Policy and Internet

Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making

Pete Burnap and Matthew L. Williams

The use of "Big Data" in policy and decision making is a current topic of debate. The 2013 murder of Drummer Lee Rigby in Woolwich, London, UK led to an extensive public reaction on social media, providing the opportunity to study the spread of online hate speech (cyber hate) on Twitter. Human annotated Twitter data was collected in the immediate aftermath of Rigby's murder to train and test a supervised machine learning text classifier that distinguishes between hateful and/or antagonistic responses with a focus on race, ethnicity, or religion; and more general responses. Classification features were derived from the content of each tweet, including grammatical dependencies between words to recognize "othering" phrases, incitement to respond with antagonistic action, and claims of well-founded or justified discrimination against social groups. The results of the classifier were optimal using a combination of probabilistic, rule-based, and spatial-based classifiers with a voted ensemble meta-classifier. We demonstrate how the results of the classifier can be robustly utilized in a statistical model used to forecast the likely spread of cyber hate in a sample of Twitter data. The applications to policy and decision making are discussed.

KEY WORDS: Twitter, hate speech, Internet, policy, machine classification, statistical modeling, cyber hate, ensemble classifier

Introduction

Research using traditional surveys and interviews has identified that crimes with a prejudicial motive are influenced in the short term by singular events such as widely publicized murders (Phillips, 1980, on homicide), riots (Bobo, Zubrinsky, Johnson, & Oliver, 1994, on race relations), and court cases and terrorism (King & Sutton, 2013, on hate crime; Legewie, 2013, on anti-immigrant sentiment). Hate crimes have been shown to cluster in time and tend to increase, sometimes dramatically, in the aftermath of an antecedent or "trigger" event (King & Sutton, 2013). The impacts of hate crime on individuals and communities are well documented (see Williams & Tregidga, 2014) and most research is preoccupied with where hate crimes happen (risky neighborhoods, demographic factors, etc.), while there is little research that looks at when they happen. King

1944-2866 © 2015 The Authors. *Policy & Internet* published by Wiley Periodicals, Inc. on behalf of Policy Studies Organization
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
Published by Wiley Periodicals, Inc., 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford, OX4 2DQ.

and Sutton (2013) report that 481 hate crimes with a specific anti-Islamic motive occurred in the year following 9/11, with 58 percent of them perpetrated two weeks following the event (4 percent of the at-risk period). Such evidence demonstrates that crimes entailing a prejudicial motive often occur in close temporal proximity to galvanizing events, such as terrorist attacks. It is during this period that decision makers, particularly those responsible for minimizing the risk of social disorder through community reassurance, local policing, and the online governance of hateful and antagonistic content, require additional information on the likelihood of disruption.

Hate crimes are communicative acts, often provoked by events that incite retribution in the targeted group, toward the group that share similar characteristics to the perpetrators (King & Sutton, 2013). Collecting and analyzing temporal data allows decision makers to study the escalation, duration, diffusion, and de-escalation of hate crimes following "trigger" events. However, decision makers are often limited in the information that can be obtained in the immediate aftermath of such events. When data can be obtained, they are often of low granularity, subject to missing information (hate crimes are largely unreported to the police), and invariably retrospective. However, the recent widespread adoption of social media offers a new opportunity to address these data problems. The continued growth of online social networks and microblogging Web services, such as Twitter, enable a locomotive, extensive and near real-time data source through which the analysis of hateful and antagonistic responses to "trigger" events can be undertaken. Such data affords researchers with the possibility to measure the online social mood and emotion following large-scale, disruptive, and emotive events such terrorist attacks in near real-time.

Twitter is a defensible and logical source of data for such analysis given that users of social media are more likely to express emotional content due to deindividuation (anonymity, lack of self-awareness in groups, disinhibition) (Festinger, Pepitone, & Newcomb, 1952). There is also a case history relating to the expression of hateful sentiment on social media in the United Kingdom, providing evidence of "real world" criminal justice response to, and therefore criminalization of, online acts of targeted hateful communication. For example, in 2012, Liam Stacey was sentenced to 56 days in prison for posting racially offensive comments on Twitter after a U.K. Premier League footballer's cardiac arrest, and in 2014, Declan McCuish was jailed for a year for tweeting racist comments about two Glasgow Rangers football players.

To date there has been very little research into the manifestation and diffusion of hate speech and antagonistic content in social media in relation to events that could be classed as "trigger" events for hate crimes. In 2013, the murder of Drummer Lee Rigby in Woolwich (London) by Islamic extremists led to an extensive social media reaction. Given the extreme terrorist motive and public nature of the actions it was conceivable that the public response might include written expressions of hateful and antagonistic sentiment toward a particular race, ethnicity, or religion, which could be interpreted as "hate speech." In this article, we present a supervised machine learning text classifier trained

and tested to identify online hate speech—or cyber hate—using data collected from Twitter in the immediate aftermath of Lee Rigby's murder. The data were annotated by human coders, who were asked to decide whether the tweets they were shown contained hateful and/or antagonistic responses toward minority groups.

As "Big Data" is a growing topic of study, and its use in policy and decision making a current subject of debate (González-Bailón, 2013), we discuss the use of supervised machine learning tools to classify a sample of "Big Data," and how the results can be interpreted for use in policy and decision making. Data from Twitter, and social media more generally, are exceptionally noisy and contain a great deal of grammatical variance, misinformation, and mundane chatter. Due to the poor veracity of such data in its raw form, its use in policymaking is somewhat hindered. A key contribution of this study is therefore the production of a machine classifier that could be developed into a technical solution for use by policymakers as part of an existing evidence-based decision-making process. Further contributions of the paper are the identification of nuanced features of cyber hate on social media using a particular type of syntactic relationship within text as a classification feature, and the application of an ensemble machine classifier to cyber hate. We include a section on how the classifier can be finely trained to suit the needs of policymakers, in order to minimize error and maximize confidence in results. We then demonstrate how the results of the classifier can be robustly utilized in a statistical model used to forecast the likely spread of cyber hate in a sample of Twitter data.

Related Work

The analysis of subjective language has been widely applied to the classification of opinions and emotions in text (Wiebe, 2005). Indeed, sentiment analysis, which aims to annotate text using a scale that is a measure of the degree of negative and positive sentiment within the text, has been applied to data collected from social media to determine emotional differences between genders on MySpace (Thelwall, Wilkinson, & Uppal, 2010b) and study levels of positive and negative sentiment in Facebook (Ahktar & Soria, 2009) and Twitter comments (Bollen, Goncalves, Ruan, & Mao, 2011; Thelwall, Buckley, & Paltogou, 2011) following real-world events.

Specifically focusing on hateful and/or antagonistic content, Greevy and Smeaton (2004) classified racist content in Web pages using a supervised machine learning approach with a bag-of-words (BoW) as features. A BoW approach uses words within a corpus as predictive features and ignores word sequence as well as any syntactic or semantic content. This approach can lead to misclassification due to word use in different contexts and, if words are used as a primary features for classification, it has been shown that combining sequential words into n-grams (list of words occurring in sequence from 1–n) improves classifier performance by incorporating some degree of context into the features (Pendar, 2007). However, an n-gram approach can suffer from the problem of high levels

of distance between related words—for example, if related words appear near the start and near the end of a sentence (Chen, Zhou, Zhu, & Xu, 2012). Dadvar, Trieschnigg, and de Jong (2013) used profane words in a social media account username, references to profanities and bullying-sensitive topics, and first and second person pronouns to classify antagonistic behavior on YouTube. Dinakar, Jones, Havasi, Lieberman, and Picard (2012) also focused on the identification of cyberbullying using a BoW approach, but also incorporated lists of profane words, parts-of-speech and words with negative connotations as machine learning features. Furthermore, they included a common-sense reasoning approach to classification by using a database that encoded particular knowledge about bullying situations (e.g., associating wearing dresses with males).

Burnap et al. (2013) developed a rule-based approach to classifying antagonistic content on Twitter and, similarly to Dinakar et al. (2012), they used associational terms as features. They also included accusational and attributional terms targeted at a person or persons following a socially disruptive event as features, in an effort to capture the context of the term use. Their results demonstrated an improvement on standard learning techniques (see also Williams et al., 2013). Chen et al. (2012) identified offensive content by using profanities, obscenities, and pejorative terms as features, weighted accordingly based on the associated strength of the term, as well as references to people. They also produced a set of rules to model offensive content, showing an improvement on standard machine learning approaches in terms of a much-reduced false negative rate.

Identifying syntactic constructs that tend to be insulting or condescending is a key function of the "Smokey" abusive message classification tool (Spertus, 1997), which uses pattern matching and syntactic positioning of words within text to classify content at a message level. Mahmud, Ahmed, and Khan (2008) followed a similar approach but also incorporated relationships between terms to identify "flaming" behavior online. The identification of syntactic relationships within text is possible via the development of parsing tools such as the Typed Dependency parser from Stanford (Marneffe, MacCartney, & Manning, 2006), though this has yet to be applied to hate speech.

Data Collection

We collected the study data set from Twitter during a two-week time window following the "trigger" event—the murder of Drummer Lee Rigby in Woolwich, London, UK on May 22, 2013. To ensure we maximized the collection of data surrounding the event we used the search term "woolwich," which would include many references to the events at Woolwich and also the main hashtag surrounding the event "#woolwich." The hashtag convention is widely used on Twitter to link an individual's thoughts and comments to an event.

The two-week data collection window was imposed based on three factors. First, existing research indicates that public interest in events typically spikes a short time after the event, and then rapidly declines (Downs, 1972). Second, this

first point was confirmed by tracking the search term "Woolwich" using the Google Trends service,[1] which records the relative number of searches performed on Google over time. Within two weeks, the use of "Woolwich" in Google searches had almost returned to preevent levels. Third, more than half of all hate-related attacks following 9/11 occurred within two weeks of the event; and we wanted to measure the immediate reaction to such events and capture data that perhaps would not otherwise be available to policy and decision makers due to the time taken to collect, record, and process hate crime results, and therefore be proactive in the first two weeks to reduce harm to targeted social groups in an appropriate manner. Social media data lend themselves to this purpose given their inherent fine-grained temporal characteristics. Tweets are produced by the second, while curated and administrative data have a much higher degree of latency in terms of both availability to decision makers and measurement of reaction. A total of 450,000 tweets were collected during the study window.

Data Annotation—Crowdsourcing

Building models to classify data according to a predefined coding scheme is an essential task in digital social research, used for the purposes of understanding social interactions, beliefs, emotions, and the like. In this research, once the Twitter data were collected, we built a supervised machine learning classifier to distinguish between hateful and/or antagonistic responses with a focus on race, ethnicity, religion, and more general responses, following the event. To complete this subjective task using large-scale data analytics, which is absolutely necessary for the volumes of data produced, we used machine classifiers to learn the features of tweets that are indicative of the class they belong to (cyber hate or general response).

Once features were learned, we applied the model to the whole data set. However, it was essential to understand and explain the limitations of the learned model by producing model-specific classification performance results, such as precision and recall per class, and confusion matrices (terms that are explained in detail later). Thus, we needed a "gold standard" against which to test the classification model. Commonly, this is obtained by sampling from a larger data set and employing human annotators to label each data point (tweet) according to a coding frame (Burnap et al., 2013). The coding frame serves as a set of categories or classes into which each data point can be classified. Computationally crowdsourcing human annotations is now becoming popular, and Web services such as CrowdFlower or the Amazon Mechanical Turk provide programmatic application programming interfaces (APIs) through which researchers can automatically upload a data set, coding frame, and set of instructions for annotation. The results of the annotation tasks can then be split into training and testing data sets for machine learning.

From the 450,000 tweets collected, we sampled 2,000 to be human coded. Coders were provided with each tweet and the question: "is this text offensive or antagonistic in terms of race ethnicity or religion?" They were presented with a

ternary set of classes—yes, no, undecided. We utilized the CrowdFlower online service that allows for Human Intelligence Tasks, such as coding text into classes, to be distributed over multiple workers. Workers can sign up to the service to participate in tasks in return for micropayments (small payments set by the task creator based on the number of tasks completed to an acceptable standard). Task creators can also specify a range of worker requirements such as location and experience, and can verify the level of expertise via test questions. Results from workers can then either be accepted or rejected, based on level of agreement with other workers.

CrowdFlower recruits from its pool of workers until each unit of analysis (in this case, each tweet) is annotated by a minimum number of workers, as specified by the task creator. We required at least four human annotations per tweet as per the convention in related research (Thelwall, Buckley, Paltogou, Cai, & Kappas, 2010a). CrowdFlower provides an agreement score for each annotated unit, which is based on the majority vote of the trusted workers (Kolhatkar, Zinsmeister, & Hirst, 2013). Because CrowdFlower continues to recruit workers until the task is complete, there is no guarantee that all workers will annotate the same set of units. Indeed, in this case we had 158 workers contribute to the task, each annotating a sample of tweets. Therefore we cannot calculate traditional interrater reliability scores, such as Krippendorf's Alpha or Cohen's Kappa to determine agreement between all annotators. However, CrowdFlower has been shown to produce an agreement score that compares well to these classic measures (Kolhatkar et al., 2013).

Based on the output from our annotator task we can determine agreement on *each unit*. The purpose of the experiments performed in this article are to establish the accuracy of a machine classifier when annotating tweets as hateful and/or antagonistic or not, and thus it is the agreement score for the unit of analysis (each tweet), and not the overall human agreement for all units that is important for validation. We removed all tweets with less than 75 percent agreement and also those upon which the coders could reach an absolute decision (i.e., the "undecided" class)—again, following established methods from related research (Thelwall et al., 2010a). The results of the annotation exercise produced a "gold standard" data set of 1,901 tweets, with 222 instances of offensive or antagonistic content (11.68 percent of the annotated sample), which could be classed as cyber hate (referred to below as the "cyber hate sample"), and 1,679 instances of nonhateful or antagonistic commentary (88.32 percent), which we will refer to as "benign." Ten percent of each class was subsequently used as a sample from which to identify appropriate features to build a cyber hate classifier. This subsample was not used when testing the classifier.

Feature Selection

It was evident from the cyber hate sample that many of the terms used in cyber hate were expletives or derogatory, targeted at specific social groups. The sample contained words that are well known derogatory terms for black, asian,

and muslim social groups, as well as derogatory adjectives (e.g., "black savages"). It was evident that the words of the tweets were going to be particularly useful features for the classification task. Using the words of the text to be classified, known as a BoW technique, is not a particularly novel approach to text classification, but the frequency of particular unigram (single word) and bigram (two word) terms were overwhelming and needed to be utilized.

Of more interest from a sociological and common sense reasoning perspective were the numerous instances in the cyber hate sample of calls for collective action and hateful incitement toward social groups exhibiting protected characteristics. For instance, there were exclamations such as "send them home," "get them out," and "should be hung." These exclamations clearly follow a pattern that could be encoded in parts-of-speech notation [e.g., Verb, Pronoun, Noun; Verb, Pronoun, Adverb; Verb, Verb, Verb(PT)]. However, the benign sample also displayed an abundance of similar patterns, such as "leave them alone," or "they are peaceful." Thus, parts-of-speech tagging to produce features to inform the machine classifier was avoided, as it seemed highly likely to cause confusion between the classes. Instead, we implemented the Stanford Lexical Parser, along with a context-free lexical parsing model, to extract typed dependencies within the tweet text (Marneffe et al., 2006). Typed dependencies provide a representation of syntactic grammatical relationships in a sentence (or tweet in this case) that can be used as features for classification. The following example explains the meaning of such relationships and how they can be used as features to inform the machine classifier.

Consider the sentence:

> "Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home."

The typed dependency parser returns the following output:

[root(ROOT-0, Send-1), nsubj(home-5, them-2), det(home-5, all-3), amod-(home-5, back-4), xcomp(Send-1, home-5)]

Within the output we can see five instances of typed dependencies. The second instance (*nsubj(home-5, them-2)*) identifies a relationship between "home" and "them," with "home" being the fifth word in the sentence and "them" appearing before "home" as the second word. Word order within a sentence is preserved in the type dependency and provides a feature for classification as well as the syntactic relationship between words. The relationship identified by the parser in this case is *nsubj*, which is an abbreviation of *nominal subject*. This will include a noun phrase ("them"), which is the syntactic subject in the sentence, and an associated relational term ("home"). Linguistically therefore, the term "them" is associated with "home" in a relational sense. Sociologically, this is an "othering" phrase, which essentially distances "them" from "us" through the relational action of removing "them" to their "home," as perceived by the author of the tweet.

Similarly, the third typed dependency (*det(home-5, all-3)*) identifies a *det* relationship, which is short for *determiner*, where a link is established between a noun phrase and its determiner. The noun phrase here being "home" (as in a place) and the determiner being "all." Again, this falls into an "othering" behavior, suggesting that the entire social group to which the Woolwich perpetrators belonged should have a relationship with "home," which we can assume means the perceived "home" of the social group by the author of the tweet (i.e., "not my country"). This combination of linguistics and sociology potentially provides a very interesting set of features for the more nuanced classification of cyber hate, beyond the BoW approach that utilizes expletives and derogatory terms. It allows a more common-sense reasoning approach to classifying cyber hate by considering the integration of "othering" terms and calls for retribution action into the classification features.

Data Preprocessing and Feature Preparation

Each tweet was computationally transformed into a word vector—a list of all the individual words (tokens) in the tweet. All tokens we transformed to lower case to avoid capitalized versions of words being treated as separate features to lower case versions of the same word. Nonalphanumeric characters other than those present in emoticons and exclamatory punctuation were removed, stop words were removed, and we stemmed each token to ensure that multiple representations and tenses of a word could be considered as a single features; for example, "attacked," "attackers," and "attacking" can all be reduced to "attack" so the machine can consider the verb as a single predictive feature, as well as the various forms of the verb. Tokens within each tweet were then clustered into sequential groups of tokens, or n-grams, ranging from one to five tokens in length to preserve an element of context for each word by encapsulating their surrounding words within a feature. Single tokens, or unigrams, were prominent in the cyber hate sample in the form of expletives or derogatory terms. Two-token combinations, or bigrams, were also present in the form of combinations of expletives, adjectives, and derogatory terms. Three-token terms (trigrams) could represent "othering" and incitements of retributional action, such as "send them home" or "get them out." Four- and five-token terms contained extended but similar phrases.

The BoW approach used here is fairly unsophisticated as a feature identification method, as it weights each n-gram equally as a feature and is likely to lead to confusion within the classification task when words occur frequently in both classes. Therefore, two experiments were conducted at the classification stage where in the first experiment all n-grams were retained as classification features, while in the second, only hateful and derogatory terms sampled from an online racial slur database[2] were retained, and the remaining n-grams were removed. Classification results were produced for each experiment. It was expected that the hateful terms would be predictive of cyber hate, but we were interested to see if other terms were also statistically significant predictors.

To produce a more sophisticated classifier capable of learning the grammatical structure of tweets containing cyber hate, each tweet was transformed into a set of typed dependencies using the Stanford Parser. Each typed dependency was considered as a unigram feature, and we again performed clustering on all the typed dependencies in a tweet to identify groups of between one and three typed dependency n-grams that represented the syntactic structure of each tweet. The number of possible typed dependency relationships produced by the Stanford model is around 50, and we suspected that not all relationships would be useful for classification.

As with the BoW experiments, at the classification stage we performed a two-step approach. The first experiment involved testing the classifier using all typed dependencies as features. We then performed a meta-analysis to better determine which features were more statistically significant at classifying cyber hate. To achieve this we ran a Bayesian Logistic Regression (BLR) using the typed dependency features extracted from the 10 percent sample of gold standard cyber hate and benign tweets. We used the model output of the BLR to establish a list of statistical coefficients relating to the probability of each typed dependencies n-gram occurring in a hateful or antagonistic tweet. The list was sorted to identify the most likely forms of typed dependency *relationship* to occur in the cyber hate class, and these relationships alone were retained as predictive features when the classifier was retrained and re-evaluated in a second experiment.

Finally, we combined both experiments and produced a final testing scenario to determine if combining the BoW, typed dependencies, and hateful and derogatory n-grams would prove to be the optimal set of features.

Model Selection

Given our feature set of specific words and syntactic features, we aimed to create a set of results and related model that could be used to inform policymakers of the risk of cyber hate spreading online following events that are likely to incur a hateful or antagonistic response toward a specific social group. To produce experimental results we used the Java Weka machine learning libraries to develop a number of supervised classifiers that were trained and tested using the features discussed in the previous section. Each tweet was transformed into a feature vector—a list of attributes that represent the tweet for the purposes of training a classifier. Each vector included the actual class the tweet belonged to based on the human annotation exercises (reduced to a binary "Yes" or "No" as to whether it was hateful or antagonistic or not), and a list of n-grams that either included words, typed dependencies, or a combination of both, depending on the feature set used to train the classifier.

Given the prevalence of individual words or short combinations of words in the cyber hate sample, it was appropriate to implement a classifier that would make decisions based on the likelihood of feature occurrence. We implemented a BLR classifier as a probabilistic approach. This classifier identifies statistical coefficients for each feature in a vector based on the likelihood of that feature

appearing in any of the classes available ("Yes" or "No") and uses this to predict the classes of previously unseen tweets.

Rule-based approaches to classifying antagonistic content have been shown to produce promising results in previous research, and the case of cyber hate seemed similar to other work in its accusational and targeted construct. Therefore, we employed a Random Forest Decision Tree (RFDT) as a rule-based approach to classification. A decision tree approach was chosen because it iteratively identifies the feature from the vector that maximizes information gain in a classification exercise—or put another way, it quantifies the significance of how using one n-gram as a rule to classify a tweet as "Yes," reduces the uncertainty as to which class it belongs to. Performing this step multiple times creates a hierarchical and incremental set of rules that can be used to make classification decisions. A Random Forest implementation of a decision tree was used because it iteratively selects a random subsample of features in the training phase and trains multiple decision trees before predicting the outputs and averaging out the results, maximizing the reduction in classification error (Breiman, 2001). The approach combines the results of a number of decision trees to identify the optimal set of rules, which was appropriate in this case because of the amount of noise and grammatical variance within the training and testing data sets.

A Support Vector Machine (SVM) was also used to determine if a spatial classification model would improve or enhance on a probabilistic or rule-based model. Feature vectors are plotted in high-dimensional space, and hyperplanes (lines that separate the data points) are used to try to find the optimum way to divide the space such that the tweets belonging to "Yes" and "No" classes are separated. Multiple hyperplanes can be used and the optimal hyperplane will be the line that maximizes the separation between classes. The rationale for the use of an SVM classifier was to determine whether cyber hate tweets and general responses to an event could be separated by spatial differences in lexical or syntactic features, as well as with probability and rules to determine predictive feature efficiency.

In addition to the three individual classifiers, we also implemented an "ensemble" classifier where a combination of all three was used to make a final classification decision. We used a voting meta-classifier, which produces a classification result for each base classifier (BLR, RFDT, and SVM) during the training phase, before making a decision on which model to use based on its prediction accuracy. A choice can be made based on the base classifier with the maximum probability or minimum probability; the results of all base classifiers can be averaged; or a majority vote can be taken. We implemented the maximum probability to make classification decisions, based on selecting the classification function that is most statistically likely to reduce error.

Classification Results

A 10-fold cross-validation approach was used to train and test the supervised machine learning methods. This approach has previously been used for building machine classifiers for short text (e.g., Thelwall et al., 2010a). It functions by

iteratively training the classifier with features from 10 percent of the manually coded data set, and classifying the remaining 90 percent as "unseen" data, based on the features evident in the cases it has encountered in the training data. It then determines the accuracy of the classification process and moves on to the next iteration, finally calculating the overall accuracy.

The results of the classification experiments are provided in Table 1 using standard text classification measures of: *precision* (i.e., for class x, how often are tweets classified as x when they should not be—a measure of false positives); *recall* (i.e., for class x, how often are tweets not classified as x when they should be—a measure of false negatives); and *F-Measure*, a harmonized mean of precision and recall. The results for each measure range between 0 (worst) and 1 (best). The formulae for calculating these results are as follows (where TP = true positives, FP = false positives, TN = true negative, and FN = false negative):

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP}$$

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN}$$

$$\text{F} - \text{Measure} = 2 \times ((\text{P} \times \text{R})/(\text{P} + \text{R}))$$

Because of the specific interest in the accurate detection of hateful and antagonistic content, the results reported in Table 1 are the precision, recall and f-measure for the *Yes* class ONLY. The number of false positives (instances where benign content has been classified as cyber hate) and false negatives (where cyber hate has been classified as benign) are also reported. Table 2 provides results for the best performing classifier and includes both *Yes* and *No* classes, as well as an overall performance score. Table 3 presents the confusion matrix for the best performing classifier with a breakdown of classifier error.

In Table 1, the bold text indicates the best performance results for precision, recall, FP, and FN for each feature set. In cases such as the *n-gram hateful terms* feature set, the whole row is bold because there was no difference between the performance of the classifiers. The shaded areas indicate the best overall performing feature set for each classifier.

The results suggest that overall the most efficient features for classifying cyber hate are n-gram typed dependencies combined with n-gram hateful and antagonistic terms. In fact, the hateful terms alone achieved the same precision performance but had a lower performance for recall. The number of false negative results (missed instances of cyber hate) was 7 percent higher when using hateful terms alone. This is an interesting result as it provides evidence to suggest that human annotators identify hateful or antagonistic content on Twitter that does not necessarily contain hateful or antagonistic terms, and requires a more nuanced representation of what is deemed cyber hate when aiming to classify tweets.

The use of a more sophisticated set of features as well as a BoW has successfully contributed to this requirement. A 7 percent improvement may seem fairly small, but considering the size of the initial corpus was 450,000, and in the annotated random sample of these data around 11 percent was considered hate speech by the human annotators, we could infer that there were around 49,500

Table 1. Cyber Hate Classification Results

	BLR			RFDT			SVM			Voted Ensemble (Max Probability)		
	P	R	F	P	R	F	P	R	F	P	R	F
n-Gram words (1–5) with 2,000 features	0.76 FP = 46	0.67 FN = 74	0.71	0.76 FP = 38	0.55 FN = 99	0.64	**0.80 FP = 38**	0.69 FP = 69	**0.74**	0.73 FP = 58	**0.71 FN = 65**	00.72
n-Gram hateful terms	**0.89 FP = 19**	**0.66 FP = 75**	**0.76**	**0.89 FP = 19**	**0.66 FN = 75**	**0.76**	**0.89 FP = 19**	**0.66 FN = 75**	**0.76**	**0.89 FP = 19**	**0.66 FN = 75**	**0.76**
n-Gram words (1–5) with 2,000 features + hateful yerms	0.75 FP = 40	0.55 FN = 100	0.64	**0.81 FP = 21**	0.42 FN = 128	0.56	0.74 FP = 50	0.65 FN = 78	**0.69**	0.68 FP = 70	**0.66 FN = 75**	0.67
n-Gram typed dependencies	0.52 FP = 50	0.25 FN = 167	0.34	0.56 FP = 36	0.21 FN = 176	0.30	0.53 FP = 48	0.24 FN = 168	0.33	0.49 FP = 57	0.25 FN = 167	0.25
n-Gram reduced typed dependencies	**1** FP = 0	**0.18** FN = 183	**0.29**	0.97 FP = 1	0.14 FN = 190	0.25	**1** FP = 0	0.17 FN = 185	0.28	**1** FP = 0	**0.18** FN = 183	**0.29**
n-Gram reduced typed dependencies + hateful terms	**0.89 FP = 19**	**0.69 FN = 70**	**0.77**	**0.89 FP = 19**	**0.68 FN = 71**	**0.77**	**0.89 FP = 19**	**0.69 FN = 70**	**0.77**	**0.89 FP = 19**	**0.69 FN = 70**	**0.77**
n-Gram words (1–5) with 2,000 features + n-Gram reduced typed dependencies + hateful terms	0.87 FP = 16	0.50 FN = 111	0.63	**0.88** FP = 10	0.32 FN = 150	0.42	0.88 FP = 18	0.59 FN = 91	0.70	0.83 FP = 27	**0.60** FN = 88	**0.70**

Table 2. Voted Classifier Full Results

	Voted Classifier		
	P	R	F
Yes	0.89	0.69	0.77
No	0.96	0.98	0.97
Overall	0.95	0.95	0.95

instances of cyber hate in the corpus. Overlooking 7 percent of these would lead to more than 3,000 hateful or antagonistic tweets being missed; so for policy-making purposes, the 7 percent improvement achieved by introducing the typed dependency features is significant if an accurate snapshot of the level of hateful and antagonistic emotive responses to an event is to be achieved.

The number of false positives in the best performing classifiers was 19, which constitutes 0.009 percent of the test data. Other classifiers reduced the number of false positives to below 19 (to zero in once instance), but the recall performance in these instances was far below that of the best performing classifiers, meaning that a reduction in false positives was also accompanied by an increase in false negatives. It is essential to retain a balance of minimized false positives and false negatives. In all cases, the voted ensemble classifier matched or improved upon the recall of each of the individual base classifiers. This suggests that combining the output of the respective probabilistic, rule-based and spatial classifiers, and selecting the classification decision of maximum probability can assist policy and decision makers in reducing the oversight of hateful or antagonistic content. While the base classifiers all achieved fairly similar results using the most efficient features set, given the improvement of recall across all other experiments when using a voted classifier, it would seem pertinent to consider the use of the voted classifier as a first choice when applying the cyber hate classifier to unseen data.

The full results of the cyber hate classifier are reported in Table 2. It is clear that the precision and recall of the non-hateful responses is very high ($P = 0.96$, $R = 0.98$). The precision of the "Yes" class is also high ($P = 0.89$), showing a low number of false positives, but there are improvements to be made to the recall of the "Yes" class ($R = 0.69$) before significant confidence can be given to the results for policy- and decision-making purposes. Table 3 shows 70 misclassifications where cyber hate was classified as a benign response by the classifier, suggesting

Table 3. Voted Classifier Confusion Matrix

	Human Coders	
	Yes	No
Machine		
Yes	152	70
No	19	1,660

Table 4. Probabilistic Features Highly Likely to be in Cyber Hate

Typed Dependency	Qualitative Description
det(religion-5 a-4)	Determiner (a specific reference to a noun phrase) discussing "a" "religion" in a particular context
amod(people-7 black-6)	Adjectival modifier (a descriptive phrase related to a noun phrase) discussing "people" who are "black"
aux(burn-6 to-5) dobj(burn-6 korans-9)	Auxiliary (a form of "be," "do," or "have") action phrase using "burn" and "korans"
amod(muslim-40 black-39)	Adjectival modifier (a descriptive phrase related to a noun phrase) discussing "muslims" who are "black"
det(muslim-40 a-38) amod (muslim-40 black-39)	Determiner (a specific reference to a noun phrase) discussing "a" "muslim" in the context of a "black" "muslim"
dobj(told-4 you-5) amod (people-7 black-6)	Direct object (an accusatory object of the verb) "told" "you" (e.g., "I told you") in the context of "black" "people"
advmod(seen-3 just-2) dobj(seen-3 video-4) dobj(getting-9 shot-10)	Adverbial modifier (a descriptive phrase related to a verb) "just" "seen," that is commenting on what has just been witnessed
dobj(burn-6 korans-9)	Direct object (an accusatory object of the verb) "burn" "korans"

Note: Using a classifier to inform a statistical model.

a further refinement is required to detect more discrete hateful and antagonistic content.

To give some insight into the qualitative narrative of cyber hate we have provided some instances of typed dependencies that were probabilistically more likely to occur in cyber hate than the benign class in Table 4. We can see that the content of tweets focuses on a response to religious and ethnic minority social groups from the wider population (e.g., black muslims). There are phrases suggestive of incitement to respond with actions (e.g., burn Korans) and claims of well founded or justified discrimination against social groups (e.g., "I told you black people..."). Given this reflective and responsive narrative it would seem pragmatic to include more semantic rules and constructs into feature identification in future in order to improve classifier performance.

Cautionary Caveat

Once a supervised machine learning classifier has been developed it can be used on a larger sample to classify new and unseen data, and inform policy decisions directly or via additional models. First and foremost it is essential to remember that supervised machine learning classifiers build models of what they perceive to be the features indicative of specific classes—in this case, hateful and antagonistic content. As a result, if new or unseen features occur, such as different types of language or content with mixed meaning, it can cause confusion in the classifier and produce inaccurate results. We can classify new instances, but we must always bear in mind the limitations in the existing model (i.e., not all instances of cyber hate were identified by our model), and that variance in the way people respond to such events may compound this.

That said, what we have tried to achieve with the classifier is to assist human decision making using a machine to handle the large volumes of data produced by the general public in response to a large-scale emotive event. The results of the cyber hate classifier are reasonably high, especially when considering that around 5 percent of our human-annotated sample had to be removed because the three out of four humans could not agree which class a tweet belonged to. It is worth remembering that while machine learned models are not always accurate in their judgment, humans are also susceptible to disagreement and confusion.

Cyber Hate and Contagion Modeling

In the following example we demonstrate how the supervised machine learning classification model of cyber hate can be applied to the whole corpus of 450,000 tweets to help determine to what degree hateful or antagonistic content is spreading—a measure of the *contagion* effect of cyber hate in response to a specific event. This could help inform those responsible for minimizing the risk of social disorder through community reassurance, local policing, and the online governance of hateful and antagonistic content, as to whether cyber hate is likely to spread.

One way to measure the impact of cyber hate on the spread of information on Twitter is to treat cyber hate as a predictive feature in a statistical regression model where the dependent variable (the outcome you are trying to predict) is the number of retweets a tweet is likely to receive. Theoretically, the more retweets a tweet receives, the more people are likely to see it, increasing the risk of public exposure and opportunity to propagate and respond to cyber hate. By measuring the statistical associated strength of cyber hate within a model of retweet counts, we can determine the likelihood of hateful and antagonistic content being retweeted, and therefore spreading to a large number of people. We can define a tweet that has been retweeted a large number of times as an *information flow* (Lotan, 2011).

Table 5 shows the result of a zero-inflated negative binomial model of information flow "size." The dependent variable is a count measure of the number of retweets a tweet actually received following the Woolwich event. The statistical predictors of the count include the number of followers of the person sending the tweet, the time of day the tweet was sent, the content of the tweet (hashtags, URLs), the sentiment polarity (+ve, –ve), the number of press headlines on the day the tweet was made, and the type of agent sending the tweet (e.g., press, police, politician). The data for these features were all derived from the data set collected from Twitter. For more details on how these were derived we recommend the reader study a related paper that examined the social media reaction in greater detail (Burnap et al., 2014). In this instance we are only interested in the impact of cyber hate as an example of how machine classification can help inform the modeling of online social reaction.

If we look at the incidence rate ratio (IRR) column in Table 5 we can see the strengths of association for each predictor variable with the dependent "retweet"

Table 5. Zero-Inflated Negetive Binomial Regression Model Predicting Counts of Retweets

	Count of Retweets		
Poisson Model (Count/True Zeros)	Coef.	SE	IRR
TimeLagRT5	0.000**	0.000	1.000
Tweet count	−0.215**	0.015	0.807
Commute morning	0.030	0.048	1.030
Work	0.014	0.038	1.015
Commute evening	0.074	0.045	1.077
Evening	−0.010	0.040	0.990
Ref: Commute night			
Sunday	−0.133**	0.041	0.875
Monday	−0.302**	0.047	0.739
Tuesday	−0.344**	0.051	0.709
Thursday	−0.365**	0.058	0.694
Friday	−0.311**	0.044	0.733
Saturday	−0.122**	0.052	0.853
Ref: Wednesday			
Hashtag	0.217**	0.025	1.242
URL	0.439**	0.026	1.551
Sentiment	0.322**	0.018	1.380
Google search	0.005**	0.001	1.005
Press headlines	0.000*	0.000	1.000
News agent	1.460**	0.044	4.304
Police Agent	1.742**	0.408	5.708
Political agent	0.670**	0.150	1.954
Far right political agent	0.632*	0.327	1.882
Ref: Other agent			
Cyber hate speech	−0.604**	0.107	0.546
Constant	0.543**	0.126	1.721
Binomial model (inflation/excess zeros)			
Number of followers	−0.899**	0.017	-
Constant	4.586	0.063	-
Model fit			
Log-L		−92,196.36	
Chi-square		2,594.57	
Sig.		$p = 0.00$	
LRT for alpha = 0		$p = 0.00$	
Vuong		$Z = 45.00$; $p = 0.00$; $N^a = 210{,}807$	

Notes: *$p < 0.05$, **$p = <0.01.1$.[a]Reduction due to removal of retweets, leaving only original tweets.

count, as indicated by the IRR. We can use the IRR to report the strength of causal associations between certain factors and the information flow size, enabling us to identify quantitatively which factors are more important than others. Where an IRR >1, the difference is associated with a positive increase in the dependent variable (retweet count), so in the case of the "URL" variable which records whether or not a tweet contains a URL, the results indicate that the rate of retweet for tweets containing a URL is 1.55 times higher than the rate for tweets without a URL. Thus, a URL increases the chances of a tweet being retweeted. Where an IRR <1, there is a negative effect. If we look at the "Cyber Hate Speech" predictor we see the IRR is 0.55 (rounded to 2 decimal places), which means that the

inclusion of hateful or antagonistic content in a tweet reduces the rate of retweet by a factor of 0.55 (or 45 percent), suggesting that a response to this event that contains a hateful or antagonistic element, as determined by the machine classifier, is in fact reducing the likelihood of the tweet being widely spread.

For policymakers, the combination of the cyber hate machine classifier with the statistical predictive model of the retweet likelihood given the features of the tweet could be useful in determining the changing dynamic of cyber hate on Twitter over time, and as an event unfolds. At any point in time a new corpus of tweets can be collected via the Twitter API, and the number of retweets each tweet has received is available from the metadata provided by Twitter. If the machine classifier is used to detect cyber hate within the corpus, and the statistical model is subsequently rerun, the difference in IRR from one period of time to another can be illustrative of the changing dynamic of cyber hate in Twitter over time.

For instance, if the IRR for the "Cyber Hate Speech" predictor in the model is 0.55 at time x, and 0.75 at time y, it suggests an increase in the rate of retweets containing cyber hate and therefore provides an indication that hateful and antagonistic content is actually spreading more at time y. One limitation of our approach is that the classification of cyber hate is dependent upon the language used in response to an event, which may not predeterminable prior to an event. Therefore, from these results we are not suggesting that the hate propagation IRR could be compared to a preevent baseline, rather that while the event is unfolding policymakers can study fluctuations within the analysis window following the event. The utility of identifying fluctuations following the event include monitoring the enabling and inhibiting factors of propagation of hate, such as further connected events (e.g., a protest march, news coverage, published opinion pieces, and political speeches).

Conclusion

In this article we have developed a supervised machine learning classifier for hateful and antagonistic content in Twitter. The purpose of the classifier is to assist policy and decision makers in monitoring the public reaction to large-scale emotive events, such as the murder of Drummer Lee Rigby in Woolwich in 2013. Previous research showed that 58 percent of hate crimes following 9/11 were perpetrated two weeks following the event (4 percent of the at-risk period). Data are available in near-real time from online social networks and microblogging websites such as Twitter, which can allow us to monitor the prevalence of hateful and antagonistic responses online in the period immediately following the event, when risk of hateful responses is highest. Hateful and antagonistic responses have led to imprisonment of the person posting the tweet—possibly as part of a risk reduction response by the judicial system.

The classification results showed very high levels of performance at reducing false positives and produced promising results with respect to false negatives. Our implementation of individual probabilistic, rule-based, and spatial classifiers

performed similarly across most feature sets, but the combination of the classification output of these base classifiers using a voted meta-classifier based on maximum probability matched or improved on the recall of the base classifiers in every experiment, suggesting that an *ensemble* classification approach is most suitable for classifying cyber hate, given the current feature sets. This could be due to the noise and variety of types of response within the data, with some features proving more effective with different classifiers.

The novel inclusion of syntactic features using typed dependencies within tweets as machine learning features reduced the false negatives by 7 percent over the baseline BoW features, providing a significant improvement when considering the volumes of data produced in response to such events. Our corpus of 450,000 tweets was collected in the first two weeks following the event, and it would be extremely difficult for human effort to manually parse these data to determine levels of public antagonism within all the responses. The improvement in machine classification using typed dependencies also suggests that cyber hate comprises content that is not instantly identifiable by words that are traditionally associated with hateful and discriminatory remarks, and requires a more nuanced approach to text classification beyond words alone. For instance, there was a prevalence of "othering" terms, such as "send them home" and "get them out," as well as incitements to undertake hateful retribution such as "burn korans" and "should be hung." The typed dependency approach was able to identify these as useful features for classification.

We developed an illustrative example using cyber hate as classified by a machine as a predictive feature in a statistical regression model. The model produced IRRs for retweet activity given a set of features for each tweet. The model showed a reduction in retweet rate ratio when a tweet contained a hateful or antagonistic response, suggesting a stemming of the flow of content on Twitter when a tweet contained cyber hate. This combination of machine classification and statistical modeling can—while accepting the limitations of machines with respect to utilizing a learned set of predictive features that are not an absolute reflection of all the possible combinations and permutations of cyber hate characteristics—produce aggregated statistics and prevalence indicators for hateful and antagonistic responses to an event on social media, including the relative spread of cyber hate on Twitter over time.

Our results are reflective of the individual event under study so we make no claims as to the generalizability of the classifier or the statistical model. However, through this case study, we have established for the first time a method and a set of results that others could replicate following similar and disparate events, in order to build up a body of work from which more generalizable results can emerge. For example, following an event prompting a hateful homophobic response, a data analyst could collect data, perform the annotation task, and replicate our method. We hope this article acts as a clarion call for further research into cyber hate and its manifestation in social media around events, and the development of technical solutions that are informed by such research.

Pete Burnap, Ph.D., Lecturer, Cardiff School of Computer Science & Informatics, Cardiff University, Cardiff, UK [burnapp@cardiff.ac.uk].
Matthew L. Williams, Ph.D., Reader, Cardiff School of Social Sciences, Cardiff University, Cardiff, UK

Notes

This work was supported by the Economic and Social Research Council and Google Data Analytics Research Grant: "Hate Speech and Social Media: Understanding Users, Networks and Information Flows" [ES/K008013/1].

1. http://www.google.com/trends/.
2. http://www.rsdb.org/.

References

Ahktar, J., and S. Soria. 2009. "Sentiment Analysis: Facebook Status Messages." Stanford University Technical Report.

Bobo, L., C. Zubrinsky, J. Johnson, and M. Oliver. 1994. "Public Opinion Before and After a Spring of Discontent." In M. Baldassare, ed. *The Los Angeles Riots: Lessons for the Urban Future.* Boulder, CO: Westview Press, 103–34.

Bollen, J., B. Goncalves, G. Ruan, and H. Mao. 2011. "Happiness Is Assortative in Online Social Networks." *Artificial Life* 17: 237–51.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Burnap, P., O. Rana, N. Avis, M.L. Williams, W. Housley, A. Edwards, J. Morgan, and L. Sloan. 2013. "Detecting Tension in Online Communities With Computational Twitter Analysis." *Technological Forecasting and Social Change* DOI: 10.1016/j.techfore.2013.04.013. http://www.sciencedirect.com/science/article/pii/S0040162513000899.

Burnap, P., M.L. Williams, L. Sloan, O. Rana, W. Housley, A. Edwards, V. Knight, R. Procter, and A. Voss. 2014. "Tweeting the Terror: Modelling the Social Media Reaction to the Woolwich Terrorist Attack." *Social Network Analysis and Mining* 4: 206. http://download.springer.com/static/pdf/6/art%253A10.1007%252Fs13278-014-0206-4.pdf?auth66=1424370056_186d118a639c2acb3eb71690339364f7&ext=.pdf.

Chen, Y., Y. Zhou, S. Zhu, and H. Xu. 2012. "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety." In *Proceedings of the Fourth ASE/IEEE International Conference on Social Computing (SocialCom 2012)*, September 3–6, Amsterdam.

Dadvar, M., D. Trieschnigg, and F. de Jong. 2013. "Expert Knowledge for Automatic Detection of Bullies in Social Networks." In *Proceedings of the 25th Benelux Conference on Artificial Intelligence, BNAIC 2013*, November 7–8, Delft, the Netherlands, 57–64.

Dinakar, K., B. Jones, C. Havasi, H. Lieberman, and R. Picard. 2012. "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2 (3): Article 18.

Downs, A. 1972. "Up and Down With Ecology—The 'Issue-Attention Cycle.'" *Public Interest* 28: 28–50.

Festinger, L., A. Pepitone, and T. Newcomb. 1952. "Some Consequences of Deindividuation in a Group." *Journal of Social Psychology* 47: 382–89.

González-Bailon, S. 2013. "Social Science in the Era of Big Data." *Policy & Internet* 5: 147–60.

Greevy, E., and A.F. Smeaton. 2004. " Classifying Racist Texts Using a Support Vector Machine." In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, July 25–29, Sheffield, UK, 468–69.

King, R.D., and G.M. Sutton. 2013. "High Times for Hate Crime: Explaining the Temporal Clustering of Hate Motivated Offending." *Criminology* 51 (4): 871–94.

Kolhatkar, V., H. Zinsmeister, and G. Hirst. 2013. "Interpreting Anaphoric Shell Nouns Using Antecedents of Cataphoric Shell Nouns as Training Data." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, October 18–21, Seattle, Washington, 300–10.

Legewie, J. 2013. "Terrorist Events and Attitudes Toward Immigrants: A Natural Experiment." *American Journal of Sociology* 118: 1199–245.

Lotan, G., E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and D. Boyd. 2011. "The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian & Egyptian Revolutions." *International Journal of Communication* 5: 1375–405.

Mahmud, A., K.Z. Ahmed, and M. Khan. 2008. "Detecting Flames and Insults in Text." In *Proceedings of the 6th International Conference on Natural Language Processing (ICON-2008)*, December 20–22, CDAC Pune, India.

Marneffe, M., B. MacCartney, and C.D. Manning. 2006. "Generating Typed Dependency Parses From Phrase Structure Parses." Paper presented at the International Conference on Language Resources and Evaluation (LREC), May 24–26, Genoa, Italy.

Pendar, N. 2007. "Toward Spotting the Pedophile Telling Victim From Predator in Text Chats." In *Proceedings of the First IEEE International Conference on Semantic Computing*, September 17–19, Irvine, CA, 235–41.

Phillips, D.P. 1980. "Airplane Accidents, Murder, and the Mass Media: Towards a Theory of Imitation and Suggestion." *Social Forces* 58: 1001–24.

Purdam, K. 2014. "Citizen Social Science and Citizen Data? Methodological and Ethical Challenges for Social Research." *Current Sociology* 62: 374–92.

Spertes, E. 1997. " Smokey: Automatic Recognition of Hostile Messages." In *Proceedings of the Eighth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI)*, July 27–29, Providence, RI, 1058–65.

Thelwall, M., K. Buckley, and G. Paltogou. 2011. "Sentiment in Twitter Events." *Journal of the American Society for Information Science and Technology* 62: 406–18.

Thelwall, M., K. Buckley, G. Paltogou, D. Cai, and A. Kappas. 2010a. "Sentiment Strength Detection in Short Informal Text." *Journal of the American Society for Information Science and Technology* 61: 2544–58.

Thelwall, M., D. Wilkinson, and S. Uppal. 2010b. "Data Mining Emotion in Social Network Communication: Gender Differences in MySpace." *Journal of the American Society for Information Science and Technology* 61: 190–99.

Wiebe, J., T. Wilson, and C. Cardie. 2005. "Annotating Expressions of Opinions and Emotions in Language." *Language Resources and Evaluation* 39 (2–3): 165–210.

Williams, M.L., A. Edwards, P. Burnap, O. Rana, N. Avis, J. Morgan, and L. Sloan. 2013. "Policing Cyber-Neighbourhoods: Tension Monitoring and Social Media Networks." *Policing and Society* 23 (4): 461–81.

Williams, M.L., and J. Tregidga. 2014. "Hate Crime Victimization in Wales: Psychological and Physical Impacts Across Seven Hate Crime Victim Types." *British Journal of Criminology* 54 (5): 946–67.