00002	0.99998	-0.00002	0.99998
Unemployed	0.01467	1.01478	0.01292	1.01301	0.01542	1.01554
African-American (AA)	0.01649	1.01663	0.01656	1.01670	0.01665	1.01679
Tweet Count					0.00053	1.00053
Anti-AA Hate Tweets			-0.36920	0.69129	-0.46888	0.62570
constant	-4.47685	0.01137	-4.41959	0.01204	-4.50259	0.01108
Fixed Model						
Tweet Count					0.00067	1.00067
Anti-AA Hate Tweets			-0.52129	0.59376	-0.58930	0.55472
			Religiously Motivated Crime			
	Model A		Model B		Model C	
Random Model	Coef	IRR	Coef	IRR	Coef	IRR
High School Diploma	-0.00148	0.99852	-0.00148	0.99852	-0.00148	0.99852
15 to 24 year olds	0.00058	1.00058	0.00058	1.00058	0.00058	1.00058
Unemployed	-0.04181	0.95905	-0.04175	0.95911	-0.04196	0.95891
White	0.02686	1.02722	0.02685	1.02721	0.02687	1.02724
Tweet Count					-0.00005	0.99995
Anti-Religion Hate Tweets			0.03897	1.03974	0.06457	1.06670
Constant	-6.18644	0.00206	-6.18882	0.00205	-6.18415	0.00206
Fixed Model						
Tweet Count					-0.00091	0.99909
Hate Tweets			0.11781	1.12504	0.18525	1.20352
			Sexual Orientation Motivated Crime			
	Model A		Model B		Model C	
Random Model	*Coef*	IRR	*Coef*	IRR	*Coef*	IRR
High School Diploma	-0.00072	0.99928	-0.00072	0.99928	-0.00072	0.99928
15 to 24 year olds	-0.00050	0.99950	-0.00050	0.99950	-0.00046	0.99954
Unemployed	0.05326	1.05471	0.05336	1.05481	0.05571	1.05729
Tweet Count					0.00041	1.00041
Anti-LGB Hate Tweets			0.00051	1.00051	-0.00938	0.99066
Constant	-5.05225	0.00639	-5.05512	0.00638	-5.11883	0.00598
Fixed Model						
Tweet Count					-0.00023	0.99977
Anti-LGB Hate Tweets			-0.04952	0.95169	-0.04687	0.95421

Notes: Table shows results of separate random and fixed effects models. To determine if RE or FE is preferred the Hausman test can be used. However, this has been shown to be inefficient, and we prefer not to rely on it for interpreting our models (see, Troeger, 2008). Therefore, both RE and FE results should be considered together.

Table 2: List of environmental features tested as risk factors in RTM analysis

No	Risk factors	N	Source of data
1	Adult education institutions	218	LA County GIS Data Portal
2	Churches	2,378	LA County GIS Data Portal
3	Colleges and universities	215	LA County GIS Data Portal
4	FDIC insured banks	1,789	LA County GIS Data Portal
5	Food assistance institutions	342	LA County GIS Data Portal
6	Health centers	125	LA County GIS Data Portal
7	Health clinics	238	LA County GIS Data Portal
8	Homeless shelters services	192	LA County GIS Data Portal
9	Hospitals medical centers	357	LA County GIS Data Portal
10	Immigration	79	LA County GIS Data Portal
11	Passports	135	LA County GIS Data Portal
12	Public elementary schools	1,208	LA County GIS Data Portal
13	Public high schools	171	LA County GIS Data Portal
14	Public middle schools	286	LA County GIS Data Portal
15	Public housing	106	LA County GIS Data Portal
16	Bank main offices	87	LA County GIS Data Portal
17	Metro stations (LA city transit)	97	LA Metro GIS Data Portal
18	Metrolink stations (commuter rail)	58	LA County GIS Data Portal
19	Alcohol-licenced places	301	Census NAICS data
20	Grocery stores	1,164	LA City Data Portal
21	Race-based hateful tweets	1,824	Identified by the research team
22	Religion-based hateful tweets	937	Identified by the research team
23	Sexual orientation-based hateful tweets	27,555	Identified by the research team

Table 3: List of environmental features tested as risk factors in RTM analysis

	n	Operationalization	Spatial Influence (m)	Relative Risk Value (RRV)
Racially motivated hate crimes				
Public Middle Schools	286	Proximity	300	6.4
FDIC Insured Banks	1,789	Proximity	450	2.4
churches	2,378	Proximity	600	2.4
Hateful tweets - racially related	1,824	Proximity	600	2.1
Religion-motivated hate crimes				
FDIC Insured Banks	1,789	Density	900	5.1
Hateful tweets - religion related	937	Proximity	750	3.1
Sexual orientation-motivated hate crimes				
Public Housing	106	Proximity	900	6.0
Passports	135	Proximity	750	4.4

Note: To illustrate how the results should be interpreted: For the Racial motivated hate crime test, risk is higher within proximity of 300 meter of a public middle school and is higher within a proximity of 600 meter of a location from where a hateful tweet originated. The RRV represents the weight of influence for each factor relative to one another. For example, places affected by a risk factor with a RRV of 6 are twice as risky compared to places affected by risk factor with a RRV of 3.

Appendix B. Hate-speech classification methodology

Religion related hate speech classification

Description of Text Pre-processing, Feature Extraction and Classifiers Training

The tweets are pre-processed by converting the words to their lower cases, removing stop words, numbers and punctuations and stemming the remaining words.

For Bag of Words (BOW) feature extraction, single-word terms (with the minimum term frequency of 2) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used the value of each feature.

For N-grams (NG) feature extraction, 1-word, 2-word and 3-word terms (with the minimum term frequency of 2) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used the value of each feature.

For Typed Dependency (TD) feature extraction, 1-dependency, 2-dependency and 3-dependency terms (with the minimum term frequency of 2) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used the value of each feature.

For Embedding feature extraction, single-word terms (with the minimum term frequency of 2) are transformed into word vectors with 100 dimensions through setting the context window size of 2 and the batch size of 512 over 50 epochs. Each tweet is transformed into a document vector by averaging the values of the associated word vectors in each dimension.

Finally, the adopted classifier is trained on the NG features by using the fuzzy approach (Fuzzy).

Details on classification Performance

Table 4. Classification Performance on F-measure for the Yes class

Feature Extraction	SVM	NB	Fuzzy
BOW	0.877	0.852	0.898
NG	0.876	0.852	**0.912**
TD	0.262	0.209	0.262
Embedding	0.675	0.288	0.608

Data Sampling

The data set contains 1146 tweets, where 146 of them are annotated as hateful ones. For collection of the hate speech instances, a public data set, which contains 80k tweets collected via Twitter API, was downloaded at: https://github.com/ENCASEH2020/hatespeech-twitter. Each of the 80k tweets was annotated as one of the four types, namely, normal, hateful, abusive and spam.

We used the IDs of the 80k tweets provided at the above web page for retrieval of the text of the tweets. Due to the case that some tweets were deleted or some users were suspended before the retrieval, we finally obtained 65898 tweets in total. We selected all the tweets annotated as hateful for subsampling of US hate speech instances. In particular, we used a list of US cities and states (in full names or acronyms) as keywords for identifying if each hateful tweet was posted in the US. Finally, we obtain 898 US hate speech instances in total.

Furthermore, we used a list of religion related keywords, such as ‘Muslim’ and ‘Islam’, for selecting instances for the hate class. Also, we found that the names “Trump” and “Obama” frequently appeared in the tweets that contain religion related keywords and thus the names were also used to increase the likelihood of augmenting hate speech sample. Finally, 146 hateful tweets were obtained and 1000 non-hateful tweets were randomly selected from the LA county data set, which make up the data set used for training the classifiers.

Race related hate speech classification

Description of Text Pre-processing, Feature Extraction and Classifiers Training

The tweets are pre-processed by converting the words to their lower cases, removing stop words, numbers and punctuations and stemming the remaining words.

For Bag of Words (BOW) feature extraction, single-word terms (with the minimum term frequency of 2) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used the value of each feature.

For N-grams (NG) feature extraction, 1-word, 2-word and 3-word terms (with the minimum term frequency of 2) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used the value of each feature.

For Typed Dependency (TD) feature extraction, 1-dependency, 2-dependency and 3-dependency terms (with the minimum term frequency of 2) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used the value of each feature.

For Embedding feature extraction, single-word terms (with the minimum term frequency of 2) are transformed into word vectors with 100 dimensions through setting the context window size of 2 and the batch size of 512 over 50 epochs. Each tweet is transformed into a document vector by averaging the values of the associated word vectors in each dimension.

Finally, the adopted classifier is trained on the NG features by using Naïve Bayes (NB).

Details on classification Performance

Table 5. Classification Performance on F-measure for the Yes class

Feature Extraction	SVM	NB	Fuzzy
BOW	0.843	0.848	0.830
NG	0.860	**0.863**	0.860
TD	0.033	0.000	0.033
Embedding	0.725	0.430	0.611

Data Sampling

The data set contains 1117 tweets, where 117 of them are annotated as hateful ones. For collection of the hate speech instances, a public data set, which contains 80k tweets collected via Twitter API, was downloaded at: https://github.com/ENCASEH2020/hatespeech-twitter. Each of the 80k tweets was annotated as one of the four types, namely, normal, hateful, abusive and spam.

We used the IDs of the 80k tweets provided at the above web page for retrieval of the text of the tweets. Due to the case that some tweets were deleted or some users were suspended before the retrieval, we finally obtained 65898 tweets in total. We selected all the tweets annotated as hateful for subsampling of US hate speech instances. In particular, we used a list of US cities and states (in full names or acronyms) as keywords for identifying if each hateful tweet was posted in the US. Finally, we obtain 898 US hate speech instances in total.

Furthermore, we used a list of religion related keywords, such as ‘racism’, ‘black’ and ‘white’, for selecting instances for the hate class. Also, we found that the names “Trump” and “Obama” frequently appeared in the tweets that contain race related keywords and thus the names were also used to increase the likelihood of augmenting hate speech sample. Finally, 117 hateful tweets were obtained and 1000 non-hateful tweets were randomly selected from the LA county data set, which make up the data set used for training the classifiers.

Sexual orientation related hate speech classification

Description of Text Pre-processing, Feature Extraction and Classifiers Training

The tweets are pre-processed by converting the words to their lower cases, removing stop words, numbers and punctuations and stemming the remaining words. For Bag of Words (BOW) feature extraction, single-word terms (with the minimum term frequency of 2) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used the value of each feature.

For N-grams (NG) feature extraction, 1-word, 2-word and 3-word terms (with the minimum term frequency of 2) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used the value of each feature.

For Typed Dependency (TD) feature extraction, 1-dependency, 2-dependency and 3-dependency terms (with the minimum term frequency of 2) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used the value of each feature.

For Embedding feature extraction, single-word terms (with the minimum term frequency of 2) are transformed into word vectors with 100 dimensions through setting the context window size of 2 and the batch size of 512 over 50 epochs. Each tweet is transformed into a document vector by averaging the values of the associated word vectors in each dimension.

Finally, the adopted classifier is trained on the NG features by using Support Vector Machine (SVM).

Details on classification Performance

Table 6. Classification Performance on F-measure for the Yes class

Feature Extraction	SVM	NB	Fuzzy
BOW	0.682	0.714	0.646
NG	**0.727**	0.713	0.660
TD	0.350	0.258	0.320
Embedding	0.317	0.211	0.470

Data Sampling

The data set contains 1182 tweets, where 182 of them are annotated as hateful ones. For collection of the hate speech instances, a public data set, which contains 80k tweets collected via Twitter API, was downloaded at: https://github.com/ENCASEH2020/hatespeech-twitter. Each of the 80k tweets was annotated as one of the four types, namely, normal, hateful, abusive and spam.

We used the IDs of the 80k tweets provided at the above web page for retrieval of the text of the tweets. Due to the case that some tweets were deleted or some users were suspended before the retrieval, we finally obtained 65898 tweets in total. We selected all the tweets annotated as hateful for subsampling of US hate speech instances. In particular, we used a list of US cities and states (in full names or acronyms) as keywords for identifying if each hateful tweet was posted in the US. Finally, we obtain 898 US hate speech instances in total.

Furthermore, we used a list of sexual orientation related keywords, such as 'gay', 'homosexual', 'heterosexual' and 'bisexual', for selecting instances for the hate class. Also, in order to increase the likelihood of augmenting hate speech sample, we also added some terms (e.g. 'man', 'men', 'woman' and 'women') that show explicitly sexual identity and could have a high likelihood of relating to sexual orientation. Finally, 182 hateful tweets were obtained and 1000 non-hateful tweets were randomly selected from the LA county data set, which make up the data set used for training the classifiers.

General hate speech classification

Description of Text Pre-processing, Feature Extraction and Classifiers Training

The tweets are pre-processed by converting the words to their lower cases, removing stop words, numbers and punctuations and stemming the remaining words.

For Bag of Words (BOW) feature extraction, single-word terms (with the minimum term frequency of 5) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used as the value of each feature.

For word N-grams (NG) feature extraction, 1-word, 2-word, 3-word, 4-word and 5-word terms (with the minimum term frequency of 5) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used as the value of each feature.

For character N-grams (NG) feature extraction, 1-character, 2-character, 3-character, 4-character and 5 character terms (with the minimum term frequency of 5) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used as the value of each feature.

For Typed Dependency (TD) feature extraction, 1-dependency, 2-dependency, 3-dependency, 4-dependency and 5-dependency terms (with the minimum term frequency of 5) are extracted as the features and Term Frequency-Inverse Document Frequency (TF-IDF) is used as the value of each feature.

For Embedding feature extraction, single-word terms (with the minimum term frequency of 5) are transformed into word vectors with 100 dimensions through setting the context window size of 2 and the batch size of 512 over 50 epochs. Each tweet is transformed into a document vector by averaging the values of the associated word vectors in each dimension.

While SVM, NB and a fuzzy approach were used, respectively, for training a classifier on each of the feature sets extracted using the above methods. The best performing individual classifier results from using the character N-grams (NG) feature extraction method and the SVM algorithm, as shown in Table 1. Based on the results, the random subspace method was used to create an ensemble of SVM classifiers trained on NG (character) features, leading to a further improvement of the performance.

Finally, the adopted classifier is an ensemble of classifiers trained on the NG (character) features by using random subspace for ensemble creation and Support Vector Machine (SVM) for training of base classifiers.

Details on classification Performance

Table 7. Classification Performance on F-measure for the Yes class using a single classifier

Feature Extraction	SVM	NB	Fuzzy
BOW	0.687	0.443	0.470
NG(Word)	0.682	0.471	0.486
NG(Character)	**0.797**	0.725	0.720
TD	0.464	0.355	0.408
Embedding	0.610	0.349	0.551

Table 8. Classification Performance on F-measure for the Yes class using an ensemble of SVM classifiers

Ensemble Setting	NG(Character)
Random Subspace (60% features for each subset)	0.804
Random Subspace (65% features for each subset)	0.807
Random Subspace (70% features for each subset)	0.807
Random Subspace (75% features for each subset)	**0.810**

Data Sampling

The data set contains 11185 tweets, where 3129 of them are annotated as hateful ones. For collection of the hate speech instances, a public data set, which contains 80k tweets collected via Twitter API, was downloaded at: https://github.com/ENCASEH2020/hatespeech-twitter. Each of the 80k tweets was annotated as one of the four types, namely, normal, hateful, abusive and spam.

We used the IDs of the 80k tweets provided at the above web page for retrieval of the text of the tweets. Due to the case that some tweets were deleted or some users were suspended before the retrieval, we finally obtained 65898 tweets in total. We selected all the tweets annotated as hateful for subsampling of US hate speech instances. In particular, we used a list of US cities and states (in full names or acronyms) as keywords for identifying if each hateful tweet was posted in the US. Finally, we obtain 898 US hate speech instances in total.

Furthermore, another Twitter hate speech data set, which contains 31962 tweets in total and 2242 hateful tweets, was downloaded at: https://www.kaggle.com/vkrahul/twitter-hate-speech.

Afterward, the 2242 hateful tweets were taken for pre-processing, i.e. removing hashtags, mentions and URLs. Some instances become empty strings after the above pre-processing so we remove such instances ending up with 2231 hateful instances. These instances are combined with the previously obtained 898 instances. Finally, 3129 hateful tweets were obtained and 8056 non-hateful tweets were randomly selected from the LA county data set, which make up the data set used for training the classifiers.