Miró-Llinares *et al. Crime Sci* (2018) 7:15
https://doi.org/10.1186/s40163-018-0089-1

Crime Science

RESEARCH Open Access

Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments

Fernando Miró-Llinares[1], Asier Moneva[1*] and Miriam Esteve[2]

Abstract

With the objective of facilitating and reducing analysis tasks undergone by law enforcement agencies and service providers, and using a sample of digital messages (i.e., tweets) sent via Twitter following the June 2017 London Bridge terror attack ($N = 200{,}880$), the present study introduces a new algorithm designed to detect hate speech messages in cyberspace. Unlike traditional designs based on semantic and syntactic approaches, the algorithm hereby implemented feeds solely on metadata, achieving high level of precision. Through the application of the machine learning classification technique Random Forests, our analysis indicates that metadata associated with the interaction and structure of tweets are especially relevant to identify the content they contain. However, metadata of Twitter accounts are less useful in the classification process. Collectively, findings from the current study allow us to demonstrate how digital microenvironment patterns defined by metadata can be used to create a computer algorithm capable of detecting online hate speech. The application of the algorithm and the direction of future research in this area are discussed.

Keywords: Metadata, Cyber place, Hate speech, Twitter, Random Forests

Introduction

Moments after Khuram Shazad Butt used a van to run down pedestrians along the London Bridge, Twitter was boiling. At 22:01,[1] before the first call for help was received, the hashtag #PrayForLondon was trending[2] on a global level; 2 min later, the first message including the hashtag #StopIslam was posted; and an hour later, 18 million tweets with the hashtag #LondonBridge had been published. In all of these digital messages, users expressed solidarity and indignation over the attack. Unfortunately, some digital content also contained messages of happiness, hatred towards certain groups, and the glorification of violence.

Academic interest inherent in the impact of hate speech on the Internet is not new (Tsesis 2001). The possibilities of cyberspace to unify users and tear down some of the spatiotemporal barriers that limit the transmission of knowledge in physical space have augured an exponential increase both in the number of potential diffusers of such types of content and its receivers (Levin 2002). Such quantitative growth, however, has taken place simultaneously with an even more relevant qualitative change. The democratisation of electronic communications and technologies (Brenner 2017) and, in particular, the emergence of social networks as a brand-new social interrelation environment that has normalised communications through instant messaging systems has created a window of opportunity in which the expression of violent messages is no longer hidden or considered uncharacteristic of an ideological or political discussion.

We reconceptualize the role social networks play in the production of criminal events (e.g. hate speech) based

*Correspondence: amoneva@crimina.es
[1] CRÍMINA Research Center for the Study and Prevention of Crime, Miguel Hernández University of Elche, Avda. de la Universidad, s/n, Hélike building, 03201 Elche (Alicante), Spain
Full list of author information is available at the end of the article

[1] Time in London.

[2] A topic is considered trending in Twitter when it is popular in a specific location at a given moment.

on an adaptation of the principles of Criminology of Place to cyberspace (Miró-Llinares and Johnson 2018). The present paper addresses the potentially massive dissemination of radicalized content via Twitter through the introduction of an algorithm for the automatic detection of contents that contribute to mitigate their impact. This research demonstrates how patterns of hate speech can be detected in metadata,[3] basing the analysis on the relation between crime and place (Eck and Weisburd 1995; Sherman et al. 1989). Cyberspace, however, is not contained in a single "place" with homogeneous characteristics, but events occur in different cyber places inside of it and at different times (Miró-Llinares and Johnson 2018). The identification of these spatiotemporal patterns may help us to improve the algorithms based solely on content analysis. This method adds to quantitative efficiency by automatizing part of the analytic process and thereby reducing the complexity of content analysis needed to identify messages of hate speech. Furthermore, it adds to qualitative efficiency by increasing the ability to limit the attention on content by private entities or public authorities to content that is actually related to high-risk activities, that is the dissemination of hatred or radical content in cyberspace.

In the following section, a review of recent literature is conducted to summarise the existing approaches to hate speech detection in cyberspace. Then, a comprehensive explanation of the concept of "cyber place" based on the idea of convergence is provided to present the theoretical framework in which the algorithm is built on. Afterwards, an empirical study is reported on to show the performance of the system proposed with a sample of tweets. The results are then interpreted and discussed in terms of efficiency and innovation to conclude with a summary of the relevant contributions and developments this work provides.

Related work

There has been a normalisation of extreme situations in an environment visited daily by millions of users to obtain the latest news and to socialise that is also used for propaganda purposes and the recruitment of radicalised subjects (Berger and Morgan 2015). This situation has led European authorities who were already focused on social control (McGuire 2017) to increase social media surveillance and specially to create and use digital tools that employ complex algorithms to detect propaganda and extremist and hate speech content (Awan and Blakemore 2016) as well as to identify individuals in the process of radicalising (Edwards 2017).

Such tools for the early detection of radical content are based on the identification of patterns, but in order to achieve this aim, they utilise a variety of techniques of content analysis, including the following: (1) manual collection (Gerstendfeld et al. 2003), and sampling methods and crowdsourcing (Chatzakou et al. 2017; Magdy et al. 2015); (2) systematic keyword searches (Décary-Hétu and Morselli 2011); (3) data mining for sentiment analysis (Cheong and Lee 2011); (4) natural language processing (Nobata et al. 2016); and (5) different machine learning procedures (Ashcroft et al. 2015; Burnap and Williams 2015; Malmasi and Zampieri 2017; Sharma et al. 2018), including logistic regression models (Davidson et al. 2017), and neural networks (Djuric et al. 2015; Dos Santos and Gatti 2014) or. Although some of these tools employ metadata analysis in combination with semantic or syntactic methods (Schmidt and Wiegand 2017; Waseem and Hovy 2016), all of them focus their attention at the core of the analysis on the content of the message, meaning the words themselves or the relations among them, which implies a major drawback when analysing communicative environments as dynamic as social networks (Serra et al. 2017). To overcome these difficulties when analysing online hate speech, in this paper we focus instead on analysing the metadata features extracted from Twitter digital microenvironments that are relevant for hate speech dissemination.

Traditional microenvironments, digital microenvironments, and hate speech

Twitter, like other social networks, is not a concrete physical location but can be accessed from many places, and criminal microenvironments are usually thought of as the locations, places, or spaces where crimes occur. Traditionally, the analysis of these micro places has served the purpose to understand how convergence allowed for a criminal event to take place. Social networks are not places in the traditional geographic sense, but they are places in a relational sense, since they are environments "that are visited" in which people converge with other people and with content in different ways, depending on the characteristics of the particular digital environment or network. The combination of the people (i.e., accounts), who say things (i.e., tweets) to other people (i.e., other accounts), define unique digital microenvironments in cyberspace. Indeed, it is in this sense of "place" where some cybercrimes occur in certain digital places more often than in others (Miró-Llinares and Johnson 2018), which implies that the basic premises of environmental criminology in general, and crime patterns in particular, may be true for certain cybercrimes.

[3] The information that defines single data items (e.g., the number of times a tweet has been retweeted, or the number of followers an account has).

In particular, this approach refers to the idea that crime distribution is not random but is based on patterns determined by the different environmental elements of the places where victims and offenders converge and by the relevance of such places to the routine activities developed in the activity spaces (Brantingham and Brantingham 1981). This is similarly valid for hate speech and for similar behaviours such as the dissemination of terrorist propaganda and radicalisation messages. It is true that in these types of crimes, the relevant convergence is not occurring between offender and victim but between the sender and receiver of the message. However, the convergence remains necessary: it needs a place where the hate message is reflected, and where another (or others, as the quantity of receivers is irrelevant) perceives it, such that hate speech or radicalisation on the internet will occur in some places more frequently than in others at both macro and micro levels, given certain environmental parameters.

From a macro perspective, that is, in comparison with other "places" or social networks, Twitter is an environment of massive, interactive and immediate communication of content. Although it allows streaming communication (through Periscope) and direct messages to concrete users out of sight of the rest of network, Twitter works essentially as a public square in which stored and forward communication is used to express content that can be observed and shared by a large number of people (Marwick and Boyd 2011). If we add that political or ideological communication has become increasingly frequent on Twitter (Bode and Dalrymple 2016), it seems understandable that this social network is commonly used to disseminate hate speech (Schmidt and Wiegand 2017) and that it has become perhaps the favourite social network of extremist and terrorist groups for propaganda and the promotion of radicalisation to a wider audience (Berger and Morgan 2015; Veilleux-Lepage 2014; Weimann 2014).

In addition, Twitter's structural configuration, in particular the restriction on the length of messages (first 140 characters, now 280), limits the possibilities for interaction among users and makes both hate speech, which will not be the same as the content expressed in a different forum or on Facebook (Awan 2016), and the activities of radicals and terrorists based on such speech less focused on recruitment and more aimed at normalising and magnifying terrorist activity for soft sympathisers (Veilleux-Lepage 2014) as well as disseminating propaganda by redirecting users to other places in cyberspace (Weimann 2014). Furthermore, Twitter allows anonymity, although it is not the most common way of interacting (see Peddinti et al. 2014). Finally, despite its constant technical modifications, Twitter has not shown much efficiency with regard to withdrawing offensive, hate-related or radical content (Weimann 2014), either because of the technical ease involved in creating accounts and the immediate publication of tweets or because of its rather vague free speech policy, which makes requests for removal different in each country (Hsia 2017).

However, Twitter is not a homogeneous place where everything occurs in the same way everywhere inside it. It is well known, for example, that the temporal distribution of messages does not occur randomly (Miró-Llinares and Rodríguez-Sala 2016); that there are some profiles with more followers than others and that not all of them publish the same number of tweets (Lara-Cabrera et al. 2017); and that there are very different degrees of identity expression on this social network (Peddinti et al. 2014). This indicates that a microanalysis of the configural elements of digital microplaces may be helpful to detect the environmental patterns that determine the occurrence of an event. In addition, it seems similarly obvious that the micro units that are essential for such an analysis are accounts and tweets.

A tweet is the essential microplace because it is where a message is expressed and shown and is where other users can interact with it, while an account is the microplace from which the publication or the viewing of such messages is made available. Like every microplace, a Twitter account has certain characteristics that differentiate it from the rest. For instance, if an account's registration information coincides with the identity of a public personality, Twitter will verify the user account with a blue badge. At the same time, a user can include a brief personal biography in one's profile and even activate an option to geolocate tweets in such a way that when publishing a message, the geographic location of where the tweet was written can be attached. Furthermore, users can include other accounts in thematic groups called "lists", which are useful for seeing only those messages published by selected accounts in chronological order. The number of lists in which an account is included is reflected in its profile together with other parameters such as the number of tweets published, the number of tweets liked, and the number of followers as well as the number of users that the account follows.

Similarly, a variety of elements configure and define a message transmitted by tweet. Tweets have a structural limitation in relation to the extension of their content that permits only a maximum number of characters, whether alphanumeric or in the shape of small icons, known as emojis. The combination of these characters with a variety of other elements will define the content of the microplace and its scope. Such elements include mentions, which act as specific personal notification when they include the @

symbol before the name of the user; Uniform Resource Locators (URL), which allow the inclusion of a hyperlink to additional content, whether an image, a video, a GIF or a link to an external site; or hashtags, which are situational elements that serve to thematically tag the content of a tweet to connect messages and create communicative trends. Indeed, the result of combining all these elements conditions the ways and the frequency with which people interact with a tweet just by seeing it or by interacting with the message and promoting its dissemination through a retweet, which is a feature that allows the dissemination of messages to the followers of an account.

In any case, the relevance of the microplaces where more or less hatred can be found lies in the premise that motivates the present work: that hate speech, similar to other crimes in physical spaces and in cyberspace (Miró-Llinares and Johnson 2018), will also be distributed in certain patterns conditioned by the characteristics of the digital microenvironments where they occur. Thus, with regard to the special nature of hate speech in the sense of its dissemination via Twitter and taking into consideration the different structural characteristics of the microplaces that integrate it, there exists an opportunity to detect environmental patterns related to hate speech that could help to detect its early appearance in order to prevent, control or mitigate its impact.

The present study

The present study introduces and evaluates a new algorithm, designed to detect hate speech, through the identification of patterns found in the situational metadata of digital messages. Existing research has discovered various types of patterns on Twitter: linguistic and temporal (Williams and Burnap 2015), sociodemographic and temporal (Marcum et al. 2012), spatiotemporal and socioeconomic (Li et al. 2013) and sociodemographic (Sloan et al. 2015), among others. In addition, patterns have been found related to the metadata on other social networks: for example, those linked to certain content for the detection of cyberbullying on Instagram (Hosseinmardi et al. 2015), or the tagging of YouTube videos to identify deviant content (Agarwal et al. 2017). What has not yet been analysed, however, is whether such patterns are related to the environmental characteristics of the social media accounts and digital messages in relation to their configuration as microplaces.

To achieve the study's aim, we required a large sample of digital messages from Twitter, upon which data mining techniques could be applied. This would enable us to determine whether characteristics of this social network's microplaces are decisive with regard to determining the types of messages that will be published from or inside them. With the aim of finding a more efficient tweet classification criterion, two classification trees were implemented: one with account metadata as inputs and another with the tweet microplace's metadata. A detailed description of the sampling strategy, variables analysed, and analytic technique follows.

Sample and procedure

The data collection was performed through the Application Programming Interface (API) of Twitter, which allows users with developer permissions access to data for reading, writing or monitoring in real-time. Researchers that work with data from Twitter are already familiar with the constant changes experienced by their API, which may compromise the process of data gathering. To address this problem and to overcome the possible changes caused by the application, an algorithm for data gathering was developed (see Additional file 1: Appendix A) that is equipped with sufficient rigidity due to an exception management system: programming techniques that enable researchers to control the appearance of anomalies during the execution of a script. Additionally, a system was implemented that provides immediate alerts if the server experiences any problems, the connection is interrupted, or the API loses or receives new permissions. Through this system, it is possible to quickly resolve any adjustment problems regarding the requests sent to the server via the code and the responses from the API when new updates modifying the composition of the dataset occur.

Once the API access is obtained and after establishing convenient authentication parameters, information about a concrete event can be collected for subsequent analysis by using certain keywords or hashtags as search criteria. In this case, the terrorist attack perpetrated on London Bridge on 3 June 2017 has been selected. Once the data collection process has begun, the API can store up to 1% of the tweets published on Twitter based on pre-set search criteria. Thus, three filtering hashtags were selected to provide balanced sampling (see Miró-Llinares 2016): #LondonBridge, which refers neutrally to the event; #PrayForLondon, for solidarity content; and #StopIslam, which is a representative hashtag for radical expressions, Islamophobia in this case. The first two hashtags were trending topics at some point during the event, while the last one was also a trending topic during previous attacks, allowing us to make comparisons with other samples collected earlier. Through this procedure, over 3 days, a sample of more than 200,000 tweets was obtained ($N = 200{,}880$) that refer directly or indirectly to the selected event.

Independent variables: microplace characteristics

In addition to the content of the tweets, the semi-structured dataset [in JavaScript Object Notation (JSON) format] contains numerous fields that provide information on different elements of Twitter, including the microplaces of accounts and tweets. Once the dataset was pre-processed and high-value dispersion variables were eliminated together with record identifiers as well as those variables with a percentage of nulls higher than 25–30% (Hernández et al. 2004), the dataset was built. To build the dataset on which the classification tree was applied, there has been selected, on one hand, those variables that are related to the anonymity and the visibility of accounts and, on the other hand, to the structure and interaction of the tweets. These variables and others that were created from the aforementioned, together with each observation (i.e. tweet), comprise the dataset analysed in the present study.

The users' account has been identified as a microplace intimately related to their anonymity and the visibility of their actions, hence relevant for hate speech dissemination. Table 1 provides a detailed description of the variables related to the anonymity and visibility of the accounts that were used in the present study. Those variables that provide information about the person behind the profile, such as their name, interests, or area of residence were included within the anonymity category. A second set of variables measuring the visibility of the users' activity in Twitter such as message posting, the user's active period on the social network, and different forms of interaction with other users were included within the visibility category. Regarding the characteristics of an account, the variable "description" has been modified because the API returned the entire text field of users' biographies, and since the analysis of its content would have implied a subjective interpretation, a dichotomisation was applied (1, the user has a biography; 0, the user does not have a biography) to enable the classification tree to operate with these data.

Table 1 Account variables related to users' anonymity and visibility. Source: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object. Accessed 13 July 2018

Variable	Type	Description
Anonymity		
Verified	Boolean	When true, indicates that the user has a verified account
Description[a]	Boolean	When true, indicates that the user has included a biography in his or her account profile
Geoenabled	Boolean	When true, indicates that the user has enabled the possibility of geotagging their tweets
Visibility		
Day_count	Numeric	The number of days since the user account was created
Listed_count	Numeric	The number of public lists in which this user is a member
Statuses_count	Numeric	The number of Tweets (including retweets) issued by the user
Followers_count	Numeric	The number of followers the account currently has
Friends_count	Numeric	The number of users the account is following. Also known as followings
Favourites_count	Numeric	The number of tweets the user has liked in the account's lifetime

[a] New variables

Tweets themselves and their associated metadata have also been identified as potential predictors of hate speech dissemination. Some of these elements are related to the interaction a tweet generates, while others determine its structure. Within the interaction category, some interactive elements that favour the users' engagement in dissemination activities were included together with the timing of the tweet publication. The structure category comprises two variables that constrain the length of the text and consequently the content of the message. The group of variables from the microplace of a tweet is shown in Table 2. Regarding these elements, a few modifications have been made (see Additional file 1: Appendix B). Because the restriction on the number of characters when publishing a tweet is one of the most distinctive characteristics of Twitter that has an obvious communicative impact, we measured the length of the text in the messages in the sample. To this effect, short scripts were elaborated to identify both the codification of the emojis on Twitter and the character chains composing URL to subsequently extract them from the body of a message. Thus, it is possible to carry out a character count to determine the actual length of a message, and two new variables are used to measure the presence of emojis and URL. With a similar method, we were able to determine the number of mentions and hashtags in each message, and we codified the results using two more numerical variables.

Dependent variable: hate speech

With regard to the dependent variable, a tailored reading and the subsequent dichotomisation were carried out to determine whether the content of each tweet was neutral or hate speech. This method was chosen over semantic or syntactic approaches (e.g., Bag of Words)

Table 2 Tweet variables related to the interaction and the structure of messages. Source: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object. Accessed 13 July 2018

Variable	Type	Description
Interaction		
Mention_count[a]	Numeric	Number of mentions included in the text of the tweet
Hashtag_count[a]	Numeric	Number of hashtags included in the text of the tweet
Url[a]	Boolean	When true, indicates that the tweet includes a URL
Re*tweet*_count	Numeric	Number of times this tweet has been retweeted
Minute_count	Numeric	Number of minutes since the event happened and the tweet was issued
Structure		
Text_count[a]	Numeric	Number of characters in the message, excluding URL, emoji, and retweet structure characters (i.e., 'RT @username')
Emoji[a]	Boolean	Indicates whether the text of the tweet includes an emoji

[a] New variables

because these have shown weaknesses when dealing with specific messages such as humour or irony (Farías et al. 2016; Reyes et al. 2013). Plenty of investigations have addressed the problem of hate speech detection in social networks with such methodologies (e.g., Burnap and Williams 2015, in Twitter; Mariconti et al. 2018, in YouTube). Although there exists a profound dogmatic discussion in that regard, in the present study, a broad concept of hate speech was used to classify such messages that comprises all the expressions considered violent or hateful communication in the taxonomy elaborated by Miró-Llinares (2016). According to this classification, for a tweet to be considered hate speech, its content must include the following categories: (1) direct incitement/threat of violence, (2) glorification of physical violence, (3) an attack on honour and human dignity, (4) incitement to discrimination/hate and (5) an offense to the collective sensitivity. This classification task was therefore based on the subjective interpretation of a text, with the limitations derived from this method. To alleviate the effect of judges' subjective analysis of the messages ($n = 100$), the Kappa coefficient (Cohen 1960), which measures the degree of agreement, was applied to ensure accordance in the assessments and thus the reliability of the classification of the tweets. As can be observed in Table 3, and according to the criteria established by Landis and Koch (1977), "almost perfect" (p. 165) agreement was obtained among the three pairs of judges (0.81–0.89).

Table 3 Results of the applications of the Kappa coefficient to the three pairs of judges

Group	Value of κ
Judges A and B	0.81
Judges A and C	0.89
Judges B and C	0.88

Although previous studies that used the same classification methodology removed all retweets from the sample to filter original messages from their redundant replicas (Esteve et al. 2018; Miró-Llinares 2016; Miró-Llinares and Rodríguez-Sala 2016), this procedure was not adequate in this study because the data collection method through the API did not guarantee that all retweets fit the original tweets that bounced back. Thus, only duplicated tweets were removed, which left 35,433 remaining unique cases to be classified. After the judges classified these messages, duplicates were folded back into the dataset to calculate the hate speech prevalence in our sample: a total of 9488 (4.7%) out of 200,880 tweets.

Analytical strategy

Regarding the characteristics of the sample, to confirm the relevance of places in cyberspace, it is necessary to apply data mining techniques. Therefore, by making use of the Random Forests classifier technique (Breiman 2001), an algorithm was implemented to create a number of classifiers for tweets that divide the sample based on the filters generated by each of the variables included in the model (i.e., nodes). These classifiers grow from a randomized data set extracted from the main sample to train the model and fit its parameters. 70% of the sample comprises the training set and the remaining 30% constitutes the test set. This division was repeated 10 times to promote randomization. The training set was then balanced favouring the minority class (i.e., hate speech tweets), while the remaining data were included within the unbalanced test set (Table 4).

This training and testing process allow to control for anomalous or less consistent nodes and, hence, growing a non-overfitted, pruned tree. To define the most appropriate parameters for our algorithm, a series of computational experiments were carried out. These parameters were adjusted to reduce the forest's sensitivity to their value (Tuffery 2011).

When going through each node, the model asks each classifier whether the sample fulfils the condition established on it, thereby filtering the main sample and creating two subsamples: one that fulfils the condition and one that does not. The model then selects the best filtering among all trees and averages their individual estimations

Table 4 Training set and test set composition

Class	Training set	Test set
Neutral	6638	184,754
Hate speech	6638	2850
Total	13,276	187,604

to produce the final output. By creating several decision trees that learn from a predetermined training set, the Random Forest produces robust predictions. When the condition that defines a node reaches maximum classifying efficiency, it means that the model has reached a leaf node, and it classifies the corresponding subsample to the same class: hate speech or neutral content. This technique intends to demonstrate that the cyber place variables selected can be used to properly classify a part of the sample, thereby contributing to the automation of the process. Additionally, to avoid results to be positively or negatively influenced by the training set composition, we used κ-fold cross validation defining $\kappa = 5$ subsamples (Kuhn and Johnson 2013).

An overview of the methodology employed in the present paper can be found in the figure below (Fig. 1).

Results

As can be observed in Table 5, two classification models were implemented and then validated for each set of cyber place variables to classify our sample: one used account variables as predictors while the other used tweet variables. Since the vast majority of accounts issued a single message ($Min = 1.0$; $Q1 = 1.0$; $Mdn = 1.0$; $M = 1.3$; $Q3 = 1.0$; $Max = 126$), their associated metadata can be treated differently and therefore the performance of the algorithm between the two models can be compared. Whereas account variables related to visibility and anonymity of users produce a rather poor model performance, the variables related to interaction and the structure of the tweets produce very promising results.

Table 5 Algorithm maximum precision and validation scores according to account and tweet models

Model	Precision	Recall	F1-score	Fivefold
Account				
Neutral	0.99	0.65	0.79	
Hate speech	0.03	0.62	0.05	
Average/total	0.98	0.65	0.78	0.63
Tweet				
Neutral	1.00	0.87	0.93	
Hate speech	0.09	0.86	0.17	
Average/total	0.98	0.87	0.92	0.86

Parameters: number of estimators = 1000; maximum depth = 10

Overall, the ability to avoid false positives (i.e., Precision) is consistently higher when including tweet variables in the algorithm. Regarding the accuracy of the model, results also support the use of tweet metadata over account metadata when it comes to the correct classification of positive cases (i.e., Recall). Mean scores resulting from fivefold validation are also included.

More detailed information about the number of correctly and incorrectly classified messages for both models can be found in the resulting confusion matrix (Table 6). Attending to the final purpose of the algorithm, effort was put into reducing the incorrect classification of hate speech messages (i.e., false negatives).

Regarding the cyber place related variables used to classify the messages, Table 7 shows their specific relevance within the models. The importance score reflects the proportion of nodes that include a condition imposed by each of the variables listed. In the case of account metadata, results show that visibility related variables are more important for the output decision, while anonymity has a negligible impact. On the other hand, two tweet variables influence the decision process over the rest: the number of retweets under the interaction category

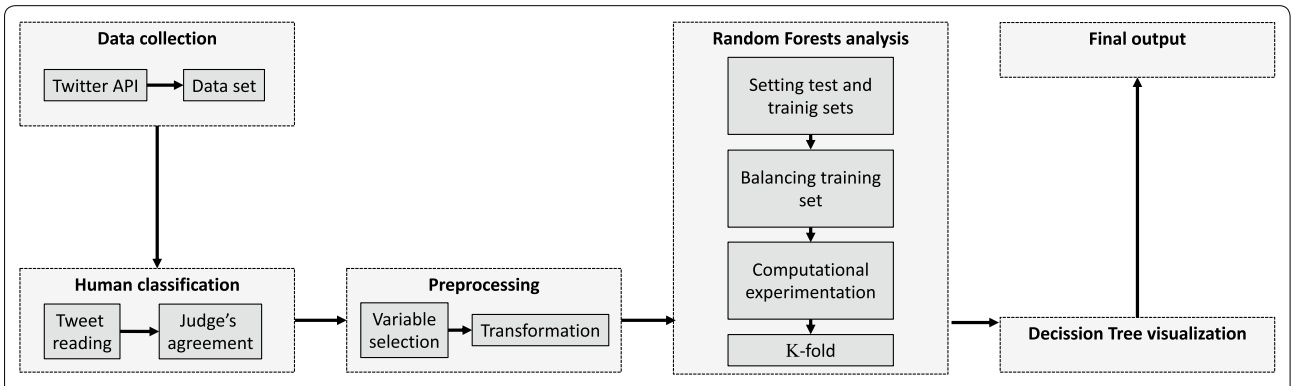

Fig. 1 Overview of the methodology employed

Table 6 Confusion matrixes according to account and tweet models

Model	Real	Prediction	
		Neutral	Hate speech
Account	Neutral	120,511	64,243
	Hate speech	1078	1772
Tweet	Neutral	160,676	24,078
	Hate speech	397	2453

(*importance* = 0.41), and the length of the text associated to the structure of the message (*importance* = 0.34).

To further understand which specific conditions a message must meet to be classified as neutral or hate speech by the algorithm, one of the decision trees produced with the Random Forests has been randomly selected and transformed into a flow chart (Fig. 2). As can be observed, the metadata patterns described by hate speech messages are different from those depicted by neutral communication. This flowchart shows some contents that describe clear patterns and can be classified using only one to three variables: retweet count, text count, and minute count. Even if temporal stamps appear to have low influence in the decision process (Table 7), they are crucial to define the content of the messages.

Table 7 Importance of the variables included in both models

Variable	Importance
Account	
Anonymity	
Verified	0.00
Description	0.02
Geoenabled	0.05
Visibility	
Day_count	0.16
Listed_count	0.12
Statuses_count	0.17
Followers_count	0.14
Friends_count	0.16
Favourites_count	0.17
Tweet	
Interaction	
Mention_count	0.02
Hashtag_count	0.08
Url	0.05
Retweet_count	0.41
Minute_count	0.08
Structure	
Text_count	0.34
Emoji	0.02

In summary, and as shown in the previous graph for the analysed sample, it is possible to define the environmental conditions that Twitter microplaces should have in order to differentiate the type of event occurring in them with certainty. These figures allow us to interpret the environmental patterns that arise from the sequential combination of account and tweet metadata associated to concrete messages. For example, if a message in our sample received between 6907 and 8138 retweets, was published 262 min after the attack, and had a text length of more than 107 characters (140 characters was the maximum allowed at the time of sampling), it was classified as a hate speech message; otherwise, it was classified as neutral (see Fig. 2).

Discussion

Based on the results of the present study, we can deduce that (1) digital microenvironment metadata can be used to detect hate speech patterns in cyberspace similar to the way spatiotemporal crime patterns in the physical environment can be found, and that (2) hate speech messages on Twitter describe environmental patterns that are different from neutral messages. This result is derived from the fact that hate speech messages are communicated via tweets, or through accounts, with specific environmental characteristics reflected in concrete metadata associated with the message. In other words, tweets and accounts containing hate speech have different characteristics from tweets and accounts containing neutral messages, which is a logical consequence of the different ways of communication currently available and messages that are expressed differently by taking advantage of the different possibilities of the digital environment.

The performance of the models reported on in this paper demonstrate that not all account variables related to the anonymity and visibility of users are relevant criteria to distinguish whether or not the content of a tweet is hate speech. This is perhaps due to the ease in proving them fake as an identifier element, and therefore, they are not relevant for differentiating between messages. More specifically, anonymity related variables have proven to be almost irrelevant for classification purposes, probably conditioned by their dichotomous categorization as the information gain is biased towards variables with large number of values (Quinlan 1986). Additionally, it does not seem entirely correct to make use of variables that describe a place where a crime will not occur just to determine the optimal environmental characteristics. As a matter of fact, the account is the microplace from which hate speech is published, but it is not where it manifests. In other words, in the present analysis, we are using the

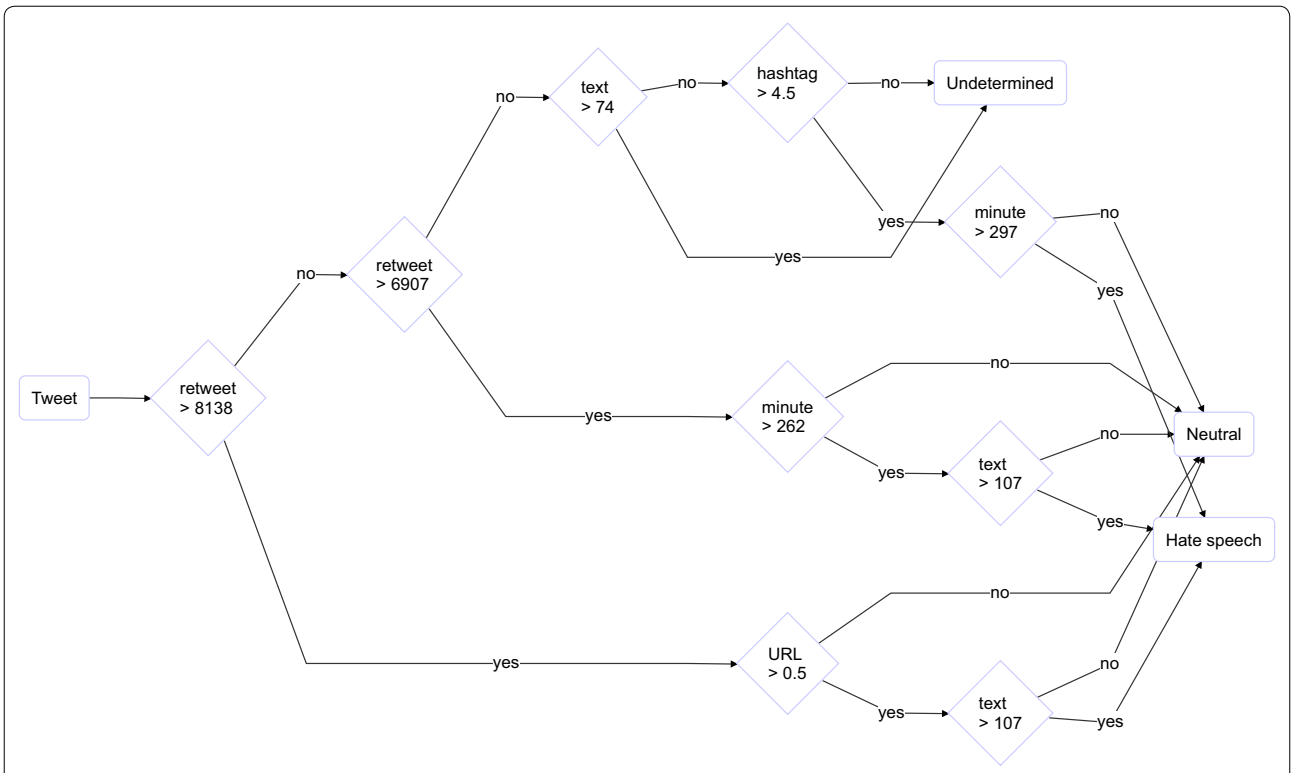

Fig. 2 Flowchart for a Random Forest classification tree according to the variables of the tweet (*depth* = 5)

characteristics of houses to define the context of a crime that occurs on that street. For this reason, we argue that the results are far from expected. We also believe that account metadata are not useful for classifying tweets because such data are associated with a dichotomised result of a particular tweet, and in this way, we might be incorrectly attributing radical characteristics to a not-so-radical place, such as an account that might have published just one hateful message. It seems reasonable to conclude that the intention of a user who posts a single hate speech message cannot be considered the same as a radical user who systematically disseminates hatred.

Conversely, in line with the work of Ferrara et al. (2016), the most important element for classifying the contents of a tweet are the retweets it receives, as they are closely related to the interaction generated and the visibility of a message. According to theory, hate speech users seek a greater dissemination of their ideas and might therefore include certain elements such as URL and hashtags that have been found to make messages more appealing to retweeting (Suh et al. 2010). On the other hand, and in the same way that the architectural design of a physical space can condition the occurrence of criminal events in certain places [for a review of Crime Prevention Through Environmental Design (CPTED), see Cozens et al. (2005)], the present study shows that the architecture of a tweet, especially the length of its text, is an essential element to determine the nature of the message. In line with previous research, tweet time stamps have shown that hate speech messages also cluster in time (Miró-Llinares and Rodríguez-Sala 2016), suggesting that certain cues activate radical responses on individuals more than others do. However, this analytical approach seems insufficient to explain why this is the case. In addition, the results confirm that tweet metadata have proved especially relevant to automatically identifying the specific microplaces where a criminal event will not occur (i.e., neutral tweets). There is no doubt these results are consistent in environmental terms, and we suggest that future investigations examine, for example, the role played by the anonymity variables of accounts in more detail, or the structural elements of a tweet regarding the dissemination of content.

Although the present study represents an initial stage of the investigation, it demonstrates the unquestionable capacity of the social sciences to provide important contributions to the fight against cyberterrorism (Maimon and Testa 2017), and, since the main goal is to automate the process of classifying messages regardless of platform, it offers relevant information in terms of ways to potentially improve the search algorithms for different

content, as it demonstrates that to detect this type of communication, we must focus not only on the content of a message but also on the environment in which it is expressed. In this sense, recent studies applying different lexical approaches for classifying tweets such as Support Vector Machines (SVM), Logistic Regression, or Random Forests, have obtained similar or inferior performances than the algorithm presented in this study, solely fed with metadata. Thus, while our Random Forest tweet model hits a F1-score of 0.92,[4] these previous attempts obtained F-measures of 0.77 (Burnap and Williams 2015), 0.90 (Davidson et al. 2017), and 0.76 (Sharma et al. 2018) respectively.

We further argue that the use of metadata to classify messages can help to overcome limitations that arise from the application of approaches such as Bag of Words to samples comprising texts in different languages. In this sense, we believe that a combination of lexical and metadata approaches would enhance the ability of state-of-the art approaches to detect radical communication in social networks. From a methodological point of view, it can also be argued that metadata yield benefit both in the extraction of variables, since they can be obtained through the API, and their simpler computation process compared to text-based variables.

It should be noted that the contribution of the present work is cross-cutting, as it goes beyond the frontiers of Twitter because all social networks host information of major importance in the metadata of their microplaces. However, this raises interesting questions regarding who has access to such metadata and whether the metadata should be made available to any user through open access systems or its access should be somehow limited. In any case, it seems that the current trend for many social networks is restrictive. Indeed, this has been the case for Facebook and Instagram, from which the extraction of information is becoming increasingly difficult. Until now, Twitter has continued to function with an open philosophy that allows researchers to collect a wide range of data.

Conclusion

Showing that environmental criminology can also be applied to cyberspace settings, this paper has introduced a brand-new theoretical framework to underpin online hate speech detection algorithms. Crime Pattern Theory principles and cyber place conceptualizations based on digital spaces of convergence (Miró-Llinares and Johnson 2018) have been adapted to identify the most relevant characteristics associated to hate speech dissemination in Twitter. This important contribution provides an analytical background that opens the way to study different forms of cybercrime relying on cyber place metadata.

Two relevant cyber places for hate speech dissemination have been identified in Twitter: accounts and tweets. Drawing on the Random Forests technique, tweet metadata proved to be more efficient in the classification of hate speech content than account metadata. This suggests that not all variables should be taken into account when building predictive models, restricting models to those variables which are supported by valid theoretical schemes for solving particular problems. In this case, and given the nature of hate speech, it is crucial to consider the essential variables for content propagation in social networks for predictive modelling. And even if this is not a methodology comparison paper, the precision scores obtained show that this approach is, at least, on par with other methods based on semantic approaches.

Although studying the entire population of digital messages on any platform is an unrealistic task, a sample of over 200,000 tweets gives us the ability to answer our research question, despite our inability to generalise the current findings to all Twitter events. This further leads to the fundamental question of whether hate speech has been properly measured, that is, whether hate speech content has been properly distinguished from what is not. Regardless of the appropriateness of the taxonomy used to identify hate speech or whether the judges properly classified the sample, it is certain that the chosen method differentiates between events, which has been shown in the aforementioned studies.

As an axiological analysis, the sample may not accurately reflect the prevalence of hate speech on Twitter, but it is true that any pragmatic analysis will never lead two researchers to draw identical conclusions given the nature of language and the circumstances of communication. In this sense, this study aimed to achieve the greatest possible accuracy between judges to enable the analysis to interpret each criterion based on an acceptable level of agreement. Further research should be conducted to be able to escalate the application of the idea behind the methodology proposed in the present study.

Finally, despite demonstrating the utility of metadata in terms of precision for classification purposes, future research should aim to (1) compare computational times when using metadata versus text variables to determine which technique is more efficient, (2) test the ability of metadata models to overcome language limitations by comparing their performance in samples of different languages, and (3) merge the application of metadata and lexico-syntactical approaches to reduce the number of false negatives and positives, and to subsequently obtain

[4] Similar F1-scores were obtained in different samples that were not included in this paper but used the same methodology.

even higher precisions with hate speech detection algorithms in cyberspace.

Additional files

Additional file 1. Appendix A: Pseudocode for obtaining the sample. Appendix B: Pseudocode for pre-processing the variables.

Additional file 2. Anonymized data set used for the present study. The anonymized variables are in red.

Abbreviations

API: Application Programming Interface; CPTED: Crime Prevention Through Environmental Design; JSON: JavaScript Object Notation; SVM: Support Vector Machines; URL: Uniform Resource Locator.

Authors' contributions

The theoretical framework and research question were initially stated by FM, while AM further developed this background. Then, ME obtained and preprocessed the sample required for the analysis. Variables were selected according to FM and AM approach and machine learning techniques were conducted by ME. Finally, FM and AM jointly interpreted the results and elaborated the discussion section and conclusions. All authors read and approved the final manuscript.

Author details

[1] CRÍMINA Research Center for the Study and Prevention of Crime, Miguel Hernández University of Elche, Avda. de la Universidad, s/n, Hélike building, 03201 Elche (Alicante), Spain. [2] Center of Operations Research, Miguel Hernández University of Elche, Elche, Spain.

Acknowledgements

We thank Dr. Timothy C. Hart from University of Tampa for his thoughts on digital microenvironments and his valuable comments that helped to improve the manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The dataset analysed in the present study is attached as Additional file 2 in format .xlsx. Please note that the original dataset was retrieved in format .json, but it is way too heavy to be attached (13 GB) and pretty difficult to read. In addition, some variables have been removed to protect the privacy of Twitter users.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 740773. This research has been funded by the Spanish Ministry of Education, Culture and Sports under FPU Grant Reference FPU16/01671.

Publisher's Note

Received: 4 May 2018 Accepted: 31 October 2018
Published online: 15 November 2018

References

Agarwal, N., Gupta, R., Singh, S. K., & Saxena, V. (2017). Metadata based multi-labelling of YouTube videos. In *7th International Conference on Cloud Computing, Data Science & Engineering-Confluence* (pp. 586–590). New York: IEEE.

Ashcroft, M., Fisher, A., Kaati, L., Omer, E., & Prucha, N. (2015). Detecting jihadist messages on Twitter. In *European Intelligence and Security Informatics Conference (EISIC)* (pp. 161–164). New York: IEEE.

Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the Facebook's Walls of Hate. *International Journal of Cyber Criminology, 10*(1), 1–20.

Awan, I., & Blakemore, B. (2016). *Policing cyber hate, cyber threats and cyber terrorism*. Abingdon: Routledge.

Berger, J. M., & Morgan, J. (2015). The ISIS twitter census: Defining and describing the population of ISIS supporters on Twitter. *The Brookings Project on US Relations with the Islamic World, 3*(20), 1–68.

Bode, L., & Dalrymple, K. E. (2016). Politics in 140 characters or less: Campaign communication, network interaction, and political participation on Twitter. *Journal of Political Marketing, 15*(4), 311–332.

Brantingham, P. L., & Brantingham, P. J. (1981). Notes on the geometry of crime. In P. J. Brantingham & P. L. Brantingham (Eds.), *Environmental criminology* (pp. 27–54). Beverly Hills, CA: Sage.

Breiman, L. (2001). Random Forests. *Machine Learning, 45,* 5–32.

Brenner, S. W. (2017). Nanocrime 2.0. In M. R. McGuire & T. J. Holt (Eds.), *The Routledge handbook of technology, crime and justice* (pp. 611–642). New York City, NY: Routledge.

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

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the Web Science Conference* (pp. 13–22). New York: ACM.

Cheong, M., & Lee, V. C. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers, 13*(1), 45–59.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cozens, P. M., Saville, G., & Hillier, D. (2005). Crime prevention through environmental design (CPTED): A review and modern bibliography. *Property Management, 23*(5), 328–356.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint* arXiv :1703.04009.

Décary-Hétu, D., & Morselli, C. (2011). Gang presence in social network sites. *International Journal of Cyber Criminology, 5*(2), 876–890.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 29–30). New York: ACM.

Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 69–78). COLING.

Eck, J. E., & Weisburd, D. (1995). Crime places in crime theory. In J. E. Eck & D. Weisburd (Eds.), *Crime and place* (pp. 1–33). Monsey, NY: Criminal Justice Press.

Edwards, A. (2017). Big data, predictive machines and security: the minority report. In M. R. McGuire & T. J. Holt (Eds.), *The Routledge Handbook of Technology, Crime and Justice* (pp. 451–461). New York, NY: Routledge.

Esteve, M., Miró-Llinares, F., & Rabasa, A. (2018). Classification of tweets with a mixed method based on pragmatic content and meta-information. *International Journal of Design & Nature and Ecodynamics, 13*(1), 60–70.

Farías, D. I. H., Patti, V., & Rosso, P. (2016). Irony detection in Twitter: The role of affective content. *ACM Transactions on Internet Technology, 16*(3), 19.

Ferrara, E., Wang, W. Q., Varol, O., Flammini, A., & Galstyan, A. (2016). Predicting online extremism, content adopters, and interaction reciprocity. In *International Conference on Social Informatics* (pp. 22–39). Berlin: Springer International Publishing.

Gerstenfeld, P. B., Grant, D. R., & Chiang, C. P. (2003). Hate online: A content analysis of extremist internet sites. *Analyses of Social Issues and Public Policy, 3*(1), 29–44.

Hernández, J., Ramírez, M. J., & Ferri, C. (2004). *Introducción a la minería de Datos*. Madrid: Pearson.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of cyberbullying incidents on the Instagram social network. *arXiv preprint* arXiv:1503.03909.
Hsia, J. (2017). Twitter trouble: The communications decency act in inaction. *Columbia Business Law Review, 2017,* 399–452.
Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, NY: Springer.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Lara-Cabrera, R., Gonzalez-Pardo, A., Barhamgi, M., & Camacho, D. (2017). Extracting radicalisation behavioural patterns from social network data. In *28th International Workshop on Database and Expert Systems Applications (DEXA)* (pp. 6–10). New York: IEEE.
Levin, B. (2002). Cyberhate: A legal and historical analysis of extremists' use of computer networks in America. *American Behavioral Scientist*, *45*(6), 958–988.
Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science, 40*(2), 61–77.
Magdy, W., Darwish, K., & Abokhodair, N. (2015). Quantifying public response towards Islam on Twitter after Paris attacks. *arXiv preprint* arXiv:1512.04570.
Maimon, D., & Testa, A. (2017). On the Relevance of Cyber Criminological Research in the Design of Policies and Sophisticated Security Solutions against Cyberterrorism Events. In G. LaFree & J. D. Freilich (Eds.), *The Handbook of the Criminology of Terrorism* (pp. 553–567). West Sussex, UK: Wiley & Sons.
Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media. *arXiv preprint* arXiv:1712.06427.
Marcum, C. D., Higgins, G. E., Freiburger, T. L., & Ricketts, M. L. (2012). Battle of the sexes: An examination of male and female cyber bullying. *International Journal of Cyber Criminology, 6*(1), 904–911.
Mariconti, E., Suarez-Tangil, G., Blackburn, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Serrano, J. L., & Stringhini, G. (2018). "You know what to do": Proactive detection of youtube videos targeted by coordinated hate attacks. *arXiv preprint* arXiv:1805.08168.
Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1), 114–133.
McGuire, M. R. (2017). Technology crime and technology control: Contexts and history. In M. R. McGuire & T. J. Holt (Eds.), *The Routledge handbook of technology, crime and justice* (pp. 35–60). New York City, NY: Routledge.
Miró-Llinares, F. (2016). Taxonomía de la comunicación violenta y el discurso del odio en Internet. *IDP. Revista de Internet, Derecho y Política, 22,* 82–107.
Miró-Llinares, F., & Johnson, S. D. (2018). Cybercrime and place: Applying environmental criminology to crimes in cyberspace. In G. J. N. Bruinsma & S. D. Johnson (Eds.), *The oxford handbook of environmental criminology* (pp. 883–906). Oxford: Oxford University Press.
Miró-Llinares, F., & Rodríguez-Sala, J. J. (2016). Cyber hate speech on Twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy. *International Journal of Design & Nature and Ecodynamics, 11*(3), 406–415.
Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 145–153). IW3C2.
Peddinti, S. T., Ross, K. W., & Cappos, J. (2014). On the internet, nobody knows you're a dog: A Twitter case study of anonymity in social networks. In *Proceedings of the Second Association for Computer Machinery Conference on Online Social Networks (COSN)* (pp. 83–94). New York: ACM.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.
Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation, 47*(1), 239–268.
Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10). ACL.
Serra, J., Leontiadis, I., Spathis, D., Blackburn, J., Stringhini, G., & Vakali, A. (2017). Class-based Prediction Errors to Detect Hate Speech with Out-of-vocabulary Words. In *Proceedings of the First Workshop on Abusive Language Online (ALW1)* (pp. 36–40). ACL.
Sharma, S., Agrawal, S., & Shrivastava, M. (2018). Degree based classification of harmful speech using twitter data. *arXiv preprint* arXiv:1806.04197.
Sherman, L. W., Gartin, P. R., & Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology, 27*(1), 27–56.
Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLoS ONE, 10*(3), 1–17.
Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing* (pp. 177-184). IEEE.
Tsesis, A. (2001). Hate in cyberspace: Regulating hate speech on the Internet. *San Diego L. Rev., 38,* 817.
Tuffery, S., (2011). *Data mining and statistics for decision making*. Wiley Series in Computational Statistics.
Veilleux-Lepage, Y. (2014). Retweeting the Caliphate: The role of soft-sympathizers in the Islamic State's social media strategy. In *6th International Terrorism and Transnational Crime Conference*.
Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of NAACL-HLT* (pp. 88–93). ACL.
Weimann, G. (2014). *New terrorism and new media*. Washington, DC: Commons Lab of the Woodrow Wilson International Center for Scholars.
Williams, M. L., & Burnap, P. (2015). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology, 56*(2), 211–238.