Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis

Journal:	*Behaviour & Information Technology*
Manuscript ID	TBIT-2018-0743.R2
Manuscript Type:	Original Research Study
Keywords:	online hate speech, intolerance prevention, Twitter, social minorities, sexual minorities

Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis

Abstract

It is increasingly assumed that cyberspace reflects patterns and practices that are enacted in offline social interactions. Though there are currently no statistics offering a global overview of online hate speech, both social networking platforms and organizations that combat hate speech have recognized that prevention strategies are needed to address this negative online phenomenon. While most cases of online hate speech target individuals on the basis of ethnicity and nationality, incitements to hatred on the basis of religion, class, gender and sexual orientation are increasing. This paper reports the findings of the "Italian Hate Map" – a national project aimed at identifying part-of-hate-speech in Tweets against six targets – women, gay and lesbian persons, immigrants, Jews, Muslims and disabled persons – and aggregating these Tweets according to geographical provenance. Using a lexicon-based method of semantic content analysis, 2,659,879 Tweets (from 879,428 Twitter profiles) were extracted over a period of 7 months; 412,716 of these Tweets contained negative terms directed at one of the six target groups. In the geolocalized Tweets, women were the most insulted group, having received 71,006 hateful Tweets (60.4% of the negative geolocalized tweets), followed by immigrants (12,281 tweets, 10.4%), gay and lesbian persons (12,140 tweets, 10.3%), Muslims (7,465 tweets, 6.4%), Jews (7,465 tweets, 6.4%) and disabled persons (7,230 tweets, 6.1%). The findings provide a real-time snapshot of community behaviors and attitudes against social, ethnic, sexual and gender minority groups that can be used to inform intolerance prevention campaigns on both local and national levels.

Keywords: online hate speech; intolerance prevention; Twitter; social minorities; sexual minorities

Introduction

Hate speech lies in a complex nexus with "free speech"; individual, group and minority rights; and dignity, liberty and equality. Although there is no universally agreed upon definition of the term, hate speech generally refers to expressions that incite harm (particularly discrimination, hostility or violence) to a particular target on the basis of the target's identification with a certain social or demographic group. It may include – but is not limited to – speech that advocates, threatens or encourages violent acts. Hate speech can also include expressions that foster a climate of prejudice and intolerance, on the assumption that such a climate may fuel targeted discrimination, hostility and violence (UNESCO 2015). Whilst traditionally hate speech has been thought to include any form of expression deemed offensive to a religious, racial, ethnic or national group, in the 1980s, these categories were broadened to include groups identifying with a particular gender, age, sexual orientation, marital status or physical capacity (Walker 1994). In a similar vein, Human Rights Watch defined hate speech as "any form of expression regarded as offensive to racial, ethnic and religious groups and other discrete minorities, and to women" (cited in Walker 1994, p. 8). Hate speech may occur in both offline and online contexts. In the latter context, it is often described as "hate speech online," "cyber harassment," "cyber bullying," "cyber abuse," "cyber incitement/threats" or "cyber hate" (Wall 2001). This paper will use the term "hate speech online," throughout.

Although no statistics offer a current global overview of hate speech online, both social networking platforms and organizations that combat hate speech have recognized that the online dissemination of hateful messages is increasing and that greater attention should be paid to this phenomenon, in order for adequate responses to be developed. According to HateBase (2017), a web-based application that collects global instances of hate speech online, most cases target individuals on the basis of ethnicity and nationality; however, incitements to hatred on the basis of religion, class, gender and sexual orientation are increasing.

Evolutionary psychology (Schaller and Park 2011) can contribute explanations ~~as to~~for why insults towards social and sexual minority groups often co-occur with reference to body parts and sexual practices that both derogate the target and express disgust towards him/her. One theory is that disgust has developed from its origin as a disease avoidance mechanism into a putative behavioral immune system comprised of cognitive, affective and behavioral tendencies to avoid sources of disease. Because the biological costs of infection are high, this behavioral immune system makes us hypervigilant and reactive to "false positive" threats. For example, the system may be triggered by people who appear "strange" to the majority because they do not conform ~~with~~ to societal and/or sexual norms (Nussbaum 2010). According to this theory, fear is transformed into hate speech towards those perceived as different. It follows that online communication has the advantage of enabling people to express intolerance towards a disgusted/feared subject from a protected position, with no direct exposure to the target.

In 2001, prompted by the growth of online hate groups and web-based hate speech (Banks 2010; Muižnieks 2017), the Council of Europe promoted the Convention on Cybercrime and, in 2003, adopted the Additional Protocol to regulate hate speech online. This legislative development occurred in parallel with a dramatic increase in microblogging (e.g., via Twitter, Facebook, Tumblr, ~~,~~Google+), through which users shifted from merely consuming media to producing, creating and curating information by ~~creating~~ building personal profiles, writing about their lives, sharing opinions and publicly discussing issues within a bounded system (Meng et al. 2017).

Twitter is the fourth most used social network platform, with 317 million monthly active users, worldwide; these users send more than 500 million status messages (called "Tweets") each day (Twitter 2017). In Italy, where this study was rooted, it is estimated that there are approximately 6.4 million active Twitter users (Twitter 2017). Because ~~users must confine their~~ Tweets must be confined to 280 characters, ~~they~~ users tend to express their reactions to current events much more quickly and dynamically on this platform than on other microblogging sites (i.e., Facebook, ~~,~~ Google+). For this reason, Twitter is an effective platform for real-time sentiment analysis.

Although Twitter forbids users to "publish or post direct, specific threats of violence against others" (Twitter 2017), hate speech towards social groups who are viewed as minorities and/or vulnerable on the basis of their religion, ethnicity, gender or sexual orientation still appears on the site (Awan 2014).

In recent years, there has been a keen interest in identifying and extracting opinions and emotions from text, in order to provide tools for information analysts in government, commercial and political domains seeking to track attitudes and feelings in the news and online forums (Wiebe et al. 2005). However, such work has mostly been limited to posts made ~~in~~ by members of online hate groups and in radical forums at the document or sentence level (Bunrap and William 2015; Djuric et al. 2015; Gitari et al. 2015), and very few studies have examined hate speech against social, ethnic, sexual or gender minority groups on Twitter, specifically (Awan 2014; Chaudhry 2015; Cisneros and Nakayama 2015; Silva et al. 2016).

In 2014, a self-administered online survey of 2,849 Web users (Pew Research Center 2014) reported that the 66% who had experienced online harassment claimed that their most recent incident had occurred on a social networking platform. Women and young adults were more likely than others to have experienced harassment on social media. When asked how upsetting their most recent experience with harassment had been, about half responded "somewhat upsetting" or "extremely upsetting." In November 2014, Twitter enabled the non-profit agency Women, Action, and the Media (WAM!) to collect reports of Twitter-based harassment, assess them and escalate the reports to Twitter, as necessary. Among the 317 genuine harassment reports that were submitted to WAM! between 6 and 24 November, 27% related to hate speech (Matias et al. 2015).

This finding echoes ~~those~~ the conclusions of ~~reported by~~ research conducted in the everyday offline context. The most up-to-date Italian report on intolerance towards social and sexual minority groups (Cox Commission on Intolerance, Xenophobia, Racism and Hate Issues 2016) shows that immigrants are the most hated group, with 65% of Italians considering refugees a burden on society because they enjoy some social and economic benefits. The second and third most hated groups are,

respectively: women, with only 43.7% of Italians recognizing that women are discriminated against in the workplace; and LGBT persons, with 25% of Italians considering homosexuality a disease. In addition, in a 2015 follow-up survey on violence against women in Italy (ISTAT 2015), 31.5% of women aged 16 to 70 (6,788,000 women) were found to have experienced some form of physical or sexual violence during their lives, and 16.1% were found to have experienced psychological violence and stalking.

In 2012, researchers from Humboldt State University launched the "Geography of Hate Map" project. In this project, they tracked and plotted 10 abusive words on an interactive map of racist, homophobic and ableist Tweets posted between June 2012 and April 2013 in the United States. By applying sentiment analysis – which refers to the task of automatically determining feelings from text (Mohammad 2016) – to Tweets on a state level and calculating the ratio of hateful Tweets to the total number of Tweets per state, the researchers revealed the ~~areas of the countr~~states ~~y where~~ in which hateful Tweets were most prominent (Stephens 2013). Such analysis may be particularly useful, as the massive amount of data emanating from Twitter is informative of ~~people's~~ users' valence and emotions towards a particular target or topic (Mohammad 2016; Pang and Lie 2008; Wiebe et al. 2005). At ~~its~~ the foundation of this analysis is Russell's (1980) circumplex model of affect, which characterizes affect according to two primary dimensions: valence (i.e., positive or negative) and arousal (i.e., degree of reactivity to a stimulus). In this vein, the application of sentiment analysis to Twitter is particularly challenging, as the base emotion~~al import~~ of a ~~sentence or utterance~~Tweet is not ~~simply~~ necessarily equivalent to the sum of the emotional associations of each of its component words. Furthermore, valence is not especially straightforward to determine, as emotions are rarely explicitly stated in Tweets and it can be difficult to determine ~~their~~ a Tweet's tone, pitch and emphasis. ~~Utterances~~ Tweets may, in fact, convey ~~more than one~~multiple emotions (to varying degrees) through the contrastive evaluation of multiple target entities. Finally, Tweets are rife with terms that are not found in dictionaries, such as

misspellings, creatively spelled words, hashtagged words, emoticons and abbreviations (Mohammed 2016).

In the current paper, we present the findings of the "Italian Hate Map" project, which aimed at expanding the "Geography of Hate Map" by identifying part-of-hate-speech in Tweets against six targets – women, gay and lesbian persons, immigrants, Jews, Muslims and disabled persons – and aggregating these Tweets according to geographical provenance (Musto et al. 2015). A lexicon-based approach to semantic content analysis was employed to determine the valence of the Tweets (Russell 1980), dealing with the abovementioned challenges in applying sentiment analysis to Twitter. The research question examined was: How might Twitter data extraction and processing enable us to detect and identify hate speech online and develop more effective prevention strategies?

The project drew on three theoretical frameworks: participatory sensing (Aggarwal and Abdelzaher 2013), evolutionary psychology (Schaller and Park 2011) and the minority stress model (Meyer 1995). Together, these enabled us to emphasize the cumulative effects of hate speech online, the psychological advantages for those expressing hate speech and the psychological costs suffered by the targeted social and sexual minorities. As the contribution of evolutionary psychology (Schaller and Park 2011) was outlined above, the remainder of the section will describe participatory sensing and the minority stress model. Participatory sensing (Aggarwal and Abdelzaher 2013) is a mobile crowd sensing approach whereby individuals contribute data on a participatory sensing platform. By sharing information online about their lives, thoughts, sentiments, habits, routines and environments, individuals provide information on larger community behaviors and attitudes towards specific groups or events. The minority stress model (Lingiardi and Nardelli 2014; Meyer 1995) relates to the juxtaposition of minority and dominant values and the resulting conflict with the social environment experienced by minority group members. Minority stress is unique, as it is experienced in addition to the general stressors experienced by all people and is caused by three factors: external objective events and conditions; expectations of such events

and the vigilance that such expectations bring; and the internalization of negative attitudes, feelings and representations ~~of oneself~~. Stigmatized persons may develop adaptive and maladaptive responses to minority stress, which may manifest in mental health symptoms (Meyer 2003).

Materials and Methods

Definition of the lexicon

To establish a corpus of terms associated with the six targets, the terms used by the Humboldt University research team (http://users.humboldt.edu/mstephens/hate/hate_map.html#) to refer to gay and lesbian persons ("dyke," "fag," "homo," "queer"), immigrants ("chink," "gook," "nigger," "wetback," "spick") and disabled persons ("cripper") were expanded. To identify additional terms, the researchers reviewed eight major Italian newspapers' coverage of current events related to the target groups between August 2015 and February 2016. In the same period, an online survey was run on Unipark.de, asking participants to indicate five negative terms they associated with each target group. As different methods of advertising were used (i.e., placing listings on websites and university bulletin boards, snowballing) it was not possible to calculate the precise response rate. However, of the 1,358 people who accessed the link online, 935 completed the survey (69%; M_{age} = 27.48, SD = 6.55). From the three methods of developing the lexicon, 76 derogatory terms were identified.

---------- Table 1 about here ----------

Data collection and analysis

To achieve the project goals, a domain-agnostic framework for the semantic analysis of social streams, called CrowdPulse (Musto et al. 2015), was employed. This framework basically extracts textual content (posts, Tweets, etc.) from social networks such as Twitter, Facebook and Instagram~~,~~ and processes this content to generate interesting insights and to draw out relevant

patternss ~~from the data~~. In our specific setting, we used the framework to extract and identify hate speech, particularly in areas of Italy where more hate speech is typically published. ~~The~~ As the framework is domain -agnostic ~~nature of the framework refers to the fact that the platform~~, it can extract and process all kinds of data, subject to the constraints that: (i) the data is publicly available on a social network and (ii) the data is in textual form (i.e., ~~neither~~ not video ~~n~~or image data ~~can be extracted or processed~~). Formally, the extraction and analysis processes take the shape of a *processing graph*; that is to say, the processes follow a sequence of steps beginning with data extraction, continuing on to the necessary processing algorithms and ending with the storage and visualization of the information. More formally, each processing graph can be figured as a set of nodes connected by edges (see Figure 1). In CrowdPulse, each node is typically referred to as a "plugin" and represents a single processing step. In the present analysis, eacha plugin ~~could be considered~~was a specific software module that performed one of the analytical steps (e.g., data extraction, sentiment analysis, semantics interpretation, etc.); thus, the sequence of nodes that composed the processing graph represented the sequence of algorithms used to ~~correctly~~ process the data and obtain the desired output.

---------- Figure 1 about here ----------

In all CrowdPulse projects, the first analytical step is carried out by an *Extraction* plugin and the final analytical step is carried out by a *Storage* plugin. The ~~first~~ Extraction plugin ~~takes care of~~performs the data ingestion process by drawing on certain defining heuristics (e.g., all Tweets containing specific hashtags or posted from a specific location) and the ~~final~~ Storage ~~one~~ plugin stores all of the (processed) data in a local MongoDB instance (https://www.mongodb.org/). ~~Given~~ However, between these constraints, users can combine particular plugins according to their analytical goals.

The processing graph we used for the ~~"~~Italian Hate Map~~"~~ project is reported in Figure 1. Specifically, the project employed the following plugins:

- *Social Extractor*: The goal of this plugin was to extract textual content from Twitter and Facebook according to specific criteria. The gathered data represented the input that triggered the analysis.
- *Semantic Tagger*: The goal of this plugin was to analyze the content returned by the Social Extractor and to understand the semantics conveyed in each Tweet. The plugin also filtered and removed ambiguous content (e.g., Tweets retrieved by the Social Extractor that were not hate speech) from the outputs.
- *Sentiment Analyzer*: The goal of this plugin was to enrich our comprehension of the content by associating each Tweet with a label indicating the opinion it conveyed (e.g., a positive, negative or neutral opinion). The plugin also filtered out all content conveying a positive or neutral opinion, since our interest was in Tweets spreading a negative message. The remaining Tweets used the abovementioned lexicon with the clear intent of spreading hate speech.
- *Localization*: The goal of this plugin was to increase the amount of geolocalized content. To this end, heuristics were applied and the geographical coordinates of each Tweet were stored along with the content.
- *Storage*: The goal of this plugin was to store and make available the results of the analysis. By querying the information available in the Storage we could access the single Tweets that composed our ~~"~~Italian Hate Map.~~"~~

The next sections provide more detail on the processing that was carried out by each plugin. To better ~~guide the reader in the comprehension of~~ illustrate the overall pipeline, ~~throughout the article~~ we use ~~the~~ three ~~following~~ Tweets as running examples throughout the article. ~~Clearly, a~~All

are written in the Italian language and include the word "midget" (in Italian, *nano*) in the lexicon. The first Tweet (henceforth identified as $\mathbf{t_1}$) discusses the opinion of the (former) Italian Minister Brunetta on recent government measures relating to the economy. The second Tweet (henceforth identified as $\mathbf{t_2}$) is about the iPod Nano (therefore matching a word in the lexicon). The third Tweet (henceforth identified as $\mathbf{t_3}$) refers to the ~~performance of the~~ short statured Italian football player Sebastian Giovinco, ~~also~~ who is more popularly known as the "Atomic Ant."

The precise translation of $\mathbf{t_1}$ is: "*If midget Brunetta said that the stability law sucks, then it is excellent.*" The precise translation of $\mathbf{t_2}$ is: "*Ipod nano orange 8gb arrived!! Thank you Apple for the nice gift! :).*" The precise translation of $\mathbf{t_3}$ is*: "Come on!!!!! The midget!!!! The atomic ant!!! #Giovinco 4-3 #ItalyJapan*."

---------- Figure 2 about here ----------

Social Extractor

The Social Extractor plugin was ~~the~~ an essential component of the pipeline ~~component~~, enabling the framework to connect to the social network and extract all content matching certain criteria. The plugin bridged with Facebook (http://developer.facebook.com) and Twitter (http://dev.twitter.com) by exploiting their official APIs. We chose these data sources because we considered Facebook and Twitter the most popular social networks; thus, we assumed that most online discussions would occur on these platforms. With respect to Twitter, we accessed content by querying the official Streaming APIs; for Facebook, due to privacy reasons, we only exploited public content from specific pages or groups.

Generally speaking, CrowdPulse extracts Tweets and Facebook posts through the application of six heuristics: 1) *Content*, which extracts all material containing a specific term; 2) *User*, which extracts all material posted by a specific user (identified by a specific user name); 3) *Geo*, which extracts all available ~~(~~geolocalized ~~)~~ material (according to a given latitude, longitude

and radius); 4) *Content+Geo*, which extracts all available geolocalized material that matches the terms indicated; 5) *Page*, which extracts all material from a specific page; and 6) *Group*, which extracts all material from a specific group. ~~Clearly, a~~All heuristics are always available for use, but~~.~~ ~~t~~The final selection ~~choice~~ of ~~the most suitable~~ heuristics~~, however,~~ is ~~a design choice~~ made by the programmer ~~depending~~ according to ~~on~~ the goals of the project. In our ~~specific~~ research ~~setting~~, we used heuristics (2) and (4)~~; that is to say,~~: Content and Content+Geo. Specifically, we asked CrowdPulse to extract all Tweets containing one ~~(~~ or more~~)~~ terms in our lexicon and ~~those~~ all Tweets containing terms in our lexicon that were also published by users in Italy.

To begin our data acquisition process, we fed the 76 terms contained in the previously defined lexicon into the Social Extractor plugin ~~with the 76 terms contained in the previously defined lexicon~~. This process generated a preliminary set of items containing potential hate speech that was further analyzed to build the ~~"~~Italian Hate Map.~~"~~ All three of the Tweets presented above were extracted by the Social Extractor module~~,~~ through the application of heuristic (2); that is to say, each of these Tweets was found to contain~~s~~ one of the terms contained in the lexicon ("midget" ~~or~~ [*nano*]).

As ~~we will~~ shown in the "Results" section, a ~~huge~~ large number of items ~~with~~ containing potential hate speech were gathered and stored in this step. However, this ~~extraction~~ step was not sufficient to achieve the goals of the project, since three main issues emerged from a preliminary analysis of the extracted Tweets. First, many of the terms in the lexicon were ambiguous and also used in non-intolerant Tweets. For example, Tweet t_2 ~~is an example of a Tweet affected by this issue~~used the~~. Indeed, the~~ term *nano* ~~has several meanings and can therefore express many intents. In this Tweet, it is used~~ to innocuously describe an Apple product. Thus, several non-hate Tweets needed to be filtered out. Second, many Tweets that matched a lexicon term were not hate speech (e.g., ironic Tweets). For example, Tweet t_3 used~~s~~ the term "midget" to refer to a person of small stature. However, the intent of the Tweet wa~~i~~s not intolerant, since the author wa~~i~~s simply

celebrating a player's goal. Such content needed to be filtered out from the output. Finally, the number of geolocalized Tweets was very low.

In order to address these problems, we introduced three more plugins into our processing graph: a *Semantic Tagger* to filter out ambiguous Tweets; a *Sentiment Analyzer* to determine the sentiment expressed by Tweets in order to filter out neutral and ironic terms and to maintain only those containing hate speech; and a *Geotagger* to increase the number of geolocalized Tweets. The following sections describe the processing carried out by each of these plugins.

Semantic Tagger

Semantic tagging was ~~carried out~~used to identify (and filter out) ambiguous Tweets. A Tweet was considered ambiguous when it contained one ~~(~~or more~~)~~ terms in the lexicon but ~~did not have~~lacked a clear intolerant intent. As described above, Tweet t_2 ~~i~~was an example of a Tweet characterized by this issue. The Semantic Tagger ~~module~~ implemented ~~a pipeline of~~ entity linking algorithms to ~~enable a~~ better ~~understanding of~~identify the meaning and intent of the content extracted by the Social Extractor. Generally speaking, the goal of entity linking is to identify the *entities* mentioned in a piece of text. ~~Clearly~~While~~,~~ a complete discussion of entity linking algorithms is beyond the scope of this paper ~~(;~~ we suggest that readers who are interested in ~~a complete discussion of the~~this topic refer to Derczynski, ~~(~~2015)~~. However~~, ~~for the sake of simplicity, we can state that~~in simple terms, the entity linking process ~~is carried out through~~uses statistical approaches ~~that~~ to map portions of the input text to one or more entities by exploiting large knowledge bases, such as Wikipedia.

In our approach, content was processed through a pipeline of state of the art entity linking algorithms. DBpedia Spotlight (http://dbpedia-spotlight.github.io/demo/), Wikipedia Miner (http://wikipedia-miner.cms.waikato.ac.nz/) and Tag.me (http://tagme.di.unipi.it/) were used to disambiguate the terms used in Tweets. An example of the processing carried out by this module is reported in Figure 3, which shows the output returned by the Tag.me algorithm on Tweet t_2. As

shown in the figure, the Semantic Tagger immediately understood the meaning and intent of the Tweet as ~~clearly~~ non-intolerant. The entity linking algorithm correctly recognized the entities mentioned in the text and detected that the term *nano* ("midget") ~~was~~ had been used to refer to a particular iPod model. Accordingly, ~~once this anomaly was detected,~~ the Tweet was filtered out from the output. This process was repeated for all of the Tweets returned by the Social Extractor. Whenever an ambiguous term was used and the Semantic Tagger detected the absence of intolerant intent, the content was filtered out. Otherwise, Tweets remained in the analysis and passed on to the next module of the processing graph.

---------- Figure 3 about here ----------

To summarize, the Semantic Tagger module was useful for identifying the meaning of terms used in Tweets and filtering out Tweets containing polysemous terms (e.g., the abovementioned Italian term *nano* or the Italian term *finocchio*, which is a translation of both "Nancy" and "queer").

Sentiment Analyzer

The goal of this plugin was to enrich our comprehension of the content by analyzing the *opinion* conveyed in each extracted Tweet. As previously explained, we were interested in maintaining only Tweets with a clear intolerant intent; that is to say, those conveying a clear *negative* opinion. In order to associate a Tweet with a positive, neutral or negative opinion, we employed *sentiment analysis algorithms*. Sentiment analysis aims at labeling textual content (or a part of it) with a sentiment score. This process can be carried out by one of two approaches: *unsupervised sentiment analysis* or *supervised sentiment analysis*. The first technique relies on polarity lexicons ~~that contain a polarity score for (most) terms in a language~~ that label~~.~~ ~~As an example,~~ terms such as "good," "love," "harmony" and "beauty" ~~are labeled with~~as containing a *positive* sentiment ~~score,~~ ~~while~~ and terms such as "bad," "hate," "anger" (or, in general, insults)~~,~~ ~~are~~

~~labeled with~~ as containing a *negative* sentiment ~~score~~. Given these polarity lexicons, unsupervised algorithms ~~would~~ calculate a Tweet's sentiment score ~~of a Tweet~~ as the *sum* of the sentiment scores of each term used in the Tweet, using heuristics to deal with negation~~s~~ and emphasis. As an example, Tweet t_1 would be labeled *neutral*, since it contains two terms with strong but conflicting polarity: "sucks" and "excellent" (here, we are referring to the translation presented in the "Materials and Methods" section); these terms cancel each other out. Similarly, Tweet t_3 would be labeled *neutral*, since no term with significant polarity occurs in the text.

Such an unsupervised approach based on polarity lexicons was implemented in our previous research (Musto et al. 2015), with unsatisfying results. Indeed, Tweet t_1 conveys a *negative* opinion while t_2 is a *positive* Tweet that celebrates a football player's goal; thus, we needed algorithms that could correctly identify sentiment.

Accordingly, in this work we employed more sophisticated sentiment analysis techniques that were able to catch *nuances* of meaning. Specifically, we exploited *supervised approaches*. Such techniques use machine learning to learn a classification model that relies on a set of labeled data and subsequently predicts the label (positive, neutral or negative) of new and unseen Tweets, according to their characteristics. In a nutshell, using this technique, a portion of the available Tweets was manually (or semi-manually) labeled "positive" or "negative" and used to feed the sentiment analysis algorithm. In turn, the algorithm learned very precise nuances of meaning and automatically learned the overall sentiment conveyed by Tweets according to their usage of terms, regardless of whether the terms were positively or negatively polarized. Unfortunately, a complete discussion of sentiment analysis techniques is beyond the scope of this paper. However, we suggest that readers refer to Pang et al. (2008) for a more in-depth discussion.

In our project, we exploited the Sentit algorithm proposed by Basile and Novielli (2015). We chose this algorithm for two reasons: (i) as shown by the related literature, supervised techniques tend to outperform unsupervised techniques; and (ii) Sentit was the best performing algorithm in the recent SENTIPOLC challenge (http://www.di.unito.it/~tutreeb/sentipolc-

evalita14/index.html), whose goal was to correctly perform sentiment analysis on Italian Tweets. As a consequence, we decided to exploit this algorithm in our project. In our research setting, the algorithm was able to correctly classify the polarity of Tweets; thus, t_1 was correctly labeled as "negative" and t_3 was correctly classified as "positive." This was due to the fact that machine learning correctly detected usage of sarcasm and complex expressions such as "Come on!" (as used in t_3) were correctly labeled as expressions conveying a positive message.

As we already explained for the Semantic Tagger, the sentiment analysis process was repeated for all Tweets. Once the sentiment of all available Tweets was calculated, we filtered out all positive and neutral Tweets, on the assumption that they did not convey an intolerant message. Following this step, all remaining Tweets were assumed to carry an intolerant intent and were thus included in our final "Italian Hate Map." Returning to our three examples, only Tweet t_1, which useds the term *nano* with an intolerant intent and conveyeds a negative opinion, was labeled as "hate speech" and included in the "Italian Hate Map." Tweet t_2 was excluded by the Semantic Tagger, which detected its non-intolerant intent and Tweet t_3 was excluded by Sentiment Analysis, which classified it as a positive Tweet.

Localization

Finally, in order to obtain the final distribution of ~~the~~ hate speech, all content that had previously been classified as intolerant was geographically aggregated and normalized to reflect Twitter use according to area. Twitter APIs can be used to tag Tweets with latitude and longitude. However, only a very small number of the collected Tweets (~~around~~ approximately 0.5%) had an explicit localization; thus, the goal of the Localization plugin was to increase the amount of geolocalized content. Specifically, the plugin queried the GeoNames API (GeoNames 2017) to map the location attribute of a user's profile and tagged that user's intolerant content with the coordinates of his/her location. Following this, all content posted by that user automatically

inherited those coordinates, on the assumption that (most of) the content posted by that user would have come from the location indicated in his/her profile.

Storage

The Storage plugin was the final plugin used. The goal of this plugin was to store processed content in a local MongoDB instance in order to enable an analytics console to access the results of the analysis in a user-friendly interface. We stored all Tweets, along with their semantically annotated content, their conveyed sentiments, their binary classifications (intolerant/not intolerant) and their associated locations (where available). In order to make the research fully reproducible, we also made available all the negative Tweets exploited in this work (data can be accessed at https://data.mendeley.com/datasets/5ky5fj7nnj/1). In the ~~f~~next~~o~~llowing section, we ~~will~~ ~~discuss~~present the ~~outcomes~~results of~~emerging from~~ the data analysis.

Results

As reported in Table 2, over a period of 7 months we extracted 2,659,879 Tweets from 879,428 Twitter profiles; 412,716 of these Tweets contained the negative search terms. In the geolocalized Tweets, women were the most insulted group, having received 71,006 hateful Tweets (60.4% of the negative geolocalized Tweets), followed by immigrants (12,281 tweets, 10.4%), gay and lesbian persons (12,140 tweets, 10.3%), Muslims (7,465 tweets, 6.4%), Jews (7,465 tweets, 6.4%) and disabled persons (7,230 tweets, 6.1%).

The distribution of hateful Tweets is shown in Figure 4. It is worth noting that the values reported in the map (with red areas indicating the origins of the greatest amount of hate speech-) do not represent a simple Tweet "count." Rather, they represent the ratio of Tweets containing hate speech to the total number of Tweets originating in the particular area. This weighting strategy was employed to correct for any natural increase in Tweets containing hate speech in highly populated areas. The f~~F~~indings indicate that personal sentiments expressed on Twitter may have been

triggered by social events that occurred in the days prior to the Tweets (see Table 3). Furthermore, the target terms were often combined with other terms, such as "shit," "cock," "dick" and other references to body parts, in order to reinforce the insult.

---------- Tables 2 and 3 about here ----------

---------- Figure 4 about here ----------

Discussion

This study applied lexicon-based semantic content analysis to the huge body of textual data on Twitter – a platform that provides a real-time snapshot of community behaviors and attitudes towards women, gay and lesbian persons, immigrants, Muslims, Jews and disabled persons. Critics might argue that hate speech on Twitter does not represent all hate speech within society. However, consistent with the Italian report on intolerance towards social and sexual minority groups in the offline context (Cox Commission on Intolerance, Xenophobia, Racism and Hate Issues 2016), our results show that immigrants, women, and gay and lesbian persons are the most frequent targets of hate speech online.

In light of these results, we would like to encourage three considerations. First, it may be speculated that increases in intolerant tweets towards a specific minority group may parallel daily events in the wider social context. For example, debates over immigration politics or same-sex marriage may stimulate negative tweets towards immigrants and gay and lesbian people, respectively, from people who are less favorable to liberalization in these policy areas. Future research should seek to verify whether peaks of intolerant tweets towards a particular target group tend to co-occur with related socio-political events.

Second, it should be borne in mind that the detection of online hate speech may not directly lead to counter-actions, because so few people report online abuse (UNESCO 2015). In part, this is

because many people are not fully aware when an online offense has been committed. Furthermore, even when such cases are reported to the police, the police have limited resources with which to pursue action. Also, in many cases, tracking the crime can present many problems from both a jurisdictional point of view (with respect to Internet service providers) and an ethical point of view (relating to, e.g., the role of free speech and the issue of online anonymity). This leads to the third consideration, which is that an effective strategy for tackling hate speech online must sensitize Internet users to the nature of online communication, enabling them to discriminate between content that is threatening and offensive and content that is not.

~~Furthermore, Twitter provides information that can further our understanding of how different forms of communication within society allow users to express intolerant sentiments. In keeping with the participatory sensing framework (Aggarwal and Abdelzaher 2013), the lexicon-based method employed in the present study limited the social desirability bias that can occur in research with other instruments (e.g., surveys, interviews), which are often costly and time-consuming. Another strength of the approach is its identification of the most commonly used intolerant terms and, more importantly, the context in which they are used.~~

Mindful of the risk of mental health symptoms in stigmatized group members (Fisher et al. 2017; Meyer 1995, 2003), the ~~"~~Italian Hate Map~~"~~ project had three progressive educational and preventative goals: to raise awareness of hate speech online by conveying and disseminating information about its consequences; to identify areas in which intolerance is more widespread; and to use geolocalized Tweets to develop prevention strategies tailored towards specific criticalities and strengths. As the project tracked the exact geographical position of Tweets, its findings may facilitate intolerance prevention on two levels: on a local level, the geolocalized Tweets reveal the social and sexual minority groups who are the most frequent victims of hate speech in specific areas; and on a national level, the Tweets detect the way in which ~~people's~~ sentiment towards these minority groups changes over time and physical distance, in relation to specific social events (e.g., immigrant landings, approval of same-sex marriage). Moreover, the finding that intolerant terms are

based on the target group's typical characteristics (e.g., "kebab" for Muslims) and often presented with other terms, such as "shit," "cock," and "dick," may be useful for the development of educational programs by informing the best linguistic strategies to deconstruct stereotypes regarding social, cultural and gender differences, both between and within groups.

However, caution is warranted when interpreting these findings, due to the methodological limitations of the approach. The research technique was based on a simple matching of terms in our lexicon with content posted on Twitter. Semantic analysis enabled us to filter out non-intolerant Tweets, but we were unable to intercept hateful content that did not contain terms in our lexicon. A methodological improvement would involve the use of our lexicon to extract seed Tweets and the use of human annotators to label these Tweets as intolerant or not intolerant. This would require a huge effort, but it would "teach" the algorithms to automatically understand the nature of the Tweets and ensure more precise outcomes, including larger vocabularies of intolerant terms and idiomatic and dialectical expressions. To this aim, building a hate speech detection system that leverages our findings is part of our future research agenda.

Notwithstanding these limitations, several strengths should be acknowledged. The analysis of Tweets provided information that could further our understanding of the way in which different forms of communication within society allow users to express intolerant sentiments. In keeping with the participatory sensing framework (Aggarwal and Abdelzaher 2013), the lexicon-based method employed in the present study limited the social desirability bias that can occur in research with other instruments (e.g., surveys, interviews), which are often costly and time-consuming. Another strength of the approach is its identification of the most commonly used intolerant terms and, more importantly, the context in which they are used.

In light of this, ~~T~~the "Italian Hate Map" project is of great importance, as it is increasingly assumed that the cyberspace reflects patterns and practices that are enacted in offline social interactions (Graham 1998). In addition, given the dramatic diffusion of hate speech online

(UNESCO 2015), the project contributes to a greater understanding of its significance and consequences and the development of effective and tailored responses.

Acknowledgments

The authors would like to thank xxx and xxx (masked for review) for their collaboration in the research project.

Declaration of Interest Statement

The authors have no conflicts of interest in relation to this work.

References

Aggarwal CC, Abdelzaher T. 2013. Social sensing. In: Aggarwal CC, editor. Managing and mining sensor data, New York (NY): Springer; pp. 237–97.

Awan I. 2014. Islamophobia and Twitter: a typology of online hate against Muslims on social media. Policy Internet. 6:133–150.

Banks J. 2010. Regulating hate speech online. Int Rev Law Comp Technol. 24:233–239.

Basile P, Novielli N. UNIBA: sentiment analysis of English Tweets combining micro-blogging, lexicon and semantic features. Proceedings of the 9th International Workshop on Semantic Evaluation; June 4–5; Denver, Colorado.

Burnap P, Williams ML. 2015. Cyber hate speech on Twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet. 7:223–242.

Chaudhry I. 2015. # Hashtagging hate: using Twitter to track racism online. First Monday. 20.

Cisneros JD, Nakayama TK. 2015. New media, old racisms: Twitter, Miss America, and cultural logics of race. J Int Intercult Comm. 8:108–127.

Cox Commission on intolerance, xenophobia, racism and hate issues. 2016. Final report. [accessed 2018 Jan]. http://www.camera.it/leg17/1313.

Derczynski L, Maynard D, Rizzo G, Van Erp M, Gorrell G, Troncy R, Petrak J, Bontcheva K. 2015. Analysis of named entity recognition and linking for tweets. Inf Process Manage. 51:32 –49.

Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, Bhamidipati N, editors. 2015. Hate speech detection with comment embeddings. Proceedings of the 24th International Conference on World Wide Web; May 18–22; Florence, Italy.

Fisher AD, Castellini G, Ristori J, Casale H, Giovanardi G, Carone N, ... Ricca V. 2017. Who has the worst attitudes toward sexual minorities? Comparison of transphobia and homophobia levels in gender dysphoric individuals, the general population and health care providers. J Endocrinol Invest, 40:263–273.

GeoNames 2017. GeoNames API. [accessed 2018 Jan]. http://www.geonames.org/export/web-services.html.

Gitani ND, Zuping Z, Damien H, Long J. 2015. A lexicon-based approach for hate speech detection. Int J Multimedia Ubiquitous Engineering. 10:215–230.

Graham S. 1998. Spaces of surveillant simulation: new technologies, digital representations, and material geographies. Environ Plann D. 16:483–504.

Hatebase 2017. Most common hate speech. [accessed 2018 Jan]. http://www.hatebase.org/popular.

ISTAT. 2015. Violence against women in Italy. [accessed 2018 Jan]. https://www.istat.it/en/files/2015/09/EN_Violence_women.pdf?title=Violence+against+women+-+23+Sep+2015+-+Full+text.pdf.

Lingiardi V, Nardelli N. 2014. Negative attitudes to lesbians and gay men: persecutors and victims. In: Corona G, Jannini EA, Maggi M, editors. Emotional, physical and sexual abuse, Cham (ZG): Springer; pp. 33–47.

Matias JN, Johnson A, Boesel WE, Keegan B, Friedman J, DeTar C. 2015. Reporting, reviewing, and responding to harassment on Twitter. Women, action, and the media. [accessed 2018 Jan]. http://womenactionmedia.org/twitter-report.

Meng J, Martinez L, Holmstrom A, Chung M, Cox, J. 2017. Research on social networking sites and social support from 2004 to 2015: a narrative review and directions for future research. Cyberpsych Beh Soc N. 20:44–51.

Meyer IH. 1995. Minority stress and mental health in gay men. J Health Soc Behav. 36:38–56.

Meyer IH. 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychol Bull. 129:674–697.

Mohammad SM. 2016. Sentiment analysis: detecting valence, emotions, and other affectual states from text. Emot Meas. 201–237.

Muižnieks N. Hate speech is not protected speech. ENARgy The European Network Against Racism's webzine 2013. [accessed 2018 Jan]. http://www. enargywebzine. eu/spip. php? article332.

Musto C, Semeraro G, Lops P, de Gemmis M. 2015. CrowdPulse: a framework for real-time semantic analysis of social streams. Inform Systems. 54:127–46.

Nussbaum MC. 2010. From disgust to humanity: sexual orientation and constitutional law. New York (NY): Oxford University Press.

Pang B, Lee L. (2008). Opinion mining and sentiment analysis. Found Trends Inf Retr. 2:1–94.

Pew Research Center. 2014. Online harassment. [accessed 2018 Jan]. http://www.pewinternet.org/2014/10/22/online-harassment/.

Russell J. 1980. A circumplex model of affect. J Pers Soc Psychol. 39:1161–1178.

Schaller M, Park JH. 2011. The behavioral immune system (and why it matters). Curr Dir Psychol Sci. 20:99–103.

Silva L, Mondal M, Correa D, Benevenuto F, Weber I. 2016. Analyzing the targets of hate in online social media. International AAAI Conference on Web and Social Media 2016. [accessed 2018 Jan]. https://arxiv.org/pdf/1603.07709.pdf.

Stephens M (2013). The geography of hate map. [accessed 2018 Jan]. http://users.humboldt.edu/mstephens/hate/hate_map.html#.

Twitter 2017. Twitter. [accessed 2018 Jan]. https://support.twitter.com/articles/18311.

UNESCO 2015. Countering online hate speech. [accessed 2018 Jan]. http://unesdoc.unesco.org/images/0023/002332/233231e.pdf.

Walker S. 1994. Hate speech: the history of an American controversy. Lincoln, NE: University of Nebraska Press.

Wall D. 2001. Crime and the Internet. London (LDN): Routledge.

Warner W, Hirschberg J, editors. 2012. Detecting hate speech on the world wide web. Proceedings of the Second Workshop on Language in Social Media; June 7; Montreal, Canada.

Wiebe, J., Wilson, T., Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Lang Resour Eval. 39:165–210.

Table 1 *Examples of terms used to detect Tweets with negative content*

Disabled	GL persons	Immigrants	Jews	Muslims	Women
Cripple	Bean flicker	Blue collar	Bagel-Dog	Bomber	Cocksucker
Freak	Dyke	Gypsy	Crikey	Cave Nigger	Slag
Fucktard	Fag	Gook	Gargamel	Kebab	Slut
Mongo	Nancy	Nigger	Kike	Landya	Trollop
Spaz	Queer	Paki	Yid	Towel-Head	Whore

Note. For each term, alternate spellings or misspellings were also considered.

Table 2 *Total number of Tweets extracted about target groups*

Target group	Total Tweets	Negative Tweets	Negative geolocalized Tweets	Total Twitter profiles
Women	1,007,540 (37.2%)	284,634 (69%)	71,006 (60.4%)	167,796 (19.1%)
Immigrants	105,727 (4%)	38,100 (9.2%)	12,281 (10.4%)	53,235 (6.1%)
GL	67,950 (2.6%)	35,207 (8.5%)	12,140 (10.3%)	30,027 (3.4%)
Muslims	1,014,693 (38.1%)	22,435 (5.5%)	7,465 (6.4%)	391,258 (44.5%)
Jews	86,102 3.2%)	6,754 (1.6%)	7,465 (6.4%)	35,602 (4%)
Disabled	377,867 (14.2%)	25,586 (6.2%)	7,230 (6.1%)	201,510 (22.9%)
Total	2,659,879	412,716	117,587	879,428

Not.: Target groups were sorted on the basis of the total number of negative geolocalized Tweets.

Table 3 *Examples of Tweets about the six target groups*

Target group	Negative Tweets
Women	[Showgil's name] the best cocksucker in the showbusiness! #cockbusiness
Gay and lesbian persons	#footballmatch #[footballer's name] #kickordance I see a nancy dancing in the football field. Kick that fuckin' ball, faggot!
Immigrants	#gipsycl(e)an #caravans You're not much different from natives when it comes to drinking ... Except your clean
Muslims	#Allah #bomber #cleansing [City's name] is crowded with stinky Camel-Fucking Cave Nigger
Jews	[Showman's name] Beautiful and poor like a Gargamel
Disabled persons	#morespacelessspaz Mongo bongo

Figure 1 *The Italian Hate Map: Processing Graph*

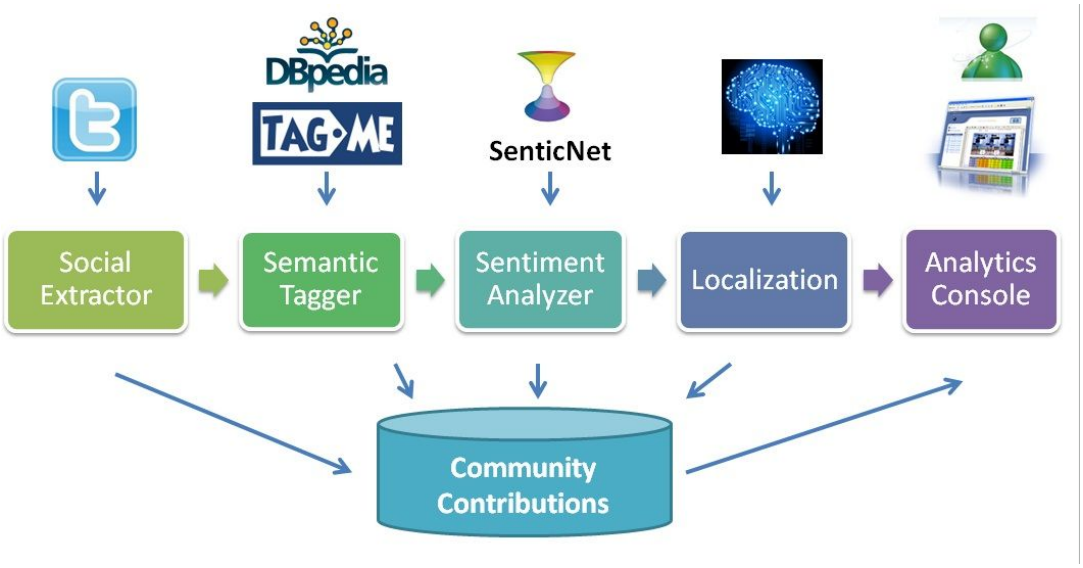

Figure 2 Three different Tweets (in Italian language) which may convey hate speech. They all match the term 'midget' (in Italian 'nano') which is in the lexicon

Figure 3 Output returned by an entity linking algorithm for a non-intolerant Tweet. The ambiguous usage of the terms in the lexicon and its non-intolerant intent immediately emerges

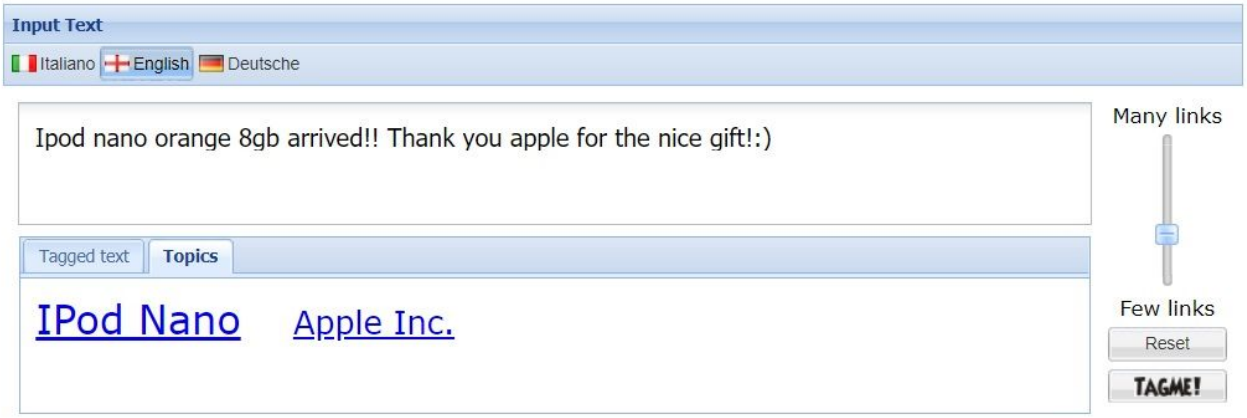

Figure 4 *Geographic Distribution of the Total Number of Intolerant Tweets about Jews, Disabled Persons, Muslims, GL Persons, Women, and Immigrants, respectively*