UNIVERSITÁ DI PISA
DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE

SOCIAL MEDIA
FOR THE SUPPORT AND IMPROVEMENT OF CITIZENS' WELL-BEING

DOCTORAL THESIS

Author
Fabio Del Vigna

Tutor (s)
Prof. Marco Avvenuti
Dr. Maurizio Tesconi

Reviewer (s)
Prof. Roberto Di Pietro
Dr. Paolo Deluca

The Coordinator of the PhD Program
Prof. Marco Luise

Pisa, May 2018

Cycle XXX

This thesis is dedicated to my parents
who gave me the energy to pursue my research.

"If you can dream it, you can do it."
Walt Disney

Acknowledgements

For this thesis work, I must first give special thanks to the Social Sensing project, Cassandra project and CRAIM laboratory, through which it was possible to deepen many of the issues discussed in this thesis work.

I also feel the need to thank all those who have been close to me during this extraordinary training period and have supported me in the important choices, starting with my family, and Serena in particular for her patience.

Finally, I thank those who supported me in the process of carrying out and revising this scientific work, my colleagues for the fruitful discussions, and the tutors for the significant support provided

Ringraziamenti

PER questo lavoro di tesi devo innanzitutto fare un sentito ringraziamento ai progetti Social Sensing, Cassandra e al laboratorio CRAIM attraverso i quali è stato possibile approfondire molte delle tematiche discusse in questo lavoro di tesi.

Sento inoltre il bisogno di ringraziare tutti coloro che in questo importante periodo di formazione mi sono stati vicini e mi hanno supportato nelle scelte importanti, a cominciare dalla mia famiglia, e Serena in particolare per la pazienza avuta.

Ringrazio infine coloro che mi hanno aiutato nel processo di realizzazione e revisione di questo lavoro scientifico, i colleghi per le proficue discussioni, e i tutori per l'importante supporto fornito.

Summary

SOCIAL Media (SM) platforms are a widespread phenomenon connected to the diffusion of the Internet. This work will exploit several SM platforms belonging to different categories to prove how collective intelligence, SM mining and data analysis can be used to improve and support citizens safety and health. Consequently, this thesis proposes a set of multidisciplinary, cost-effective approaches to investigate and sustain public health and well-being through SM.

For example, SM platforms are becoming increasingly crucial in providing uncensored information on drug sourcing mechanisms or, in case of emergency, revealing their usefulness to support population and to get more situational awareness.

In fact, in recent years, a significant increment of SM activity has been observed in the aftermath of mass convergence and emergency events. To this regard, microblogs such as Twitter, Weibo, and Instagram are favored channels of information diffusion because of their ubiquity and simplicity. During emergencies, people usually report their experience on these media, which are consequently overwhelmed by information concerning the unfolding scenario.

This SM feature is explored in this work through a system able to process incoming data to identify useful information for these purposes, analyzing the data using a two-fold perspective. On the one hand, we explore damage detection techniques to detect messages reporting damage to infrastructures or injuries to the population. On the other hand, we propose a message geolocation component that performs the geoparsing task by exploiting online semantic annotators and collaborative knowledge-bases.

Furthermore, this work measures the relevance of drugs diffusion and advertisement, as well as user engagement using SM and proposes a semi-supervised approach to support health departments in identifying novel substances timely.

Unfortunately, SM platforms are also the ideal plaza for the proliferation of other harmful information. Cyberbullying, sexual predation, self-harm practices incitement are some of the effective results of the dissemination of malicious information on SM. The hate can be directed towards wide groups of individuals, discriminated for some features, like race or gender. To study and monitor the phenomenon of hate in SM we propose a methodology to prevent the critical social consequences of massive on-

line hate campaigns and will compare the approach with the results of other academic works.

Finally, concerning these kinds of attacks and news spreading, we will discuss the problem of censored identities and propose a methodology to spot them.

Sommario

I Social Media (SM) sono piattaforme di ampio successo, collegate alla diffusione di Internet. Questo lavoro sfrutta diverse tipologie di SM per mostrare come l'intelligenza collettiva, la trasformazione e raffinazione dei dati in informazioni e la loro analisi possano essere utilizzati per migliorare la sicurezza, la salute e le condizioni di vita dei cittadini. Di conseguenza, questa tesi fa uso di approcci multidisciplinari per investigare i contenuti delle piattaforme di SM e ricavarne un beneficio per la società.

Ad esempio, i SM stanno diventando sempre più cruciali per l'investigazione, il tracciamento e la comprensione del ciclo di vita delle sostanze stupefacenti oppure, in caso di calamità naturale, nella gestione delle comunicazioni e dei flussi di informazione per supportare la popolazione e per migliorare ed aumentare la conoscenza del contesto venutosi a creare.

Infatti, negli ultimi anni, si è osservato come a seguito di disastri naturali o incidenti causati dall'uomo, l'attività degli utenti sui SM aumenti significativamente, specie per quanto concerne la condivisione di informazioni preziose circa l'emergenza.

A questo proposito, i microblog come Twitter, Weibo e Instagram possono essere considerati canali di diffusione delle informazioni privilegiati, a causa della loro ampia diffusione e semplicità. Durante le emergenze, le persone di solito riportano e condividono la loro esperienza attraverso questi mezzi, che sono pertanto inondati da informazioni relative allo scenario che si va via via delineando.

Questo lavoro esplora quindi questa peculiarità dei SM proponendo un sistema in grado di elaborare i messaggi degli utenti per identificare informazioni utili alla gestione delle emergenze, analizzandoli con una duplice prospettiva. Da un lato vengono testate tecniche di rilevamento automatico dei danni basate su Machine Learning per identificare i messaggi che denunciano la presenza di danneggiamenti alle infrastrutture o lesioni fisiche a persone; dall'altro, viene proposto un componente di geolocalizzazione dei messaggi che esegue l'attività di geoparsing mediante gli annotatori semantici, strumenti in grado di collegare le sequenze significative dei testi a basi di conoscenza collaborative come Wikipedia.

Inoltre, questo lavoro propone alcuni strumenti di rilevazione della diffusione e studio delle nuove droghe, attraverso un approccio semi-supervisionato, per supportare i

servizi di assistenza sanitaria e per identificare tempestivamente le nuove sostanze.

Purtroppo, le piattaforme SM sono anche il punto di aggregazione dove avvengono numerosi comportamenti al limite della legalità. Il cyberbullismo, ricatti a sfondo sessuale, incitamento a pratiche di auto-lesionismo sono soltanto alcuni degli effetti negativi che possono essere causati dalla diffusione di determinati contenuti sui SM. L'odio e la violenza verbale possono essere rivolti ad ampi gruppi di individui, discriminati per alcune loro caratteristiche, come la razza o il genere. Per studiare e monitorare il fenomeno dell'odio sui SM, si propone una metodologia volta a prevenire le conseguenze critiche delle campagne di odio condotte online dagli haters, e si confronta l'approccio qui proposto con i risultati di lavori simili in letteratura.

Infine, riguardo queste tipologie di attacchi, viene affrontato il problema delle identità censurate (anche connesso con il problema della libertà di stampa e della diffusione libera delle notizie sui SM), proponendo una metodologia per individuare le identità reali nascoste dai nomi in codice o nei messaggi alterati.

List of publications

International Journals

1. Avvenuti, M., Cresci, S., Del Vigna, F., & Tesconi, M. (2016). Impromptu crisis mapping to prioritize emergency response. Computer, 49(5), 28-37.

2. Avvenuti, M., Cresci, S., Del Vigna, F., & Tesconi, M. (2017). On the need of opening up crowdsourced emergency management systems. AI & SOCIETY, 1-6.

3. Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., & Tesconi, M. (2018). CrisMap: a Big Data Crisis Mapping System Based on Damage Detection and Geoparsing. Information Systems Frontiers, 1-19.

4. Del Vigna, F., Petrocchi, M., Tommasi, A., Zavattari, C., & Tesconi, M. (2018). Who framed Roger Reindeer? De-censorship of Facebook posts by snippet classification. Online Social Networks and Media, Elsevier (Accepted).

International Conferences/Workshops with Peer Review

1. Del Vigna, F., & Cresci, S. (2015). Social Media for the Common Good: the case of EARS.

2. Avvenuti, M., Del Vigna, F., Cresci, S., Marchetti, A., & Tesconi, M. (2015, November). Pulling information from social media in the aftermath of unpredictable disasters. In Information and Communication Technologies for Disaster Management (ICT-DM), 2015 2nd International Conference on (pp. 258-264). IEEE.

3. Del Vigna, F., Avvenuti, M., Bacciu, C., Deluca, P., Petrocchi, M., Marchetti, A., & Tesconi, M. (2016, October). Spotting the diffusion of New Psychoactive Substances over the Internet. In International symposium on intelligent data analysis (pp. 86-97). Springer International Publishing.

4. Del Vigna, F., Petrocchi, M., Tommasi, A., Zavattari, C., & Tesconi, M. (2016, November). Semi-supervised knowledge extraction for detection of drugs and

their effects. In International conference on social informatics (pp. 494-509). Springer International Publishing.

5. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. Proceedings of the First Italian Conference on Cybersecurity (pp. 86-95).

Others

1. Bellomo, S., Cresci, S., Del Vigna, F., La Polla, M. N., & Tesconi, M. (2015). A platform for gathering eyewitness reports from social media users in the aftermath of emergencies. Technical Report IIT-CNR

2. Bacciu, C., Del Vigna, F., Marchetti, A., Tesconi, M. & Deluca, P. (2016). Towards an Automated Analysis of the Online Supply Chain of Novel Psychoactive Substances. In OCommWEBIST 2016 abstracts book (poster session).

3. Cresci, S., La Polla, M. N., Mazza, M., Tesconi, M. & Del Vigna, F. (2016). #selfie: mapping the phenomenon. IIT TR-08/2016

4. Del Vigna, F., Petrocchi, M. & Tesconi, M. (2016). Technical report on the methodology used for the analysis of websites. IIT TR-10/2016

5. Del Vigna, F., Petrocchi, M., Deluca, P. & Tesconi, M. (2016). Main social media analysis outcome of the CASSANDRA project Main social media analysis outcome of the CASSANDRA project. IIT TR-12/2016

6. Cresci, S., Del Vigna, F. e Tesconi, M. (2017) I Big Data nella ricerca politica e sociale, in Andretta, M., and Bracciale, R., (eds.), Social Media Campaigning: Le elezioni regionali in #Toscana2015, Pisa: Pisa University Press, 113-140

Contents

CHAPTER 1

Introduction

Social Media (SM) platforms are a broad phenomenon connected to the diffusion of the Internet, and in particular of the "weblogs" or "blogs" which evolved in those platforms that will become popular as Social Networks (SN) [101]. Websites and applications like MySpace[1] and Facebook[2] made this category of content providers successful and soon we observed the rise of different services, all sharing a common aspect: the content is produced by users.

According to Kaplan definition of SM [101], there exist different kinds of SM that exhibit different features: we all agree that Wikipedia[3], YouTube[4], Facebook, and Second Life[5] are all platforms in which users act as content producers but they can be distinguished through their level of social presence (or media richness) and social processes (like self-presentation and self-disclosure) [105]. If we try to map some common SM platforms according to this categorization, we end up with an interesting classification that generates a matrix of typologies of SM (see Figure 1.1): blogs, social networking sites, virtual social worlds, collaborative projects, content communities and virtual game worlds. SNs and content communities are particularly rich in textual and multimedia data and represent a dynamic and fast source of information with a deep penetration in society, which enables a broad understanding of cultural and economical aspects of communities, if properly analyzed. This rapid information sharing and propagation are connected to the "Small World" phenomenon [183] and the high density of links. SM platforms allow users to easily establish links and interact with other users to share contents, thus reducing to few clicks distances that in the past were con-

[1] https://myspace.com
[2] https://www.facebook.com
[3] https://www.wikipedia.org
[4] https://www.youtube.com
[5] http://secondlife.com

		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/ Self-disclosure	High	Blogs	Social networking sites (e.g., Facebook)	Virtual social worlds (e.g., Second Life)
	Low	Collaborative projects (e.g., Wikipedia)	Content communities (e.g., YouTube)	Virtual game worlds (e.g., World of Warcraft)

Figure 1.1: *Classification of SM by social presence/media richness and self-presentation/self-disclosure.*

sidered important barriers in communication. Moreover, SM users are likely to share information about what they perceive or think [165, 169], thus producing a distributed collective stream of consciousness and feelings. This information sharing encourages and supports a rapid spreading of ideas and news and favors interaction either in the sense of human interactions or as economic transactions. This features contributed to new developments in global public health surveillance (infoveillance[6]), open source intelligence, and dissemination of information and alerts in real-time.

Unfortunately, SM, as well as other online communities, might also become the ideal environments to create, share and exchange information on illicit goods such as child pornography, weapon transactions, drugs, false documents and stolen identities, etc. Also, they are crucial in providing uncensored information on drug sourcing mechanisms and advice around optimal use, hosting discussions around popular choices, experiences and harm reduction practices, as well as promoting Novel Psychoactive Substances (NPS). Fortunately, SM platforms represent also a way to improve our understanding of how emerging trends in recreational drug use occur, aside from the early detection of the availability of novel psychoactive substances for sale online and their potential effects. In fact, if properly analyzed, SM provide a better understanding of how the NPS supply chain works and how their diffusion among users occurs and the role that the Internet has.

Besides SM find a broader field of application in real-time context, in which timing can be significant for public health. For example, in case of emergency, SM revealed their usefulness to support the population and to get more situational awareness. Critical events usually unchain people involvement in participation and reporting [179]. In fact, in recent years, a significant increment of SM activity has been observed in the aftermath of mass convergence and emergency events [13]. To this regard, microblogs such as Twitter[7], Weibo[8], and Instagram[9] are favoured channels of information diffusion because of their ubiquity and simplicity [95]. During emergencies, people usually report their experience on these media, which are consequently overwhelmed by information concerning the unfolding scenario. Also, messages shared on these media are often complemented by comments, images, or videos [150]. Such astonishing amounts of data, when collected and aggregated, converge around meaningful information that can be used for decision making [7]. This collective participation in reporting timely crucial information is the consequence of people attitude to evaluate the situation and

[6] https://en.wikipedia.org/wiki/Infoveillance
[7] https://twitter.com
[8] https://weibo.com
[9] https://www.instagram.com

act accordingly to achieve the overall goal for the common good. This behavior represents a form of *collective intelligence* through which people self-organize and adapt to deal better with the situation [118].

In this work, we will exploit messages and metadata collected from SNs and content communities to prove how collective intelligence, SM mining and data analysis can be used to improve and support citizens safety and security. Consequently, we propose some multidisciplinary, cost-effective approaches to investigate and support public health and well-being through SM.

Whilst the problem of NPS poses serious problems to societies as introduced in Section 1.1, SM can provide a valid support in contrasting the phenomenon. In Chapter 2 we will dissect how some of the aforementioned kinds of SM (i.e., blogs, SNs, and forums) contribute to the drug diffusion but we will discuss also how society can benefit from monitoring such sources of information to contrast the effects of NPS. In particular, we will propose a broad analysis of two large communities of drug consumers and a novel methodology to timely detect new substances, in order to stop early their diffusion. This thesis proves also the effectiveness of SM in supporting the population during mass emergencies through the proposition of a useful application to detect and map crises in real-time using SM messages (see Section 1.2 for additional references). We describe the crisis management system in Chapter 3, with particular regard to the damage detection and mapping components. Furthermore, this study will care also about certain harmful behaviors that may occur in SM like hate speech and censorship. While the first one is undoubtedly dangerous to social relationships, the latter is crucial in the news spreading process and can be used by governments to limit the information diffusion.

Remarkably, the analyses proposed in this thesis are mainly based on SM data. This often implies to deal with sensitive personal data that falls under the privacy regulations. To preserve at best the security of personal data, we tried to limit as much as possible the quantity of information that is stored in our databases. All best practices have been adopted in order to prevent unauthorized accesses to the system, with particular regards to data. All software was run in machines not exposed to the Internet and whose access was protected by a password, and the access to databases was protected by a password. Moreover, we avoided maintaining sensible information when possible and kept only aggregated information, or use them to train Machine Learning models before removing it.

Indeed, on 25th May 2018 the new EU regulation on Data Protection[10] will take effect and will enforce a deeper and accurate data treatment, requiring system designers to provide Privacy-By-Design. All software run in this manuscript is not based on sensitive information or can be executed without it. This makes the adaptation of the current work easy to adapt and work with the EU regulatory.

1.1 The Novel Psychoactive Substances and Social Media

NPS, which represent for US and European countries a raising emergency, lay in a grey area of legislation and are spreading rapidly through the Internet. The risks connected to this phenomenon are high: every year, hundreds of consumers get overdoses of

[10] https://www.eugdpr.org/eugdpr.org.html

these chemical substances and hospitals have difficulties in providing effective countermeasures, given the unknown nature of NPS. Government and health departments are struggling to monitor the market to tackle NPS diffusion, forbid NPS trade and sensitize people to the harmful effects of these drugs[11]. Unfortunately, legislation is typically some steps back and newer NPS quickly replace the old generation of substances. Also, the abuse of certain prescription drugs, like opioids, central nervous system depressants, and stimulants, is a widespread and alarming trend, which can lead to a variety of adverse health effects, including addiction[12].

The described phenomena are being exacerbated by the fact that online shops and marketplaces convey NPS through the Internet [158]. Moreover, SM, with particular regard to specialized forums, offer a fertile stage for questionable organizations to promote NPS, as a replacement of well-known drugs, whose effects have been known for years and whose trading is strictly forbidden. Furthermore, forums are contact points for people willing to experiment with new substances or looking for alternatives to some chemicals, but also a discussion arena for those having first experiences with drugs, as well as trying to stop substance misuse or looking for advice regarding doses, assumption, and preparation.

Online shops and marketplaces convey NPS through the Internet [158], without any (or with very few) legal consequences. Quite obviously, this attracts drug consumers, who can legally buy these drugs without risk of prosecution. Furthermore, products sold over the Internet with the same name may contain different substances, as well as possible changes in drug composition over time [55]. In addition, it is possible to observe the opposite phenomenon: the same substance can be sold across different stores with different names.

EU have supported several projects over the past 10 years that have investigated the role of the Internet in shaping drug use. These include Psychonaut 2002, Psychonaut EWS, ReDNet, CODEMISUSED, ALICE RAP and Cassandra[13]. This one, in particular, has been active and effective in investigating the NPS supply chain, lifecycle, and endorsement, through the analysis of popular SM, drug forums and online shops. Such analysis is vital to timely detect NPS diffusion: this will support governments and health agencies in confining the progress of substance abuse, prohibiting NPS sales and improving the awareness of citizens towards unhealthy and harmful behaviours.

Given the relevance of drug dealing on public health, in Chapter 2 we will present a methodology used to collect data related to NPS from different SM and show how to inspect data to extract actionable information. In particular, SM data reveals to be useful in spotting the diffusion of novel substances and monitoring the diffusion and life cycle of drugs among drug communities. To shorten the gap between drug market and law enforcers/health departments, in Section 2.3 we propose a classifier to detect new drugs using just a small set of occurrences of some known drugs. In this way we expect that using a classifier trained with known NPS, it will be able to identify unknown substances, helping EU to ban such chemicals timely.

[11] http://www.emcdda.europa.eu/start/2016/drug-markets#pane2/4
[12] https://www.drugabuse.gov/publications/research-reports/prescription-drugs/director
[13] http://www.projectcassandra.eu

1.2 Social Media and disaster management

The Himalayan earthquake[14] (Nepal, 2015), the Hurricane Sandy[15] (Central and North America, 2012), and the Emilia earthquake[16] (Italy, 2012) are only some of the major disasters that occurred in the era of technology-mediated social participation [146] during which affected individuals shared their experiences through SM platforms. Thanks to the ubiquitousness and real-time data sharing enabled by widespread mobile devices, in the aftermath of disasters SM become rapidly overwhelmed by messages conveying actionable and time-sensitive information [73]. This information meets the needs of both the helping and the affected communities, as proven by the interest of emergency responders in envisioning innovative approaches able to merge data collected from traditional physical sensors with that crowdsourced through networks of humans, or social sensors [127].

However, crowdsourced information is often unstructured, heterogeneous and fragmented over a large number of messages. To make it usable, therefore, relevant content must be mined and aggregated in order to provide contextual information to emergency responders [9]. Crisis mapping is an effective method for increasing situational awareness as it allows the real-time gathering and visualization of data contributed by a large number of individuals. Crisis maps can also help supporting resource allocation and prioritization during emergencies, when key resources are overwhelmed by the sudden increase in the demand curve [21]. During recent disasters, civil protection agencies developed and maintained live, Web-based crisis maps to help visualize and track stricken locations, assessing the damage and coordinating the rescue efforts [130].

The possibility to exploit SM data for crisis mapping has been first envisioned in works such as [79, 128]. Since then, there has been a growing interest in all areas related to crisis mapping: from data acquisition and management, to analysis and visualization [7]. Among currently widely adopted crisis mapping platforms are Ushahidi [17], ESRI ArcGIS, and CrisisCommons [17]. The main features of these platforms are related to data acquisition, data fusion and visualization. Such platforms represent hybrid crowdsensing systems where users can voluntarily load data onto the system in a participatory way, or the system can be configured so as to automatically perform data acquisition in an opportunistic way. Recent scientific literature has instead mainly focused on SM data analysis, with novel solutions being proposed to overcome typical crisis mapping challenges such as geoparsing and extracting situational awareness from microtexts [49].

Early tools for producing crisis maps from crowdsourced data usually crawl SM (e.g., Twitter) for crisis-specific keywords, and geolocate messages based on GPS latitude and longitude coordinates metadata. However, statistics show that only up to 4% of SM messages carry GPS geospatial metadata [38]. To overcome this limitation that drastically reduces the number of useful messages and results in very sparse maps, new systems *geoparse* the textual content of emergency reports to extract mentions of known places. A common approach to geoparsing, as done by the state-of-the-art

[14] https://en.wikipedia.org/wiki/April_2015_Nepal_earthquake
[15] https://en.wikipedia.org/wiki/Hurricane_Sandy
[16] https://en.wikipedia.org/wiki/2012_Northern_Italy_earthquakes
[17] https://www.ushahidi.com/

system [130], looks up a number of preloaded resources (e.g., the Geonames[18] and GEOnet Names[19] global gazetteers) containing all the possible matches between a set of toponyms and their geographic coordinates. This approach requires an offline phase where the system is set to work in a geographically-limited region. Apart from scalability issues – the wider the area covered, the higher the amount of data to load and manage – this approach poses limits on the area covered by the system since incidents occurring outside the monitored area cannot be mapped. Another issue related to geoparsing is that of toponymic polysemy. A toponym might have multiple meanings, some of which might refer to different locations while others might not refer to places at all ("Washington" may refer to the first US president, to the US state, to the US capital, etc.). Crisis maps are then generated by comparing the volume of SM messages that mention specific locations with a statistical baseline. Other solutions for the geoparsing task have been recently proposed in [57, 74, 75] where authors experimented with heuristics, open-source named entity recognition software, and machine learning techniques.

Other works emphasized the extraction of actionable and time-sensitive information from messages. Authors of [177] apply natural language processing techniques to detect messages carrying relevant information for situational awareness during emergencies. In [98] a technique is described to extract "information nuggets" from tweets – that is, self-contained information items relevant to disaster response. While these works present fully automatic means to extract knowledge from texts, in [178] a hybrid approach is proposed exploiting both human and machine computation to classify messages. All these linguistic analysis techniques for the extraction of relevant information from disaster-related messages have however never been employed in a crisis mapping task. Recently a survey [96] presented an extensive review of current literature in the broad field of SM emergency management, and can be considered for additional references.

To overcome these limitations, we built a general and flexible SM-based crisis mapping system able to create a situational description impromptu, through crowdsourced reports, and without any prior knowledge of the location and extension of the stricken area, as described in Chapter 3. Furthermore, to better support resource prioritization during emergency response, the system ranks identified stricken areas according to the estimated amount of damage they suffered, thus going in the direction of supporting *the greatest good for the greatest number* [21].

To reach these goals, the proposed crisis mapping system works solely with SM data related to unfolding emergencies, detects damage to infrastructures or injuries to the population by exploiting linguistic features and a machine learning classifier, geoparses messages by relying on readily available online semantic annotation tools and collaborative knowledge-bases, and produces Web-based interactive crisis maps. In particular, in Chapter 3 we compare two different approaches to the damage detection problem, one based on the Natural Language Processing (NLP) technique and the other one based on word embeddings, which exhibits performances that are close the NLP solution but with some advantages related to language independence. The output of the system has been repeatedly validated against authoritative data released by Civil

[18] `http://www.geonames.org`
[19] `http://geonames.nga.mil/gns/html/index.html`

Protection and local administrations regarding a series of large-scale natural disasters that struck Italy in past years. Finally, in Section 3.4 we discuss a general approach to SM crawling and analysis oriented to Big Data, and illustrates the state-of-the-art technologies able to cope with massive data streams like Twitter.

1.3 Social Media: the house of haters?

SM user base varies a lot from platform to platform, and per country basis. More specifically, some SM exhibit a young community, while others are more balanced across different age ranges. This statistic might change a lot accordingly to local communities, cultural reasons and market trends. Also, sex balance can differ among SM and between countries. In particular, some governments adopt restrictions in the usage of these platforms. It is not a secret that China and other countries apply a severe censorship to news and limit or control SM access and contents [39, 106]. These restrictions can be applied also to Web browsing. In some countries the limitations are not applied to the whole population but only a part, i.e., women. However, the restrictions or censorship applied to SM may also occur in more open-minded and democratic countries, like Italy. Censorship may be used for different reasons: "the military may censor the identity of personnel and the judiciary may censor the identity of minors and victims" [160]. In the latter case, newspapers may alter the real names or leave just their initials in the news articles. Notably, people often tend to censor (or hide) their identity also when conducting attacks, offenses or malicious actions against other users. The reluctance of SM platforms to share real identities of users behind SM profiles makes their identification a hard task. This behaviour obstacles the law enforcers in containing the hate.

Unfortunately, SM are the ideal plaza for the proliferation of harmful information. Cyberbullying, sexual predation [108], self-harm practices incitement [35] are some of the effective results of the dissemination of malicious information on SM. Many of these attacks are often carried out by a single individual, but they can also be managed by groups. The motivations of such behaviors can be various, i.e., a political or racial difference can be sufficient to motivate groups of haters to attack victims. Frequently, these attacks occur directly against victims. In some other circumstances haters self-organize in SM groups or pages that foster hate with the high risk of generating riots across cities[20,21]. Many young people, corrupted or confused by SM content, joined riots and took part in episodes of violence.

The issue is not new to SM although it has never been seriously considered by platforms. It has been tackled in the past with different approaches, laying somehow in the middle between pre-emption and mending. In particular, some techniques developed in the past were aimed at preventing the publication of inappropriate material or a fast removal of content, but recently the large diffusion of the content, even cross platform, has limited the application of this approach. Moreover, such restriction is often perceived by the user as a censorship from the platform. However, there are few practical tools and instruments to limit this raising plague.

[20] http://www.lastampa.it/2015/06/28/cronaca/no-tav-a-exilles-in-centinaia-alla-partenza-del-corteo-movimenti-kajhKIwkxF5l5noaCfCn6N/pagina.html

[21] http://www.ilfattoquotidiano.it/2015/05/01/expo-2015-corteo-no-expo-milano-vetrine-spaccate-idranti-e-lacrimogeni-della-polizia/1642882/

To detect the diffusion of hate speech on SM and react properly to the diffusion of hateful content, we propose an approach valid for the Italian language, presented in Chapter 4. In particular, we analyze the phenomenon of hate speech on Facebook and propose a novel taxonomy to cluster hate speech as well as a classifier model to identify posts and comments conveying a hateful message.

Considering the potentially dangerous consequences of hate, either emotional and/or physical, originated from SM content, we think that such a problem is worth analyzing and requires special treatment in order to improve the common good and the quality of our relationships (virtual or real).

Additional benefit may come from the identification of censored identities in SM messages. Concerning this issue, in Chapter 5 we analyze a possible method to circumvent censorship applied to Facebook posts (but easy to generalize to other SM platforms), with particular regard to news spreading on SM.

CHAPTER 2

Drugs and SM

The aim of this Chapter is to describe the study of the diffusion of the Novel Psychoactive Substances (NPS) on the Internet with a special regard to SM platforms. The tools and methodologies described hereafter are mainly derived from [58] and [60].

Recently, academia has started investigating the massive use of SM and online forums to advertise and discuss psychedelic substances and drugs and how the preferences of online communities can affect those of consumers. Large forums drew attention, being a primary source of information about NPS and a good sample of consumer tastes [54]. Some forums and SM have already been scanned to look for evidence of drug consumption and discussion. For example, work in [117] considers the Flashback forum and traces the trend of the discussions, especially in relation with the scheduling of a substance ban, highlighting how volumes of discussions drop when a ban is scheduled. In [192], the authors focus on new drugs detection and categorisation by scanning online shops and the dark net. A complete list of the known effects of new drugs, to the publication date, is given in [92, 158].

Work in [167] analyses small subsets of the contents of the Drugsforum and Bluelight forums, which will be deeply analysed later in this work in Chapter 2, highlighting how large forums embody a cumulative community knowledge, i.e., a stratified knowledge built over years of forum activities, and showing that drug effects and dosage are among the most discussed topics. Such SM have recently been dissected in detail in two works [58, 60] which describe the structure of the forums and used posts of user to train a classifier for detecting drugs, possibly NPS.

Other studies explored the abuse of medicines and how these are advertised, e.g., on Twitter, and sold by online pharmacies, with no authorisation [70, 102]. Twitter features a rapid spread of contents, especially through small communities of users, which share common interests and tastes. This is the main reason why it has been investi-

gated to mine patterns of drug abuse, also for non-medical purposes, e.g., to improve students performances in study [87, 88]. Furthermore, Twitter allows analysts to comprehend rapid disease diffusion and health issues [142], as well as prices and effects of new drugs [140]. Nevertheless, SM play an important role also for contrasting drug diffusion [148] and for preventing end users from further consumption [99]. Twitter was also extensively mined to detect geographical diffusion of drug consumers over time [24].

The Web is not the only marketplace where NPS are advertised and sold. Indeed, the TOR network[1] has drawn much attention from drug consumers and resellers, who search for a channel to buy and sell drugs that guarantees their anonymity. This aspect affects trustworthiness of peers, especially when it is not possible to assess users reputation at all. In [90], the authors investigated the impact of reputation in Silk Road, one of the most popular marketplaces for drugs in the dark net. Data analysis often deals with the quality of the results obtained when searching the web. The work in [145] describes the possibility to improve the recall of queries issued to search engines by exploiting all variants and misspelled words.

With respect to related work, this thesis addresses a finer-grained, more detailed picture of NPS data sources and NPS data available on the Internet to understand the NPS phenomenon and build tools to timely detect when new chemicals enter the market and monitor them. As an example, the analysis of forums carried on in [167] was limited in time and quantity. This work overcomes this limitation, by analysing more than one decade of data, posted by users all over the world. Overall, we dealt with more than 4 million and a half posts and more than 500,000 users and integrating more than one source, by monitoring two forums, Twitter, and a number of online shops. The results of the analysis are conveniently conveyed to the reader via a set of interactive visual web interfaces, which are being integrated into a dashboard that will help researchers mine the wealth of gathered data. Ultimately, this work is aligned with recent advances in data analysis leading to applications in pattern mining of, e.g., medical records and human anatomies [56, 91].

In this Chapter an insight into two popular forums will be provided, Bluelight[2] and Drugsforum[3], hosting drug discussions for more than one decade and a mapping of the NPS sales (as monitored on online shops) and NPS diffusion and distribution (as monitored on discussion forums) as proposed in [58]. Section 2.3 will deepen the analysis of forums and will propose a novel methodology to identify possible NPS, detailed in [60].

2.1 Data sources

This section presents the data sources for our analysis. The data collection occurred through ad-hoc software, which scrapes websites and uses API to crawl SM.

2.1.1 Forums

Bluelight and Drugsforum are two large forums, which host more than a decade of discussion about drugs and addiction. Being particularly rich in information, the two

[1] https://www.torproject.org
[2] http://www.bluelight.org
[3] https://drugs-forum.com

Forum	First post	Last post	Tot posts	Users
Bluelight	22-10-1999	09-02-2016	3,535,378	347,457
Drugsforum	14-01-2003	26-12-2015	1,174,759	220,071

Table 2.1: *Drug forums: Posts and Users*

forums provide a historical, worldwide background of drug consumption, comprising that related to NPS. Similar to Google Flu Trends[4] efforts to detect spreading of diseases, the analysis of the forums' content and structure is significant to understand how psychoactive substances have spread out and to study new infoveillance strategies, to timely detect drug abuse.

The two forums have a hierarchical structure, which enables proper content categorisation. The root of both forums organises content into sub-forums, which can be nested up to several levels of depth. The forums' structures were subject to different content re-organisations over time.

Through the Web scraping activity it was possible to create a dump of the entire database of discussions from the two forums, following the links between the forums' sections. During the storage phase, the scraper kept track of the forums' hierarchy and structure, maintaining all the tags and metadata associated to each post and thread. Table 2.1 summarises the amount of data available from the two forums.

The available data comprises more than half a million users and more than 4.6 million posts. Data was stored in a relational database for further querying. These forums were earlier and partially analysed in [167] and then explored in detail [58].

2.1.2 Online shops

The forums introduced in Section 2.1.1 are a primary source of information about drugs reviews, feelings, effects and preparation, but little information is available about the drug markets, such as prices and bulk quantities. Thus, to understand the link between communities of drug consumers and the supply chain, the work focuses the attention also on other data sources, dealing with drug trading.

Online shops sell both legal and illegal substances. Among others, those that sell NPS have grown in popularity, given the relatively low risks in trading such substances. Many online shops accept payments in pounds, euros and dollars. Also, bitcoins are often accepted. This opens up the possibility to track price trends and, indirectly, to estimate the popularity and quality (or purity) of drugs. Furthermore, many of the marketplaces are advertised and mentioned on forums and SM.

Products, quantities and prices have been collected through an intense scraping activity on a set of online shops to monitor the market availability of different substances. Online shops can be quite easily found through simple queries to search engines (e.g., "legal highs" and "smart drugs").

Recently UK promulgated the "Psychoactive Substances Act 2016"[5] with the aim of limiting the proliferation of NPS and other hallucinogenic chemicals. The act "makes it an offence to produce, supply, offer to supply, possess with intent to supply, possess on custodial premises, import or export psychoactive substances; that is, any substance

[4] https://www.google.org/flutrends/about/

[5] https://www.gov.uk/government/collections/psychoactive-substances-bill-2015

Website	Substances found
http://chem-shop.co.uk	7
http://researchchemist.co.uk	45
http://researchchemistry.co.uk	56
http://sciencesuppliesdirect.com	43
http://www.bitcoinhighs.co.uk	4
http://www.buylegalrc.eu	17
http://www.legalhighlabs.com	33
http://www.ukhighs.com	51
https://www.buyanychem.eu	78
https://www.iceheadshop.co.uk	68

Table 2.2: *Monitored online shops and number of substances they sell*

intended for human consumption that is capable of producing a psychoactive effect". Nevertheless, this act is limited to UK only and at the time of the scraping activity it was not yet effective.

A battery of scrapers collected the information that are present on shop showcases for some months to provide a deep understanding of the market over time. Data is collected on a weekly basis, and stored in a relational database, to be easy queryable. Table 2.2 shows the monitored shops.

2.1.3 Twitter

Twitter is extensively used by resellers and "pharmacies" to advertise psychoactive substances, and by consumers to discuss their effects and share feelings with others [70, 102]. A crawler collected about 14 million tweets, over the period March 16, 2015 - February 2, 2016, using the Streaming API[6], which allows applications to gather tweets in real time fashion. The crawler that gets data relies on a set of ad-hoc keywords. The crawler also followed a series of Twitter accounts associated to online shops. The next Section will detail the monitored keywords, chosen among known emerging substances.

2.2 Data analysis and Visualisation

This section shows the analysis carried out over the data sources described in Section 2.1, with the purpose of figuring out forums' structural features, how their content is organised, and the geographical distribution of their users.

Furthermore, the mining activity over forums' textual contents will show a simple methodology to detect possible candidates of new substances mentioned in recent discussions.

Finally, this section provides a picture on the NPS substances sold on online shops, correlating them with mentions on Twitter and the forums.

2.2.1 Forums: Structural and geographical features

To facilitate the investigation of the forums structural features, the system offers a set of visual interfaces. Figure 2.1 depicts the screenshot of a zoomable treemap of the two

[6] https://dev.twitter.com/streaming/overview

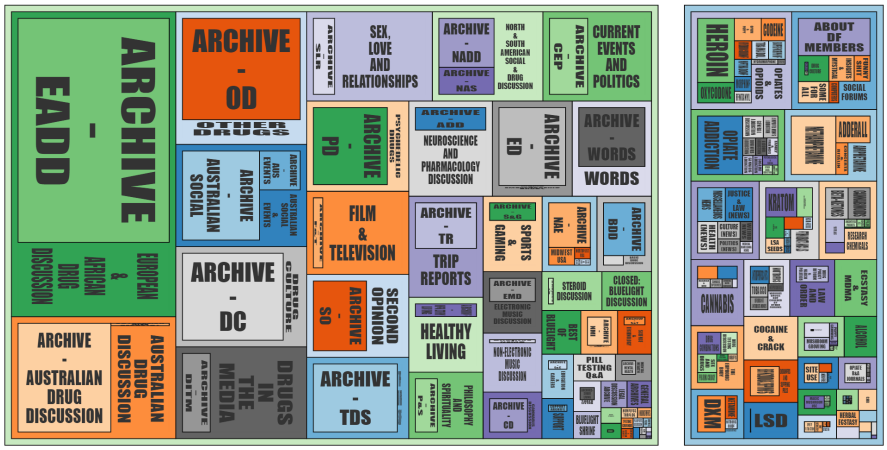


Figure 2.1: *The structure of Bluelight (left) and Drugsforum (right). Bluelight is about three times bigger.*

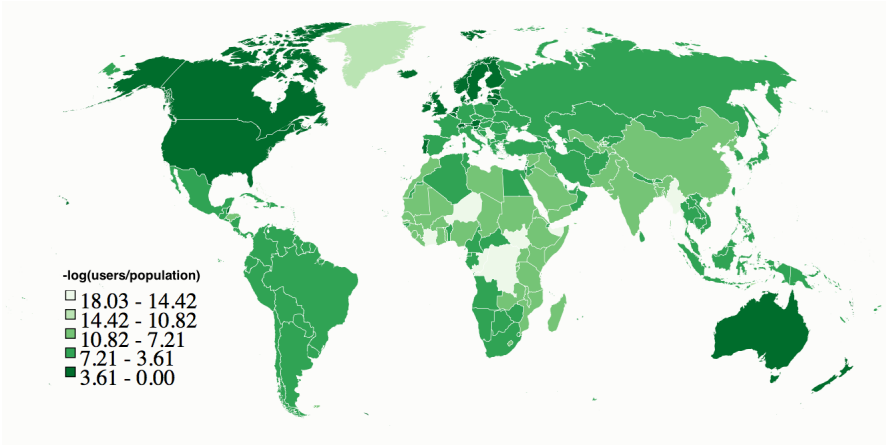


Figure 2.2: *Geographical distribution of Drugsforum users.*

forums. Nested subsections are represented as nested rectangles, the area of which are proportional to the number of posts a subsection contains. Quick visual comparisons of the forums' size and structure may gather meaningful information. For example, compared to Drugsforum, whose structure is quite complex, Bluelight has a shallow organisation. Also, the names of the subsections suggest that the discussion on Drugsforum is mainly focused on drugs and it follows a rigid categorisation, based on the kind of the substance, while the topics on Bluelight are broader and less related specifically to drugs.

Figure 2.2 shows the worldwide distribution of the Drugsforum users. The information has been extracted from the users' profiles (when available). Looking at the figure, it comes to the attention that drug discussions on forums is a wide phenomenon, quite naturally leading to a widespread word of mouth. The colours in the figure are proportional to the density of users. Noticeably, the most involved areas are North America, Australia, UK, and Scandinavia.

The investigation of the forums includes some topological aspects, like the number of posts per user and the number of posts per thread, on both forums (Figure 2.4). The

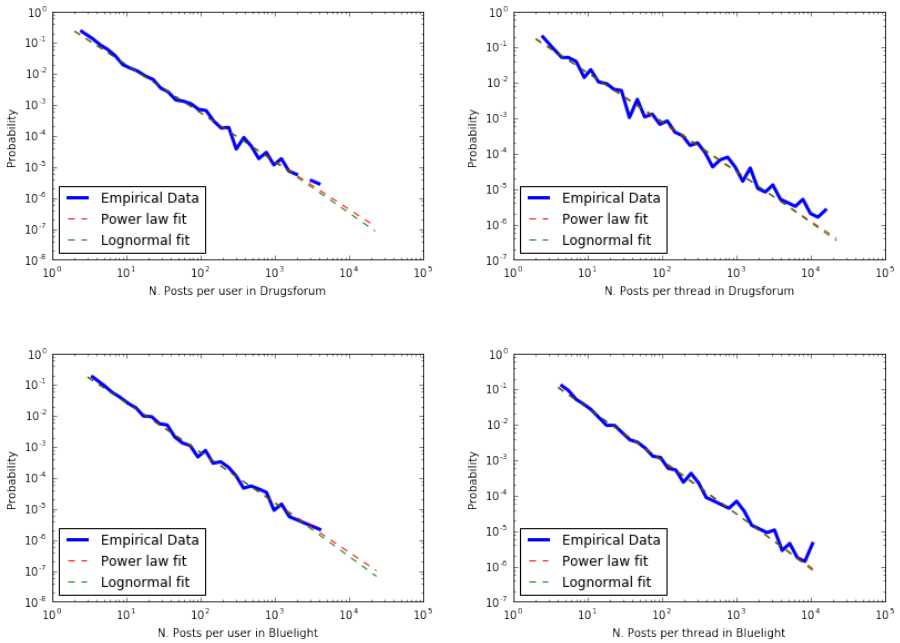


Figure 2.3: *Probability Density Function of the real data and the different distributions*

powerlaw Python package [2] has been used to compare the four real data distributions with the exponential, power law, truncated power law and lognormal distributions. The tool measured the *xmin*, no more than 4 for all the cases. Furthermore, with a p-value less than 10^{-8} for all the distributions, the power law distribution results in a better fit than the exponential one, as expected [134]. With regard to the lognormal and truncated power law distributions, the lognormal distribution fits slightly better than the power law one, while the truncated power law distribution fits better than the lognormal one. We can conclude that the (truncated) power law distribution assumption holds, as shown in sections 2.2.1 to 2.2.1. These results highlight that there is a small amount of users responsible for most of the activity, on both forums.

It is worth noting that, even if Bluelight has about 0.6 times the number of users Drugsforum has (see Table 2.1), the number of active users (i.e., that have written at least one post) is almost the same for both. As for the distribution of posts per thread, shown in Figure 2.5, Bluelight features a large number of threads having 1,000 posts. This is due to a limit on the maximum number of posts for certain threads: when exceeding the threshold, the moderators start a new thread for the discussion.

2.2.2 Content analysis

A text analysis that is really useful in our scenario is the measurement of volumes of discussion over time, given a term. This investigation helps to determine whether some drugs raise in popularity and in which section of the forum this happens, possibly obtaining some clues about the nature of the substance (being a NPS or not).

Figure 2.6 shows the frequency of the term "mephedrone" over time, normalised to the whole volume of discussion, for Drugsforum (top) and Bluelight (bottom). Even if not identical, the shapes of the spike are similar, meaning that the substance has gained

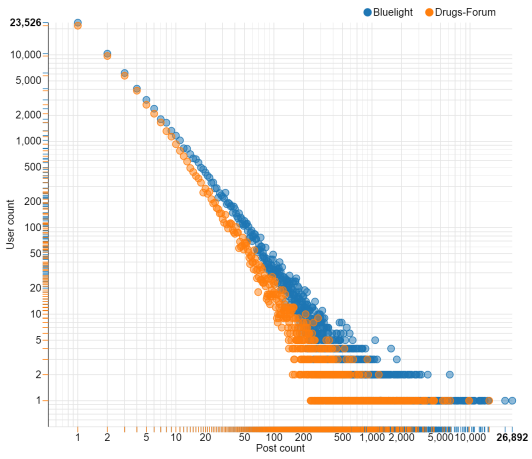


Figure 2.4: *Posts per user.*

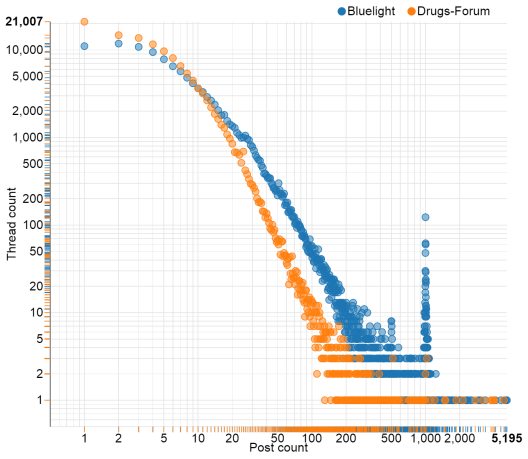


Figure 2.5: *Posts per thread.*

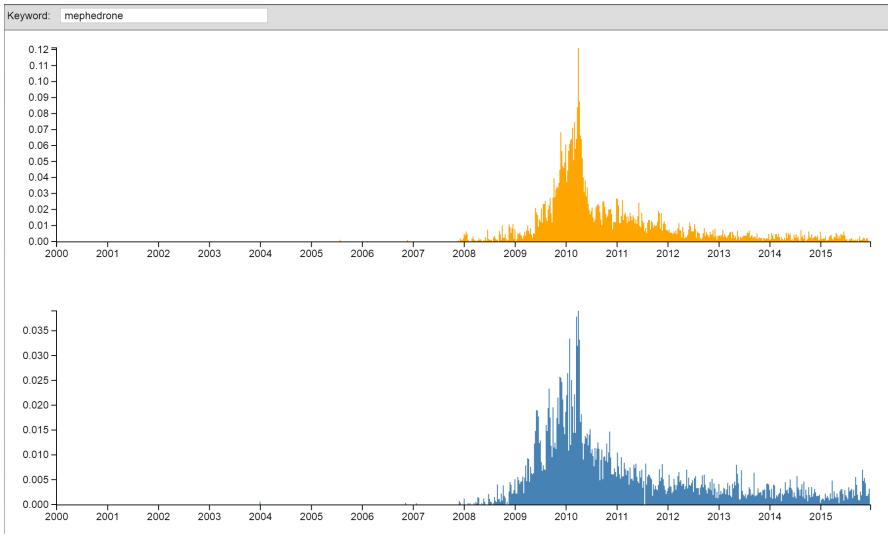


Figure 2.6: *Frequency of "mephedrone" over time, normalised to the whole volume of discussion, for Drugsforum (top) and Bluelight (bottom).*

popularity within both communities approximately at the same time.

Figure 2.7 shows a higher level of detail: each line represents a subsection of the forum. As shown in the top-left part of the screenshot, the user can choose which forum to analyse. A darker colour indicates a higher frequency of the term, for the corresponding time frame. The search for "mephedrone" in Drugsforum shows a high volume of discussion in the first half of 2010 in a series of subsections, particularly in the one called "Beta-Ketones". This indicates the category of the substance.

As shown in the example of Figure 2.8, computing the terms that co-occur with a given one gives interesting insights. Indeed, the generated wordclouds may provide knowledge on substances that are similar, with similar effects and market trends. In the figure, each word occupies an area that is proportional to its frequency. The wordclouds can be generated for both Twitter and the two forums.

Really endorsed drugs are presented and discussed in forums. To timely detect NPS,

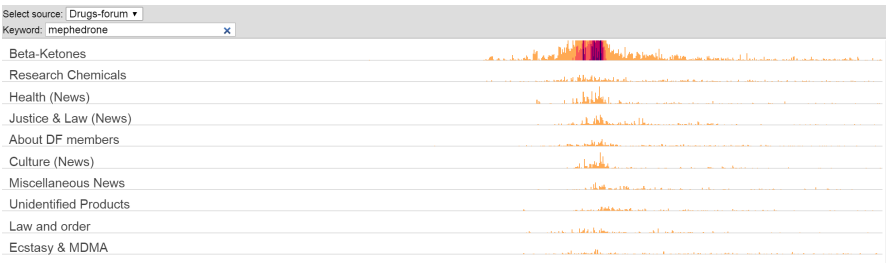


Figure 2.7: *Horizon charts showing the frequency of a given term over time, for each subsection of the chosen forum.*

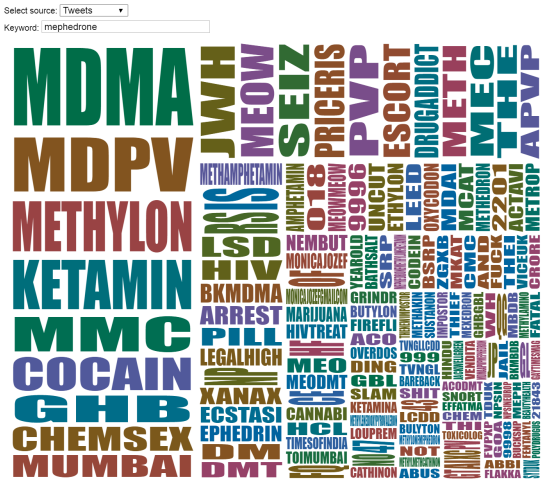


Figure 2.8: *Zoomable wordcloud showing the most frequent terms co-occurring with "mephedrone" in the Twitter dataset.*

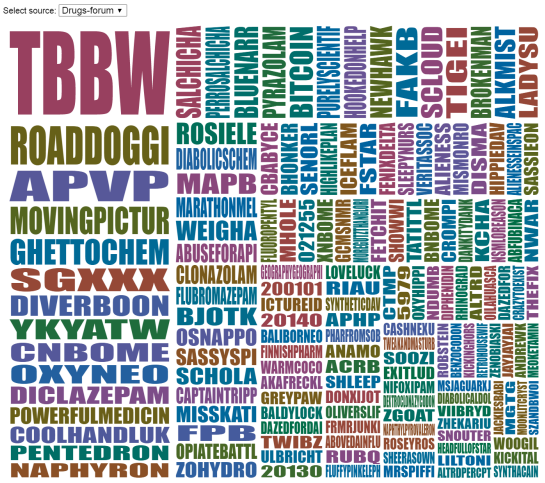


Figure 2.9: *Zoomable wordcloud showing new terms in Drugsforum after 2010.*

the work has investigated neologisms and terminology on both forums, to discover new names. As an example, Figure 2.9 represents the Drugsforum terms that appeared only after 2010. The result clearly indicates a lot of new drugs on the market from 2010 to 2015. It is possible to notice the name of some new drugs and medicines, such as α-PVP, Diclazepam, Pentedrone, Naphyrone.

2.2.3 NPS trading

As a final set of analyses, the work has explored the hyperlinks on the forums. Then, it has compared them with a comprehensive list of NPS online shops and with the links in the posts of monitored Twitter accounts. Not surprisingly, they do not overlap, meaning that forum discussions do not link shops. This is mainly due to the specific policies of the forums.

A search in the forum has also tested which are the NPS sold on the shops and also mentioned in forums, finding that almost every substance is mentioned. It is not possible to estimate the trade volume of NPS from online shops, but it is possible to infer some information about popularity by observing the discussions in forums. Checking the words frequency, in conclusion they are the very same substances also advertised through Twitter. Table 2.3 reports an excerpt of some substances, with a

Drug	Tweets	Post BL	Post DF	Online shops	First seen
MDAI	913	3507	775	1, 3, 4, 9	Bluelight
MDPV	791	11304	3631	9	Drugsforum
Methylone	679	8254	5116	9	Bluelight
AB-CHMINACA	584	16	33	4, 6, 9	Drugsforum
Methiopropamine	515	329	232	2, 3, 7, 8, 9, 10	Bluelight
1P-LSD	483	612	69	1, 2, 3, 4, 9	Bluelight
Etizolam	1592	8629	2630	2, 4, 9	Bluelight
Ethylphenidate	965	2502	1268	2, 7, 9	Bluelight
Synthacaine	217	124	60	3, 4, 9, 10	Drugsforum
Diphenidine	193	779	80	2, 3, 4, 9	Bluelight
Mexedrone	39	113	14	1, 2, 3, 4, 9, 10	Bluelight

Table 2.3: *An excerpt of monitored substances, with no. tweets, posts and shops. Bluelight (BL) and Drugsforum (DF) are the two forums analysed in this work. The last column highlights the forum where the substance has appeared first.*

measure of the discussion activity about them on Twitter, in forums and on online shops. In the table, the numbers in the column of online shops are the IDs of the shops, as in Table 2.2. The meaning is: the drug is mentioned on those shops.

2.3 Detecting drugs: a picture of the Academia effort and a novel approach

In Section 2.1 we discussed the phenomenon of drugs across different types of SM and in particular we focused the attention on the problems introduced by the trade of Novel Psychoactive Substances which pose new perspectives in the fight to drugs. Noticeably, online shops are particularly focused on NPS and forums data contain a lot of information about the new generation of substances, and the tools proposed in [58] and 2.2 help to analyse the phenomenon to spot the introduction of new substances.

However, the detection of novel drugs can be somehow automated and the following Sections describe the DAGON (DAta Generated jargON) algorithm [60], a novel, semi-supervised knowledge extraction methodology, that has been applied to the posts of the drug forums, with the main goals of: i) detecting substances and their effects; ii) putting the basis for linking each substance to its effects. A successful application of our technique is paramount: first, it introduces the possibility of shortening the detection time of NPS; then, it will be possible to group together different names that refer to the same substance, as well as to distinguish between different substances, commonly referred to with the same name (such as "Spice" [157]) and timely detect changes in drug composition over time [55]. Finally, knowing the effects tied to novel substances, first-aid facilities may overcome the current difficulties to provide effective countermeasures.

While traditional supervised techniques usually require a large amount of hand-labeled data, our proposal features a semi-supervised learning approach in order to minimize the work required to build an effective detection system. Semi-supervised learning exploits unlabeled data to mitigate the effect of insufficient labeled data on the classifier accuracy. This specific approach attempts to automatically generate high-quality training data from an unlabeled corpus. With very little information, DAGON solution is able to achieve excellent detection results on drugs and their effects, with an

FMeasure close to 0.9.

Recently, Academia has started mining online communities, to seek for comments on drugs and drug reactions [190]. Indeed, forums and SNs offer spontaneous information, with abundance of data about experiences, doses, assumption methods [54, 58]. Authors in [137] realized ADRMine, a tool for adverse drug reaction detection. The tool relies on advanced machine learning algorithms and semantic features based on word clusters - generated from pre-trained word representation vectors using deep learning techniques. Also, intelligence analysis has been applied to SM to detect new outbreaking trends in drug markets, as in [182]. A raising phenomenon connected to the consumption of psychoactive substances is the adoption of nonmedical use of prescription drugs [123], such as sedatives, opioids, and stimulants. Even these drugs are often traded and advertised online by fake pharmacies [70, 102].

The amount of data available nowadays has made automated text analysis veer towards more machine learning-based approaches. Because complex tasks might require many training examples, however, there is a vivid study on unsupervised and semi-supervised approaches. Our task encompasses identifying names in text, something often associated with named-entity extraction. Unsupervised methods such as [166] use unlabeled data contrasted with other data assumed irrelevant - to use as negative examples - in order to build a classification model. Instead, we use seeds, a small set of examples, because the writers on forums often attempt not to mention drugs explicitly, resorting to paraphrases or nicknames, making a purely contrastive approach difficult to apply. Also, multi-level bootstrapping proved to be a valid improvement in information extraction [149]; these techniques feature an iterative process to gradually enlarge and refine a dictionary of common terms. Our approach, instead, splits the problem of finding candidate terms and classifying them in two separate subproblems, the second of which is fed with a small number of annotated examples, i.e., the seeds. Co-training is a common technique [23] to evaluate whether to use an unlabeled piece of data as a training example: the idea is building different classifiers, and using the label assigned by one as a training example for another. The application described in this Chapter instead leverages the redundancy among the data, to ensure candidate examples are selected with a high degree of confidence. Relation extraction is an even more complex task which seeks the relationships among the entities. This is relevant here, because substances can only be identified by basing on their role in the sentence (since common names are often used to refer to them). Work in [151] proposes a method based on corpus statistics that requires no human supervision and no additional corpus resources beyond the corpus used for relation extraction. Our approach does not explicitly address relation extraction but it exploits the redundancy of a substance (or effect) being often associated with other entities to identify them. KnowItAll [66] is a tool for unsupervised named entity extraction with improved recall, thanks to the pattern learning, the subclass extraction and the list extraction features that still includes bootstrapping to learn domain independent extraction patterns. In the current scenario, common mention patterns are also strong indicators of the substance or effect class; however, we do not use patterns to extract, but only, implicitly, for classification purposes. Furthermore, [29] pursues the thesis that much greater accuracy can be achieved by further constraining the learning task, by coupling the semi-supervised training of many extractors for different categories and relations; we use a single multiclass classifier to

achieve the same goal. Under the assumption that the number of labeled data points is extremely small and the two classes are highly unbalanced, the authors of [187] realized a stochastic semi-supervised learning approach that was used in the 2009-2010 Active Learning Challenge. While the task is similar, our approach is different, because we do not need to use unlabeled data as negative examples. The framework proposed in [34] suggests using domain knowledge, such as dictionaries and ontologies, as a way to guide semi-supervised learning, so as to inject knowledge into the learning process. In this case the system did not rely on rare expert knowledge for the task, arguing that a few labeled seeds are easier to produce than dictionaries or other forms of expert knowledge representations. A mixed case of learning extraction patterns, relation extraction and injecting expert knowledge is in [18], which also shows the challenge of evaluating a technique when few labeled examples are available.

Work	**Target**	**Technique**
Smith [166]	Train on unlabeled data.	Contrastive estimation.
Riloff [149]	Build of a multi-level bootstrapping algorithm that generates both the semantic lexicon and extraction patterns simultaneously.	Mutual bootstrapping technique to alternately select the best extraction pattern for the category and bootstrap its extractions into the semantic lexicon, which is the basis for selecting the next extraction pattern.
Blum [23]	Use a large unlabeled sample to boost the performance of a learning algorithm when only a small set of labeled examples is available.	Two learning algorithms are trained separately on each view, and then each algorithm's predictions on new unlabeled examples are used to enlarge the training set of the other.
Rosenfeld [151]	Use corpus statistics to validate and correct the arguments of extracted relation instances, improving the overall relation extraction performance.	The method used is based on corpus statistics and requires no human supervision and no additional corpus resources beyond the corpus that is used for relation extraction.
Etzioni [66]	The paper presents an overview of KnowItAll's novel architecture and design principles.	The paper presents three distinct ways to address information extraction challenges without hand-labeled data and evaluates their performance: Pattern Learning, Subclass Extraction, List Extraction.

Carlson [29]	The paper pursues the thesis that much greater accuracy can be achieved by constraining the learning task, coupling the semi-supervised training of many extractors for different categories and relations.	Characterize several ways in which the training of category and relation extractors can be coupled.
Xie [187]	Build a stochastic semi-supervised learning approach to tackle the binary classification problem under the condition that the number of labeled data points is extremely small and the two classes are highly imbalanced.	The algorithm starts with only one positive seed given by the contest organizer. Then, it randomly picks additional unlabeled data points and treats them as "negative" seeds based on the fact that the positive label is rare across all datasets. A classifier is trained using the "labeled" data points and then is used to predict the unlabeled dataset. The final result is the average of n stochastic iterations.
Chang [34]	Study a method for incorporating domain knowledge in semi-supervised learning algorithms.	Use of a framework, based on constraints, that unifies and exploits several kinds of task-specific constraints. It scores 80.6% accuracy.
Bellandi [18]	Build of an ontology-driven system that performs relation extraction over textual data.	The system exploits expert knowledge of the domain, including lexical resources, in the form of an ontology to drive the extraction of patterns using manually annotated texts. Such patterns are then applied in order to identify candidates for relation extraction. Paired with basic, reliable named-entity-level text annotation, this results in the discovery of relations among entities in Italian newspaper articles.

Attardi [3]	Categorization by context.	Exploit both the structure of Web documents and Web link topology to determine the context of a link.
Del Vigna [60]	Semi-supervised approach to knowledge extraction, **applied to the detection of drugs and effects**.	Based on the very small set of initial seeds, the work highlights **how a contrastive approach and context deduction are effective** in detecting substances and effects from the corpora, with F-score close to 0.9.

Table 2.4: *Selection of relevant literature for the training of a classifier from a non-labeled corpus of data, highlighting this Thesis contribution.*

As said above, the problem of building a model with a limited set of information, but with a large enough amount of data, has been tackled by various angles. Table 2.4 summarizes all the relevant approaches to face a classification task when the training set is unlabelled or scarcely labelled. As starting point thus we have: a) the availability of a large set of unlabeled data, and b) the availability of a small set of labeled substance and effect names.

2.4 Datasets for experiments

To validate the approach described in Section 2.5, the algorithm is tested over two different large data sources, in order to consider a variety of contents and information, and to push the automatic detection of drugs. Experiments are conducted over the very same dataset of section 2.1 which include more than a decade of posts from Bluelight and Drugsforum, as shown in Table 2.1. This dataset and the seeds constitute the starting point for the algorithm introduced in Section 2.5.

2.4.1 Seeds

For the sake of the experiments we will make use of a list of 416 drug names of popular psychoactive substances downloaded from the Internet, including the slang which is adopted among consumers to commonly name them, from the website of the project *Talk to Frank*[7] and a dataset containing 8206 pharmaceutical drugs retrieved from Drugbank[8]. This list constitutes a ground truth for known drugs.

Also, in the experiments we will make use of a list of 129 symptoms that are typically associated to substance assumption.

[7] http://www.talktofrank.com
[8] http://www.drugbank.ca

2.5 The DAGON methodology (DAta Generated jargON)

This section, introduces DAGON, a methodology that will be applied in Section 2.6 for the task of identifying new "street names" for drugs and their effects. A street name is the name a substance is usually referred to amongst users and pushers.

The task of name identification can be split into two subtasks:

(a) Identifying text chunks in the forums, which represent candidate drug names (and candidate drug effects);

(b) Classifying those chunks as drugs, effects, or none of the above.

The first subtask - identification of candidates - could be tackled with different approaches, including a noun-phrase identifier[9], usually based on a simple part-of-speech-based grammar, or on a technique akin to the identification of named entities, as in [125].

In this work, the identification of candidates is based on domain terminology extraction techniques based on a contrastive approach similar to [143]. Essentially, we identify chunks of texts that appear to be especially significant in the context of drug forums. Based on the frequency in which terms appear both in the posts of drugs forums and in contrastive datasets dealing with different topics, we extract the most relevant terms for the forums. We have extracted unigrams, 2-grams, and 3-grams. This approach does not require English specific annotated resources and, thus, it can scale easily to different languages.

The second subtask is a classification problem. Following a supervised approach would have required having annotated posts and using them as the training set for our classifier. Instead, we have chosen to work on unlabeled data (i.e., the posts on the drugs forums, see Section 2.4) and to exploit the external list of seeds introduced in Section 2.4.1.

We represent a candidate by means of the words found along with it when it was used in a post, selecting windows of N characters surrounding the candidate whenever it was used in the dataset. Hereafter, we call *context* (of a candidate) the text surrounding the term of interest.

Thus, we have shifted the problem: from classifying candidate street names to the classification of their contexts, which are automatically extracted from the unlabeled forum datasets.

It is worth noting that, in the drugs scenario, there would be at least 3 classes, i.e., Substance, Effect, and "none of the above" - the latter to account for the cases where the candidate does not represent substances and effects. However, the seed list at our disposal consists of flat lists of substances/effects names, provided with no additional information (Section 2.4.1). Therefore, in the following, we will first automatically identify positive examples for the two classes (Substance and Effect), training a classifier on them, and then we will tune the classifier settings to determine when a candidate does not fall in either.

Summarising, we have split the task of classifying a candidate into the following sub-tasks:

(a) Fetch a set of occurrences of the term along with the surrounding text (forming in such a way the so called contexts).

[9] A noun-phrase is a phrase that plays the role of a noun such as"the kid that Santa Claus forgot".

(b) Classify each context along the 2 known classes (Section 2.5.3).

(c) Determine a classification for the term given the classification result for the context related to that term (obtained at step (b)).

The single context classification task [3] falls within the realm of standard text categorization, for which there is a rich literature.

Hereafter, the training phase for the classifier will be detailed (2.5.1), and the choice of seeds will be discussed (2.5.2), specifying the procedure for classifying a new candidate (2.5.3), and illustrating a simple approach to link substances to their effects (2.5.4).

2.5.1 Training phase

We are equipped with a list of examples for both the drugs and the effects categories, as described in Section 2.4.1. This list of entry terms is the training set for the classification task and we call it *list of seeds*. Each post in the target drug forums was indexed by a full-text indexer (Apache Lucene[10]) as a single document.

The training phase, whose aim is to build a classifier for candidates, is as follows:

(i) Let T_S and T_E be the set of example contexts, for the Substance and Effects classes respectively, initialized empty.

(ii) From the lists of seeds, pick a new seed (a drug name) for the Substance class and one (an effect name) for the Effects class. A seed is therefore an example of the corresponding class taken from the seed list (Section 2.4.1). See Section 2.5.2 for the heuristic to select a seed out of the list.

(iii) Use the full-text index to retrieve M posts containing the seed s; use only the bit of text surrounding the seed. Section 2.6, will show how results change by varying M. Pick a window of 50 characters surrounding the searched seed.

(iv) Strip s from the text, replacing it always with the same unlikely string (such as "CTHULHUFHTAGN"), in order to avoid the bias carried by the term itself, but maintaining the position of the term in the phrase for classification purposes. Call the texts thus obtained ctx_s (context of seed s).

(v) Add the texts thus generated to the set of training examples for the class C the seed belongs to (either T_S for substances or T_E for the effects)

(vi) Use the training examples to train a multiclass classification model M^{ctx}, which can be any multiclass model, as long as it features a measure (e.g., a probability) interpretable as a confidence score of the classification. Section 2.6 will show results when using SVM with linear kernel [33].

At the end of these steps, we have obtained a classifier of contexts (M^{ctx}), but as seeds (not contexts) are labeled, we are unable to assess its performance directly. We therefore define a classifier of candidate terms (M^{trm}) using the method described later in Section 2.5.3, the performance of which we can assess against the seed list. This allows us to optionally iterate back to step (ii), in order to provide additional seeds to extend the training sets, and improve performances.

[10] `http://lucene.apache.org/`

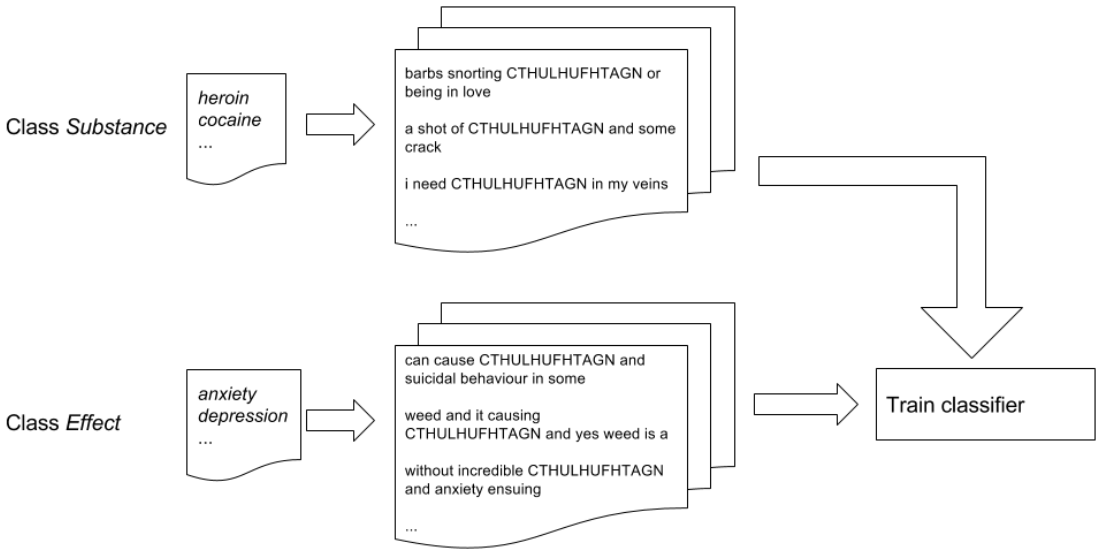


Figure 2.10: *Training phase*

The rationale behind this process is that drug (and effects) mentions will likely share at least part of their immediate contexts. Clearly, when a very small number of seeds is provided (e.g., 1 per class) there will be a strong bias in the examples ultimately used for training, which means that the resulting model will be overly specific to the type of drug used in the training. By providing more seeds, and with enough variety, the model will eventually become more generic to encompass the various drug types, and the relative differences in the contexts in which they are mentioned in the dataset.

2.5.2 Choosing a seed

Obtaining a large seed list is often costly, since it may require having to manually annotate texts, or provide the algorithm with an initial set of words. Thus it is important to design a system with high performances that uses the minimum amount possible of seeds for the training phase. Choosing an effective seed is paramount, and, in doing so, there are various aspects to consider:

(a) Is the seed mentioned verbatim enough times in the data collection? Failing this, the seed will only serve to collect a small number of additional training elements, and it will not impact the model enough;

(b) Is the seed adding new information? The most effective seeds are those whose contexts are misclassified by the current iteration of the classification model. In order to pick the most useful one, we could select, from the list of available unused seeds, those whose contexts are frequently misclassified. Using these seeds, the model is modified to address a larger number of potential errors.

In information retrieval, Inverse Document Frequency [156] (idf) is often used along with term frequency (tf) as a measure of relevance of a term, capturing the fact that a term is frequent, but not so frequent to be essentially meaningless (non-meaning words, such as articles and conjunctions, are normally the most frequent ones). A common

way to address point (a) would therefore be to use a standard tf·idf metric. However, because our seeds list is guaranteed to only contain meaningful entries, we can safely select the terms occurring in more documents first (i.e., with an increasing idf). Point (b) is left as future work.

2.5.3 Classification of a new candidate

At the end of the training phase, the classifier M^{ctx} has been trained - on contexts of the selected seeds - to classify as either pertaining to substances or effects. Here, the procedure is described by which, given a new candidate c, we establish what class (Substance or Effect) it belongs to. The new candidates are chosen from the terms which are more relevant for the forums. Such terms are extracted according to the contrastive approach described in Section 2.5, subtask (a).

The training phase produces a model M^{ctx} by which contexts in which the term appears are classified – we define here a model M^{trm} by which the term itself is classified into either Substance, Effect, or "none-of-the-above". M^{trm} is defined as a function of a candidate c and the existing model M^{ctx} as follows:

1. Apply steps (iii) and (iv) of the algorithm described in 2.5.1 to obtain the contexts for c (ctx_c).

2. Classify the elements of ctx_c using M^{ctx}. We discard all categorisations whose confidence, according to the model, falls below a threshold θ_p, which we have experimentally set to 0.8 as a reference value. In the experiments the confidence is provided by the SVM classifier adopted for M^{ctx}.

3. Consider the remaining categorisations thus obtained. If a sizeable portion of them (θ_c, initially set to 0.6, we will show you how results vary along with its value) belongs to the same class C, then c belongs to C; otherwise it is left unassigned.

In Figure 2.11 you can catch a high level graphical description of this process.

2.5.4 Linking substances to effects

Here a simple procedure is outlined by which we can associate the substances mentioned in the drug forums to the effects they produce.

When indexing a post, the significant terminology elements found in the post are linked to it as metadata. As introduced, the terminology elements have been extracted following a contrastive approach, as in [143].

We assume to have already tagged the terminology elements found in each post as referring to substances or effects, using the method described in Section 2.5.3. Thus, when searching for mentions of a particular substance, we can correspondingly fetch, for each post the substance mention is found in, the relative metadata. Then, from the metadata, we can sort the list of effects by frequency – it is very likely that those effects are related to the searched substance. Despite consumers can perceive differently the effects of a substance, it is likely that there exists a set of effects that are expected from a substance or a category of substances. As a consequence, when a substance is classified by the system, we can associate the most frequent effects to it. Moreover, we expect to find similar effects for substances belonging to the same category.

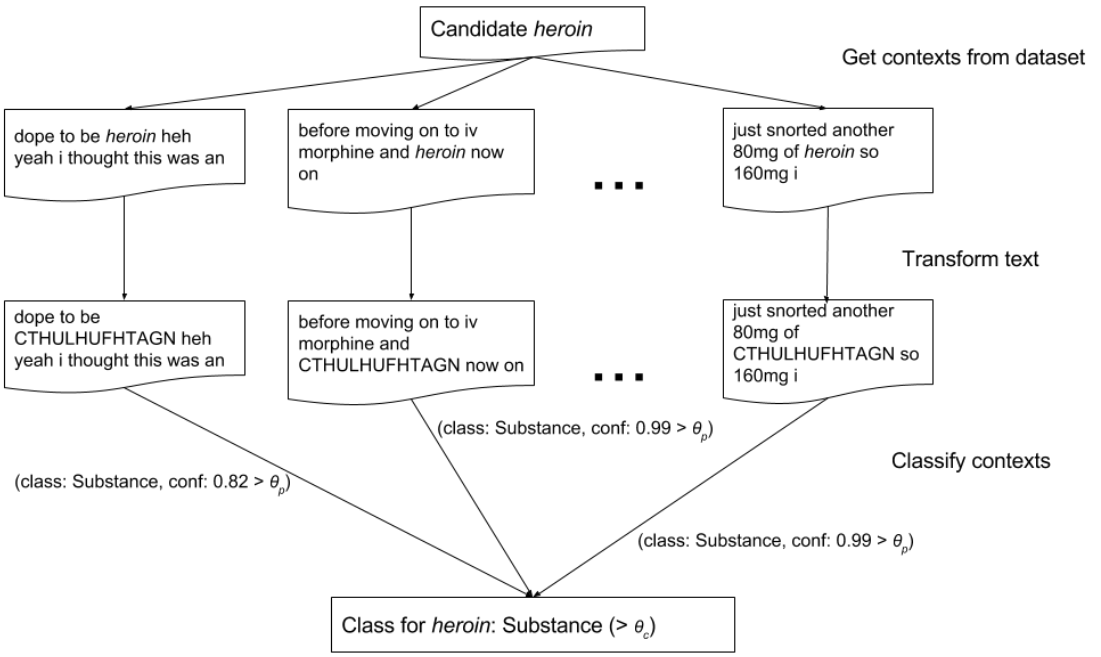


Figure 2.11: *Classification of a new term*

As a simple example, let's suppose to have a single post, with Text: *heroin gave me a terrible headache*; Substances: [heroin]; Effects: [headache].

Intuitively, we can assume that [headache] is an effect of [heroin]. If we consider all the posts in our datasets where the substance [heroin] is among the metadata, and we count the most frequent metadata effects associated to [heroin], we can have an indication of the links between substances and effects. However, many substances may appear in the same text. Thus, it is necessary to filter out the rarest substance-effect links since they are often due by chance. Those rare associations occur orders of magnitude less frequently than common ones. Section 2.6 will report on some findings we were able to achieve for our datasets about drugs and their effects.

2.6 Experiments

This section shows a set of experiments on the data described in Section 2.4.

First, from all the posts, we need to identify a list of candidates (unless we want to try and classify every term – a possible, but undesirable strategy), out of which to pinpoint substances or effects. Candidates are selected using a contrastive terminology extraction [143], to identify terms and phrases common within the community and yet specific to it; this is the first subtask outlined in section 2.5. Then, we apply the M^{trm} classifier, described in Section 2.5.3, to assign to candidates either the class Substance or Effect or none of the above, and evaluate the performance of the classification. The intermediate M^{ctx} classifier was trained using SVM with linear kernel [33].

Here are reported experiments and results for the Bluelight forum. The lists used to select seeds and to validate results have been described in Section 2.4.1. These lists represent 2 classes: Substance and Effect.

It is worth noting that, for the experiments, we consider the intersection between the lists of seeds and the extracted terminology. This is necessary because: i) items that are present in the lists may not be present in the downloaded dataset; ii) many terminological entries might be neither drug names nor drug effects. The intersection contains 226 substances and 89 effects. Some of these will be used as seeds, the rest of the entries to validate the results.

The results are given in terms of three standard metrics in text categorisation, based on true positives (TP - items classified in category C, actually belonging to C), false positives (FP - items classified in C, actually not belonging to C) and false negatives (FN - items not classified in C, actually belonging to C), computed over the decisions taken by the classifier: precision[11], recall[12] and F1-micro averaged[13].

The first results are in Table 2.5 and Figure 2.12. The outcome highlights the impact of the number of seeds selected on the classifier performance. Even though the training set is limited to a small number of entries, the results are interesting: with only 6 seeds, the proposed methodology achieves a F1 score close to 0.88 (on the 2 classes - Substance and Effect), while we left the other for validation.

With the aim of monitoring the diffusion of new substances, the result is quite promising, since it is able to detect unknown substances without (or limited) human supervision. The small seed number required to achieve good performance puts in evidence the capability to generalise of the algorithm.

# of seeds	Recall	Precision	F1
1	0.502	0.649	0.566
2	0.576	0.734	0.645
3	0.65	0.827	0.728
4	0.769	0.891	0.826
5	0.823	0.909	0.864
6	0.832	0.926	0.876

Table 2.5: *Classification results for substances and effects, varying the number of seeds*

Dealing with "the rest". Finding mentions of new substances or effects means classifying candidates terms in one class or the other. Playing with thresholds, we can discard some candidates, as belonging to none of the two classes (see Section 2.5.3).

Thus, within the extracted terminology, about 100 entries have been manually labelled as neither drugs nor effects, and we have used them as candidates. This has been done to evaluate the effectiveness of using the parameter θ_c to avoid classifying these terms as either substances or effects. Performance-wise, this resulted in a few more false positives given by terms erroneously assigned to the substance and effect classes, when instead these 100 candidates should ideally all be discarded. The results are in Table 2.6 and Figure 2.13. We can observe that, when we include in the evaluation also those data that are neither substances nor effects, with no training data other than the original seeds, and operating only on the thresholds, the precision drops significantly.

To achieve comparable performances, additional experiments have been conducted, changing the number of seeds and θ_c used to keep relevant terms. The results are

[11] $precision = \frac{TP}{TP+FP}$

[12] $recall = \frac{TP}{TP+FN}$

[13] harmonic mean of $precision$ and $recall$: $F1 = 2 \cdot \frac{precision \cdot recall}{precision+recall}$

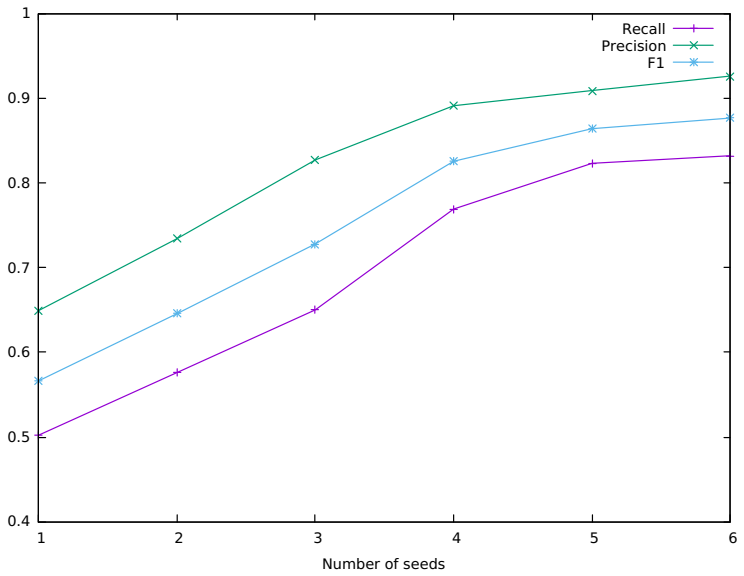


Figure 2.12: *Recall, precision and F1 varying the number of seeds*

# of seeds	Recall	Precision	F1
1	0.502	0.502	0.502
2	0.576	0.563	0.569
3	0.650	0.628	0.639
4	0.769	0.694	0.730
5	0.823	0.723	0.770
6	0.832	0.733	0.779

Table 2.6: *Classification results for substances and effects, including the "rest" category*

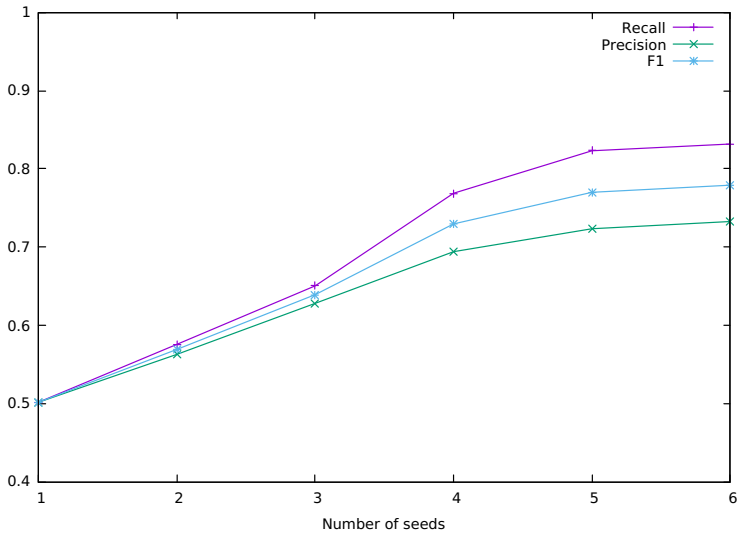


Figure 2.13: *Recall, precision and F1 including the "rest" category*

shown in Table 2.7 and Figure 2.14. The higher the threshold, the higher the precision, while increasing the number of seeds improves the recall, which is to be expected:

# of seeds	Recall 0.75	Precision 0.75	F1 0.75	Recall 0.8	Precision 0.8	F1 0.8
5	0.607	0.755	0.673	0.508	0.787	0.618
10	0.759	0.852	0.803	0.654	0.889	0.754
15	0.811	0.837	0.824	0.705	0.874	0.781
20	0.833	0.854	0.843	0.753	0.866	0.805

Table 2.7: *Precision, Recall and F1 with θ_c set to 0.75 and 0.8 (incl. "rest" category)*

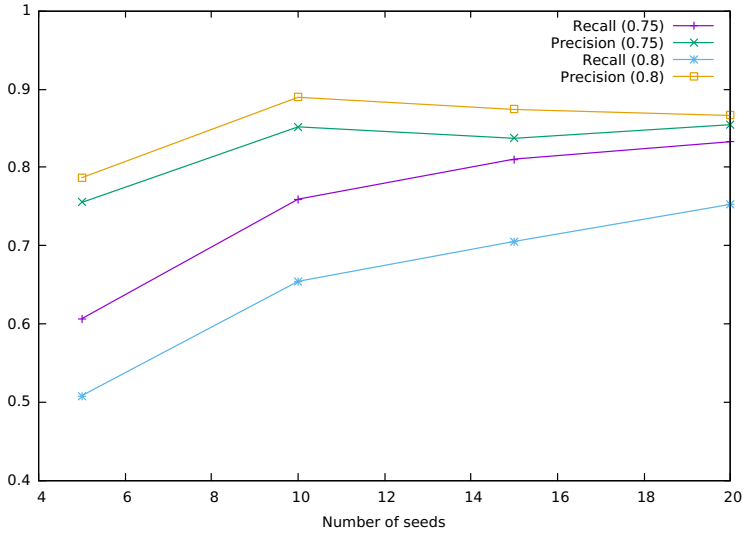


Figure 2.14: *Precision and Recall with θ_c set to 0.75 and 0.8 (incl. "rest" category)*

adding seeds "teaches" the system more about the variety of the data. Moreover, recall augments when we increase the number of contexts per seed used to train the system (Table 2.8 and Figure 2.15).

It is worth noting that increasing the number of contexts used to classify a new term seems to have no effect after a few contexts, as shown in Table 2.9 and Figure 2.16). This indirectly conveys information on the variety of contexts present on the investigated datasets.

Interestingly, the automated drug detection reported 1846 drugs in Bluelight and 1857 in DrugsForum, with 1520 drugs in common between the two forums. Moreover, some drugs appear exclusively in one of the two forums, like the *triptorelin*, *can-*

# of contexts	Recall	Precision	F1
100	0.437	0.709	0.541
1000	0.675	0.802	0.733
2000	0.742	0.830	0.784
3000	0.759	0.852	0.803
4000	0.769	0.838	0.802
5000	0.817	0.867	0.841
6000	0.831	0.851	0.840

Table 2.8: *Results varying the number of contexts per seed*

# of contexts	Recall	Precision	F1
10	0.763	0.758	0.760
50	0.746	0.815	0.779
100	0.759	0.852	0.803
150	0.763	0.852	0.805
200	0.753	0.854	0.800
300	0.759	0.852	0.803

Table 2.9: *Results varying the number of contexts per new term*

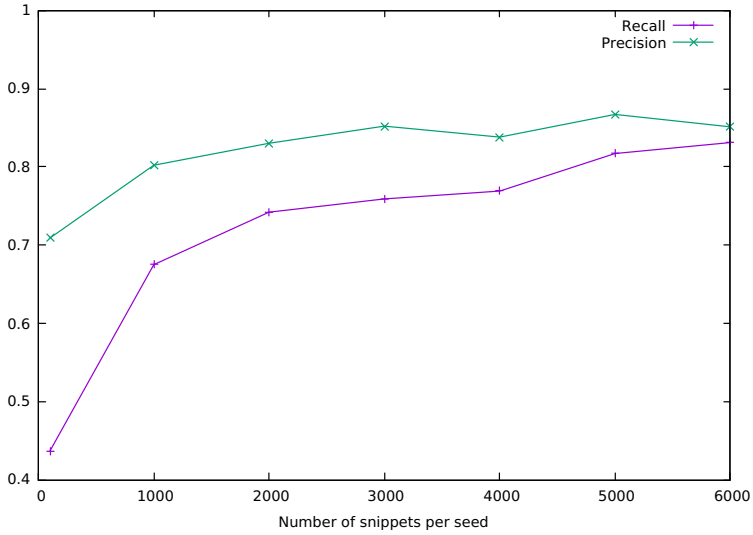


Figure 2.15: *Recall and precision varying the number of contexts (snippets) per seed, 10 seeds used*

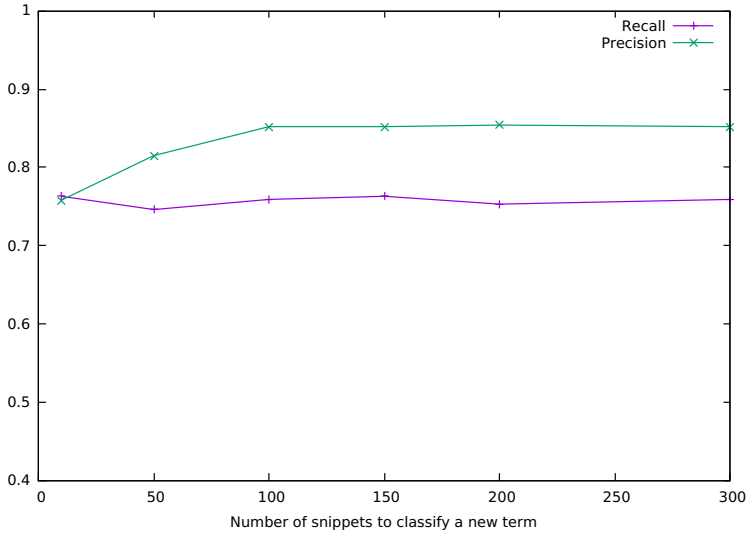


Figure 2.16: *Recall and precision varying the number of contexts (snippets) per new term, 10 seeds used*

Drug	Effects
heroin	anxiety, euphoria
cocaine	euphoria, anxiety, comedown, paranoia, psychosis
ketamine	euphoria, anxiety, visuals, comedown, hallucinations, nausea
methadone	anxiety, euphoria
codeine	euphoria, anxiety, nausea
morphine	euphoria, anxiety, analgesic, nausea
amphetamine	euphoria, anxiety, comedown, psychosis, visuals
oxycodone	euphoria, anxiety
methamphetamine	euphoria, anxiety, psychosis, comedown, paranoia
dopamine	euphoria, anxiety, comedown, psychosis

Table 2.10: *Main effects of the most discussed drugs on Bluelight*

desartan and *thiorphan* in Bluelight and the *lymecycline*, *boceprevir* and *imipenem* in Drugsforum, although the majority is shared.

Finally, upon training the system with the seeds, for every post it is possible to link the drugs to their effects. An example of links is in Table 2.10.

2.7 Discussion

Today, New Psychoactive Substances (NPS) lie on a grey area, not precisely addressed by current regulations. NPS rapidly appear on - and suddenly disappear from - the market, with a consistent and continuous introduction of new surrogates, which leaves little margin for intervention by healthcare institutions and governments. This Chapter has put in evidence some unique features of online NPS forums and shops. Monitoring such websites and elaborating the available data made it possible to explore a large quantity of information, also across platforms, allowing analysts to perform comparisons among them. We also gave a measurement of the relevance of NPS diffusion and advertisement, as well as user engagement. Furthermore, we showed how trading and discussions are correlated, through terms used by both online shops, SM, and forums, despite the prohibition, which holds on forums, to post explicit links to shops. Noticeably, co-occurrence analysis and temporal analysis of neologisms are a valid support for NPS detection.

Currently, the analyses are led by the data scientist, who is assisted by the developed software. The analyses are applicable both to offline datasets and online streaming sources. The DAGON methodology proposed here supports researchers in the automatic identification and classification of substances and effects from posts of drugs forums, making use of a semi-supervised text mining approach. Human intervention is required for the creation of a small training set, but the algorithm is able to automatically discover substances and effects with very few initial information. We believe our proposal will help to sensitize drug consumers about the risks of their choices and will contrast the diffusion of NPS, which spread on the online market at an impressively high rate. Furthermore, shortening the latency of the drug detection since its introduction helps in limiting its diffusion along the streets and online shops. Moreover, we expect to improve the recognition of novel drugs by exploiting the effects that are commonly associated to them.

The possibility to rely on publicly available data shared on SM platforms, allows

intelligent systems to extract knowledge that is increasingly important to automate the process of drug detection and classification. The openness of data allows intelligent systems to aggregate and distill information to produce knowledge that is essential for monitoring purposes and intelligence, but SM policies may change during time, also affected by business and legislation with important implications for applications development and effectiveness.

CHAPTER 3

Social Media for disaster management: a support to citizens during mass emergencies

3.1 Social Media as source of data

Most of the work presented here is connected to the application of SM data to crisis management situations as presented in [6, 7, 10]. With this Chapter we want to show how SM can be effectively mined to support the population in critical situations and help emergency responders to react quickly and help save as many lives as possible.

Nowadays SM represent a powerful way to investigate preferences, tastes and activities of groups of users. Now more than ever, people continuously share comments and multimedia content about their lives, interests, feelings and opinions [194]. To this regard, platforms such as Twitter, Weibo and Instagram are privileged channels of information diffusion because of their large user base, interactive nature and ease of use while on the move. Furthermore, SM encourage citizens to participate in the process of citizen journalism, which has been proven to be much faster than traditional media in spreading news [165]. SM users can thus be considered as sensors able to convey valuable information about situations and facts, as asserted by the social sensing (or Human as a Sensor) paradigm [169].

The amount of information shared on SM in the aftermath of mass convergence or emergency events is even bigger, showing bursts of messages describing the unfolding scenario, often complemented with images or videos [53]. Within this context, SM data revealed to be particularly valuable in the aftermath of those events, typically natural and man-made disasters, which trigger massive participation of affected communities in sharing time-sensitive and actionable information [4, 73]. Both types of disasters require a timely intervention by emergency responders, who are in charge of providing

support and relief to the affected population. It is therefore possible to exploit social sensors either to augment conventional emergency detection and monitoring systems that rely on pre-deployed ad-hoc sensing equipment, or to substitute them in the lack of such equipment [172]. However, the sheer SM activity in the aftermath of emergencies renders manual analyses of data unfeasible. Also, SM data is often noisy and bursty, with text often fragmented and unstructured [49]. Thus, automated analysis of such content asks for preprocessing steps before data can be effectively used. A data-filtering step is usually advisable and machine-learning algorithms can benefit from such operation allowing to achieve overall better performances.
ICTs now enable a new class of decision support systems and tools that aim at improving the capabilities of specialists in detecting and preparing a prompt and effective response to crises. Effectiveness of intervention is closely related to the knowledge of the event's intensity (i.e., its effects on people and infrastructures), whose evaluation can highly benefit from machine-enabled mining of the large volumes of data coming from social sensors.

In recent years, applications based on user-generated information for disaster management [9] were largely adopted either from Academia or practitioners [181]. In many practical situations, the scarcity of key resources – temporal, economic, and human resources above all – imposes dire limitations to the extent and the effectiveness of the emergency management process. For this reason, tools capable of supporting resource allocation and prioritisation can have a significant impact towards the effectiveness of emergency management operations. Among these tools there are the SM-based crisis mapping systems, which increase situational awareness by enabling the real-time gathering and visualisation of data contributed by many SM users. Such a type of system is a platform able to collect text and multimedia content from a variety of sources, such as Twitter and Facebook, to analyse and aggregate collected data, and to visualise relevant facts on a map. Notably, during many recent disasters, civil protection agencies developed and maintained live Web-based crisis maps to help visualise and track stricken locations, assess damage, and coordinate rescue efforts [130]. As a matter of fact, so far, civil protection agencies do not fully exploit tools based on SM contents and still heavily rely only on traditional technologies (e.g., seismographic networks, on-the-ground surveys, aerial or satellite images) in the most intense phases of emergency response [6].

Indeed, recent work demonstrated the possibility to create crisis maps solely using geolocated data from SM, to better understand and monitor the unfolding consequences of disasters [6,79,130]. All these SM-based crisis mapping systems face the fundamental challenge of *geoparsing* the textual content of emergency reports in order to extract mentions of places/locations, thus increasing the number of messages to exploit. Geoparsing involves binding a textual document to a likely geographic location which is mentioned in the document itself. State-of-the-art systems, such as [130], perform the geoparsing task by resorting to a number of preloaded geographic resources containing all the possible matches between a set of place names (toponyms) and their geographic coordinates. This approach requires an offline phase where the system is specifically set to work in a geographically-limited region. Indeed, it would be practically unfeasible to load associations between toponyms and coordinates for a wide region or for a whole country. Moreover, not all geolocated data is useful towards understanding the

severity of the emergency and, indeed, only a small fraction of messages convey information about the consequences of the emergency on communities and infrastructures. Current crisis mapping systems typically detect the most stricken areas by considering the number of messages shared and by following the assumption that more emergency reports equals to more damage [130, 184]. Although this relation exists when considering densely and uniformly populated areas [120], it becomes gradually weaker when considering wider regions or rural areas. These challenges, related to the detection of damage and to geoparsing, are reflected by the current limitations of state-of-the-art SM-based crisis mapping systems [6].

3.2 Emergency management in literature

The possibility to exploit SM data for crisis mapping was first envisioned in a few trailblazing works [73, 79, 128] and further backed up by recent research [4]. Since the early works, there has been a growing interest by both practitioners and scholars in all areas related to crisis mapping: from data acquisition and management, to analysis and visualization [6].

3.2.1 Practical experiences

The great interest of practitioners and emergency responders towards the exploitation of SM data is testified by the efforts of the Federal Emergency Management Agency (FEMA) and the United States Geological Survey (USGS), which already led to interesting results with practical applications to earthquake emergency management[1] [26, 64]. Regarding already deployed applications, well-known crisis mapping platforms are Ushahidi[2], Mapbox[3], Google's Crisis Map[4], ESRI ArcGIS[5], and CrisisCommons[6] [17]. The main features of these platforms are related to data acquisition, data fusion, and data visualization. Such platforms represent hybrid crowdsensing systems where users can voluntarily load data onto the system in a participatory way, or the system can be configured so as to automatically perform data acquisition in an opportunistic way. The same hybrid data collection strategy has also been employed in a fully automatic system recently benchmarked in the earthquake emergency management field [8]. Another already-deployed application that exploits crowdsourced data is USGS's "Did You Feel It?" (DYFI) system[7]. This system, although not relying on SM data, exploits citizen reports and responses to earthquakes in order to automatically assess potential damage. It is foreseeable that in the near future such system could instead be fed with SM data. Indeed, there is already interesting research – from both USGS itself and from other laboratories – moving towards this direction [47, 82, 109, 110].

[1] https://blog.twitter.com/2014/using-twitter-to-measure-earthquake-impact-in-almost-real-time
[2] https://www.ushahidi.com/
[3] https://www.mapbox.com/
[4] https://www.google.org/crisismap/
[5] http://www.esri.com/arcgis/
[6] https://crisiscommons.org/
[7] http://earthquake.usgs.gov/research/dyfi/

3.2.2 Academic works

Recent scientific literature has instead switched the focus from data acquisition and data fusion to in-depth data analysis. This is typically done by leveraging powerful machine learning techniques, and resulted in novel solutions being proposed to overcome critical crisis mapping challenges such as geoparsing and extracting situational awareness from microtexts [49].

Specifically, [130] presents a state-of-the-art system that matches preloaded location data for areas at risk to geoparse real-time tweet data streams. The system has been tested with data collected in the aftermath of New York's flooding (US – 2012) and Oklahoma's tornado (US – 2013) and achieved promising results. Among the key features of [130] is the possibility to match toponyms at region-, street-, or place-level. This is achieved by preloading already existing geographic databases (e.g.: the Geonames and GEOnet Names global gazetteers) for areas at risk, into the system. Crisis maps are then generated by comparing the volume of tweets that mention specific locations with a statistical baseline. Although presenting state-of-the-art solutions, [130] still has several drawbacks. The system can only work on a specific geographical area at a time since it has to load and manage external data for that area. Tweets mentioning locations outside the predefined area cannot be geolocated and consequently disasters cannot be monitored outside the area's boundaries. Moreover, the width of the area covered by the system has direct implications on the amount of data to load and manage. This impacts on system's performances thus resulting in limitations on the maximum geographical area that can be monitored with [130]. Furthermore, the system in [130] does not take into account the problem of toponymic polysemy [6,49]. In addition, crisis maps generated by [130] only consider tweet volumes and may result less accurate than those obtained by analyzing the content of tweets. For example in the case of severe earthquakes, where the shaking is perceived also hundreds of kilometers far from the epicenter, the majority of tweets come from densely populated areas, such as big cities. Anyway, locations that have suffered most of the damage might be small villages in rural areas around the epicenter, which risk remaining unnoticed if the analysis only considers tweet volumes [6, 49].

In addition to the fully functional crisis mapping system described above, other solutions for the geoparsing task have been recently proposed in [57,74,75] where authors experimented with heuristics, open-source named entity recognition software and machine learning techniques. Furthermore, other works emphasized the extraction of actionable and time-sensitive information from messages. For instance, authors of [177] apply natural language processing techniques to detect messages carrying relevant information for situational awareness during emergencies. In [98] a technique to extract "information nuggets" from tweets is described – that is, self-contained information items relevant to disaster response. While these works present fully automatic means to extract knowledge from texts, in [178] a hybrid approach exploiting both human and machine computation to classify messages is proposed. All these linguistic analysis techniques for the extraction of relevant information from disaster-related messages have however never been employed in a crisis mapping system.

Finally, a survey presented an extensive review of current literature in the broad field of SM emergency management, and can be considered for additional references [96].

Table 3.1 compares the most famous applications and approaches in the emergency

System	Approach	SN based	Real-Time	Geolocation
Ushaidi	Participatory	Partially	✓	GPS
MapBox	Participatory	✗	✗	GPS
Google Crisis Map	Opportunistic	✗	✗	GPS
ESRI ArcGIS	Participatory	✗	✓	GPS
DYFI	Participatory	✗	✗	GPS
EARS [11]	Opportunistic	✓	✓	GPS
Middleton [130]	Opportunistic	✓	✓	Geoparsing + global gazetteers
Crismap [10]	Opportunistic	✓	✓	Geoparsing + knowledge-base

Table 3.1: *Comparison between the most relevant crisis mapping systems. The majority of them is based on GPS data or user input about the event area.*

management and crisis mapping field and describes their main features.

3.3 Datasets

The datasets used for this work are composed of Italian tweets, collected in the aftermath of 5 major natural disasters. For our experiments we considered different kinds of disasters, both recent and historical: 3 earthquakes, a flood, and a power outage. Specifically, the `L'Aquila` and the `Emilia` datasets are related to severe earthquakes that struck rural areas of Italy in 2009[8] and 2012 respectively[9]. The `Amatrice` dataset is related to a recent earthquake that struck central Italy in 2016[10]. The `Sardinia` dataset has been collected in the aftermath of a flash flood occurred in the Sardinia island in 2013[11]. Finally, the `Milan` dataset describes a power outage which occurred in the metropolitan city of Milan (northern Italy) in 2013. To investigate a wide range of situations we picked disasters having variable degrees of severity: some caused only moderate damage, while others produced widespread damage and casualties.

The datasets were created by using the Twitter's Streaming API[12] for recent disasters, and the Twitter resellers' Historical API[13] (GNIP) for past disasters. The API give access to a global set of tweets, optionally filtered by search keywords. We exploited a different set of search keywords for every different disaster in order to collect the most relevant tweets about it. Whenever possible, we resorted to hashtags specifically created to share reports of a particular disaster, such as the *#allertameteoSAR* hashtag for the `Sardinia` dataset. In this way, we were able to select only tweets actually related to that disaster. However, for historical disasters we couldn't rely on specific hashtags and had to exploit generic search keywords already proposed in literature, see [5, 11, 154]. This is the case of the `L'Aquila` dataset, for which we exploited the "terremoto" (*earthquake*) and "scossa" (*tremor*) Italian keywords. In addition, we only used "fresh" data shared in the aftermath of the disasters under investigation. For instance, all the 3,170 tweets in the `Emilia` dataset were posted in less than 24 hours from when the earthquake occurred.

[8] https://en.wikipedia.org/wiki/2009_L'Aquila_earthquake
[9] https://en.wikipedia.org/wiki/2012_Northern_Italy_earthquakes
[10] https://en.wikipedia.org/wiki/August_2016_Central_Italy_earthquake
[11] https://en.wikipedia.org/wiki/2013_Sardinia_floods
[12] https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html
[13] http://gnip.com/sources/twitter/historical

Tweets in the L'Aquila, Emilia, and Sardinia datasets have been manually annotated for mentions of damage according to the 3 following classes: (i) tweets related to the disaster and carrying information about damage to infrastructures/communities (*damage*); (ii) tweets related to the disaster but not carrying relevant information for the assessment of damage (*no damage*); (iii) tweets not related to the disaster (*not relevant*). The inclusion of a class for tweets that are not related to a disaster (*not relevant*) is necessary because the automatic data collection strategy we adopted does not guarantee that all the tweets collected are actually related to the disaster under investigation. This is especially true for the datasets collected with generic search keywords and represents a further challenge for our classification task. The manual annotation of damage mentions among tweets is exploited to train and validate our damage detection classifier, as thoroughly explained in Section 3.7. Furthermore, following the same approach adopted in [74, 130], we carried out an additional manual annotation of 1,900 random tweets of the aforementioned datasets with regard to mentions of places/locations. This further annotation is exploited to validate our geoparsing results, as described in Section 3.8. The Milan dataset is also used for a comparison of geoparsing techniques in Section 3.8, since it was already exploited in previous work [130]. The Emilia and Sardinia datasets are also used in Section 3.9 to quantitatively validate our crisis maps against authoritative data. Finally, the Amatrice dataset is exploited in Section 3.9 as a case study of our system in the aftermath of the recent earthquake in central Italy.

Notably, the total number of 15,825 tweets in our datasets, shown in Table 3.2 along with other details, is greater than those used in other related works, such as [130] (6,392 tweets across 4 datasets), and [75] (2,000 tweets for a single dataset).

				tweets					used in
dataset	**type**	**year**	**users**	*damage*	*no damage*	*not relevant*	*GPS*	*total*	**sections**
L'Aquila	Earthquake	2009	563	312 (29.4%)	480 (45.2%)	270 (25.4%)	0 (0%)	1,062	3.7, 3.8, 3.9
Emilia	Earthquake	2012	2,761	507 (16.0%)	2,141 (67.5%)	522 (16.5%)	205 (6.5%)	3,170	3.7, 3.8, 3.9
Milan	Power outage	2013	163	-	-	-	15 (3.8%)	391	3.8
Sardinia	Flood	2013	597	717 (73.5%)	194 (19.9%)	65 (6.6%)	51 (5.2%)	976	3.7, 3.8, 3.9
Amatrice	Earthquake	2016	7,079	-	-	-	21 (0.2%)	10,226	3.9

Table 3.2: *Characteristics of the Datasets.*

To better understand the importance of geoparsing in a crisis mapping task, in Table 3.2 we also reported the number of tweets natively geolocated (GPS column). Geolocation of these tweets is performed directly by Twitter whenever a user enables GPS or WiFi geolocation. Statistics on our datasets confirm previous findings reporting that only a small percentage ($1\% \div 4\%$) of all tweets are natively geolocated [38, 49]. As introduced in Section 3.1, the low number of natively geolocated tweets drastically impairs crisis mapping, hence the need of a geoparsing operation. Noticeably, none of the 1,062 tweets of the L'Aquila dataset, dating back to 2009, are natively geolocated.

3.4 System

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis." [113]. With this

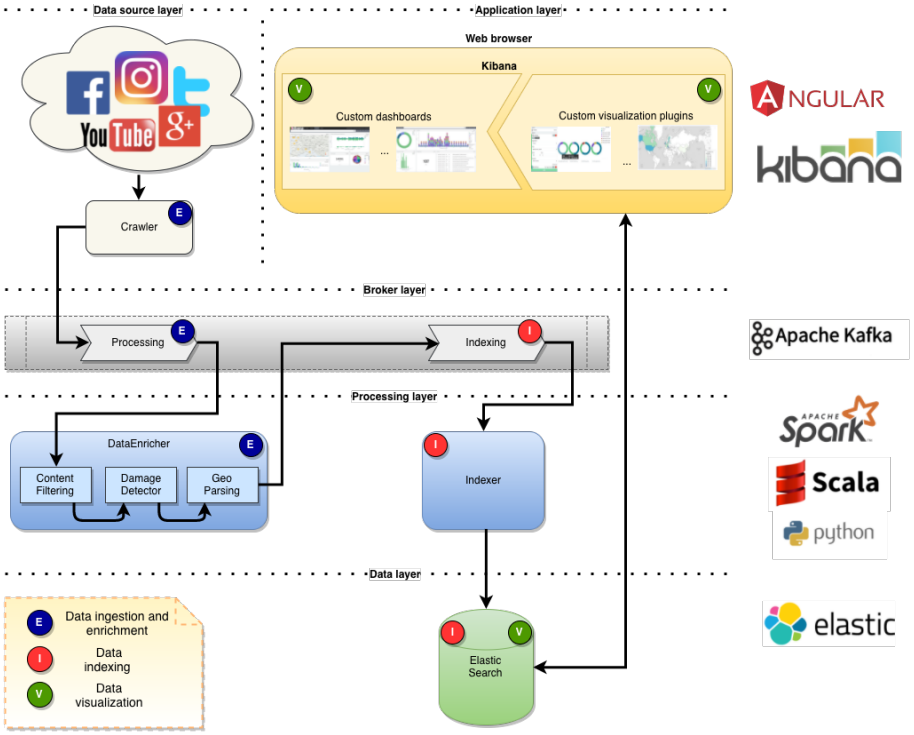


Figure 3.1: *The logical architecture of our crisis mapping system. On the left-hand side, the architecture is organised in a stack of layers: from the topmost data source and application layers of our system, to the broker, processing, and data layers. On the right-hand side, for each layer the technologies adopted to design and implement the components of the system are reported.*

in mind, SM are an incredibly rich source of big data and require that analysts adopt a proper set of tools to collect, process, store and analyse such an astonishing quantity of data. Big data exhibit some characteristics marked as 3Vs [112] (Volume, Velocity, Variety), introduced by Gartner report[14] for data in 2001 and then become a standard in Information Technology.

SM data collection imposes rigid constraints and performance requirements either for crawlers and scrapers or other technologies and stresses the scalability aspects of software algorithms and permanent storage solutions. Technology dealing with SM must take into account either the sustained arrival rate of information as well as the requirements imposed by real-time data processing and visualisation. Foremost, the proposed solution considers scalability aspects and the searchability of information. Without these two constraints it is not possible to perform massive data processing nor is it conceivable to get results with low latency.

The software architecture adopted for SM crawling, analysis and mapping is presented in Figure 3.1. As with any system that needs to cope with the massive amount of data collected from SN, special requirements are imposed in the design by both the real-time constraints and the heterogeneity of data. For these reasons, our `CrisMap` system deploys several Big Data technologies to process incoming SM data efficiently, without sacrificing scalability and fault-tolerance.

[14] `https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf`

The functional workflow is divided into three logical steps: (i) Data Ingestion and Enrichment (labelled with a blue circle), where SM data is collected and processed in order to select and geoparse messages containing information about damages; (ii) Data Indexing (labelled with a red circle), in which useful data is conditioned and saved into the internal storage; (iii) Data Visualization (labelled with a green circle), in which stored data is retrieved and used to create maps or, more generally, to provide results to the end users through a Web dashboard.

Following sections will give an overall description of the functionality of each of these steps. A detailed technical description and evaluation of the solutions adopted and implemented to address the main issues related to the design of a crisis mapping system, namely *mining messages to search for damage*, *geoparsing* and *visualise data*, are given in Sections 3.7, 3.8 and 3.9, respectively.

3.4.1 Data Ingestion and Enrichment

The first logical step of the system consists in ingesting data coming from available SM data sources, possibly enriching it with additional information not directly available from the data source and which can provide useful information exploitable subsequently during the visualisation phase. Data ingestion occurs using platform specific crawling/scraping software. For the sake of simplicity, in this section the attention is focused solely on Twitter. However, nothing prevents the adopter to deploy the system using a different data source.

The system is able to process both real-time data fetched from the Twitter stream and historical data acquired from data resellers, as well as Facebook posts and Instagram media. In a practical application scenario, the system is fed with real-time data, while historical data can be used to run simulations on past events.

Typically crawlers make use of SM API or adopt some scraping technique to collect data. In case of Twitter, for example, the *Crawler* component exploits Twitter's Streaming API[15] to perform data acquisition from the SN. The Streaming API gives low latency access to Twitter's global stream of messages. Collected messages can be optionally filtered by search keywords. Other data sources may not offer stream API and thus it is necessary to implement alternative policies, such as polling.

Collected messages are then forwarded toward the Processing layer through a specific queue (named "Processing") placed at the Broker layer. The implementation choice fell on Kafka[16], a distributed publish-subscribe messaging platform. As described on the website, Kafka "is used for building real-time data pipelines and streaming apps. It is horizontally scalable, fault-tolerant, wicked fast, and runs in production in thousands of companies". Its characteristics help the system to sustain very high traffic rates by mitigating the back-pressure data problem [171] and to process data resiliently, with no loss in case of system failures (e.g. hardware faults).

SM messages are fetched from the *Processing* queue and processed by an enrichment component called *DataEnricher*. As depicted in the figure, the *DataEnricher* is organised internally into a pipeline of three sub-modules that process, filter, and enrich incoming data. The *ContentFiltering* sub-module analyses the tweets in order

[15] `https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html`

[16] `https://kafka.apache.org`

to select those that are relevant to the problem (e.g., whose contents relate to natural disasters). Relevant messages are then analysed by the *DamageDetector* sub-module, which in turn discriminates between data carrying or not carrying damage information. The messages containing damage information are finally forwarded to the *GeoParsing* sub-module, which possibly enriches each tweet with information about the discovered geolocation. Upon completion of the pipeline, the enriched SM messages are pushed into another queue, the *Indexing* queue, waiting to be indexed and stored into the system. The *DataEnricher* is implemented using Spark[17] and uses a streaming approach to quickly process incoming data from the Processing queue.

3.4.2 Data Indexing

In the second step, enriched data is moved from the *DataEnricher* to a permanent storage through the *Indexing* queue. Data fetched from the queue is instantaneously parsed and pushed into a search and analytics engine. The choice of Elasticsearch[18] (ES) was driven by its capability to scale horizontally, as well as to perform fast search and aggregation on textual data leveraging its internal Apache Lucene[19] engine. In fact, the integration of ES with other software like Kibana and Logstash makes it a valid solution as storage system.

The role of the *Indexer* module is to map the data structure of a Twitter message into an optimised ES index, where each field type is treated differently to exploit the features of the engine to provide fast search and real-time analytics on stored data (e.g., ES tokenizes textual fields and provides an inverted terms list to efficiently search for in the related fields). Data stored in ES is available for queries in near-real time, introducing a latency of just 1 second before the indexed data can be retrieved by users. Indeed, the query latency is negligible in our context, having no practical consequences for the purposes of crisis mapping.

3.4.3 Data Visualization

In the third step, data stored in the index can be browsed and queried through the Kibana[20] software. With Kibana, it is possible to build real-time visualizations with very useful insights, like real-time volumes, maps, bar charts and word clouds. Those visualizations can be also joined to build complex dashboards that allow the user to track changes over time for the most important metrics of the dataset. Moreover, Kibana supports the possibility to easily extend the internal visualization types in order to customize graphical views using the most appropriate visual analytics for data. To our purposes, the native visualization maps have been replaced with a specific plugin. The new plugin[21] supports multi-resolution choropleth maps and the possibility to normalize data according to population, offering different options for what concerns the scales and their customization.

[17] `http://spark.apache.org`
[18] `https://www.elastic.co/products/elasticsearch`
[19] `https://lucene.apache.org/`
[20] `https://www.elastic.co/products/kibana`
[21] The plugin is publicly available at **`https://github.com/marghe943/kibanaChoroplethMap.git`** .

Figure 3.2: *Twitter examples of noise in collected messages*

3.5 Message Filtering

Using search keywords to query SM platforms allows the acquisition of messages potentially related to an emergency. However, not all the messages captured are actually related to an outbreaking emergency. A message selection procedure only based on the presence of certain keywords in the text is insufficient to ensure that the message is relevant [11]. As previously introduced, some messages can be misleading for the emergency detection and monitoring tasks and must be filtered out as noise. By noise we refer to the messages containing the search keywords but which are not related to the type of emergency under investigation. In [5] we identified two different sources of noise: (i) messages in which the keyword is used with a different meaning than the one related to the searched emergency event and (ii) messages in which the keyword refers to a past emergency. We collected the tweets shown in Figure 3.2 while looking for earthquake-related messages, and we report them as examples of these two kinds of noisy messages. Excessive levels of noise in collected messages lead to false detection by the system. However, too much filtering may result in the loss of useful messages and thus in the impossibility to detect important events. Therefore this task presents another crucial trade-off related to the accuracy of the filtering process. To overcome this trade-off we propose a solution employing data mining techniques to train a machine-learning classifier. The classifier exploits the characteristics of tweets to discriminate between relevant messages and noisy messages to the task of emergency detection. Specifically, during the offline training phase the classifier is trained using two distinct sets of messages: those relevant and those not relevant (i.e. noisy messages) for an outbreaking emergency event. Messages of the training set must be manually annotated to ensure the correctness of the training examples. In the online mode of operation, the trained classifier is able to predict the class (relevant or noisy) of any new message thus implementing the filtering functionality. The classifier bases its decision for the class to

	Predicted Class	
Actual Class	Relevant	Non Relevant
Relevant	1197	501
Non Relevant	400	3371

Table 3.3: *Confusion matrix for English tweets*

assign to a message on a series of features. Please note that this distinction is necessary to quickly detect crisis in real-time fashion. We demand to Section 3.7 a broad discussion on how to filter content to detect damage in SM messages for mapping purposes. Our analysis on the messages reporting actual emergencies has highlighted a few interesting characteristics that help distinguish between relevant messages and noisy messages. Relevant messages sent by eyewitnesses are generally very short, they present few punctuation and often contain slang or offensive words. This is due to the fact that social sensors reporting an emergency are usually scared and the contents of their messages tend to represent this emotional state [5].

3.5.1 Experiments

We set up two different experiments on English and Italian tweets in order to assess the performances of the proposed filtering approach. The training set for the English language is composed of more than 5000 manually annotated tweets containing at least one occurrence of the "quake" or "shaking" (sub)strings. Instead, for the Italian language we collected more than 1400 tweets matching the "terremoto" (earthquake) or "scossa" (tremor) (sub)strings. We developed an ad-hoc Web annotation tool specifically designed for the annotation of tweet-based training sets. For the manual annotation we employed 3 human operators and each of them annotated both datasets. We included in the final training sets only those tweets which received the same annotation by all 3 annotators. We also designed 24 distinct features based on the results of our previous analyses about tweets' characteristics. Such features take into account many structural characteristics of tweets like words count, the presence of mentions, 'RT' string in case of retweets, urls, punctuation, uppercase letters and slang/offensive words. We ran feature selection and ranking algorithms to only include in our classifiers the most influential features. Algorithms employed for the feature selection are *Information Gain* and *Pearson's Correlation Coefficient*. As a result 9 features were selected for the English language and 7 features were selected for the Italian language. Both the English and Italian classifiers were generated with the Weka framework [84] using the decision tree J48, corresponding to the Java implementation of the C4.5 algorithm [147] with a 10-fold cross validation. We measured classifiers' performances by means of standard evaluation metrics such as True Positives (TP) count, True Negatives (TN) count, False Positives (FP) count and False Negatives (FN) count. Classification results are reported in Tables 3.3 and 3.4, presenting the so-called confusion matrices. Greyed-out cells in the confusion matrices highlight the numbers of correct classifications (TP and TN), while the other cells represent the numbers of incorrect classifications (FP and FN). Overall filtering results show an Accuracy of 83.5% (4568 correctly classified tweets out of 5469 total tweets) for English tweets and an Accuracy of 90.1% (1272 correctly

	Predicted Class	
Actual Class	Relevant	Non Relevant
Relevant	654	52
Non Relevant	88	618

Table 3.4: *Confusion matrix for Italian tweets*

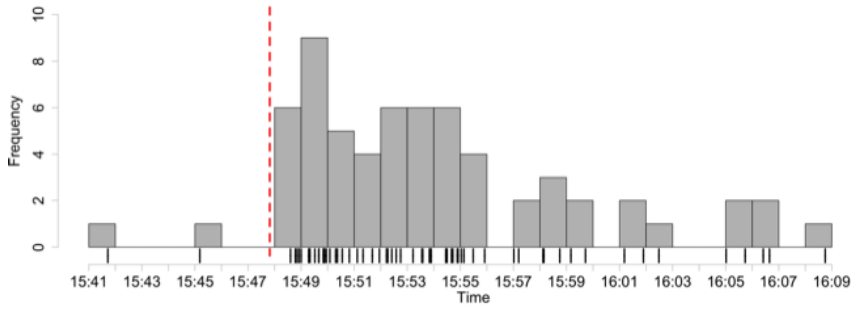


Figure 3.3: *A burst of tweets reporting a moderate earthquake in Italy*

classified tweets out of 1412 total tweets) for Italian tweets.

3.6 Emergency Detection

The detection of an emergency is triggered by an exceptional growth in the number of relevant messages captured by the system. The better the filtering phase, the easier the task of emergency detection. Event detection in SM is a topic that has been widely studied for a broad variety of purposes. Among commonly adopted event detection techniques are Bayesian statistics [155] and peak detection algorithms [64]. In our system we adopt an approach based on a burst detection algorithm. A burst is defined as a large number of occurrences of a phenomenon within a short time window [65]. Figure 3.3 displays a rug plot of the arrival times of earthquake-related relevant tweets in Italian language, as well as a histogram plot showing their frequency per minute, during a 3.4 magnitude earthquake occurred at 15:47:49, August 9 2014, in Tuscany regional district (Italy). After the occurrence time of the earthquake, indicated by the red vertical dashed line, a big burst of tweets was recorded. These bursts are caused by the large number of messages shared on SM by the people who actually felt the shaking. This "bursty" behavior is not constrained only to earthquakes, but the same applies to many other kinds of emergency situations, possibly serving as a red flag for the occurrence of an emergency.

3.6.1 Experiments

Building on the "bursty" characteristics of emergency reports, we adapted the burst detection algorithm originally proposed in [65] and we applied it to the detection of earthquakes in Italy. The detection is performed solely from Twitter data. The detec-

tion of a burst is based on the current frequency of relevant messages recorded during a short-term sliding time window. A burst is detected, and consequently the detection of an emergency is triggered, when such frequency exceeds a given threshold. The threshold to trigger a burst depends on a reference frequency calculated over a long-term sliding time window. In our experiments we tried different combinations of settings and we achieved the best detection results with the following settings:

- short-term sliding time window: 1 minute
- long-term sliding time window: 1 week

The tuning phase of the algorithm led us to set the threshold for the current frequency of relevant messages as ten times the *reference frequency*. This technique is exploited to measure how large the current instantaneous message arrival rate (computed every minute) is with respect to the average arrival rate (computed over a one week window). This threshold is an optimal value for earthquake related messages and can adaptively remove noise from SM. We tested our earthquake detection procedure with a dataset of Italian tweets collected over 70 days from 2013-07-19 to 2013-09-23. We computed the thresholds over the datasets and tested their effectiveness in contrasting the SM flooding of non relevant information about earthquakes. Over this testing period we were able to detect 47 earthquakes with this procedure. To validate our detections we exploited authoritative earthquake reports released by the Italian National Institute of Geophysics and Volcanology (INGV), which is the government agency responsible for monitoring seismic events in Italy. We therefore cross-checked all our SM-based detections against those obtained by INGV with their seismic network. We classified our earthquake detection results as in the following:

- True Positives (TP): earthquakes detected by our procedure and confirmed by INGV
- False Positives (FP): earthquakes detected by our procedure, but not confirmed by INGV
- False Negatives (FN): earthquakes detected by INGV, but not by our procedure

True Negatives (TN) are meaningless, as it would mean counting the number of earthquakes that did not happen and that our procedure did not detect. In addition we also computed the following evaluation metrics:

- Precision: ratio of correctly detected earthquakes among the total number of detected earthquakes
- Recall: ratio of correctly detected earthquakes among the total number of occurred earthquakes
- F-Measure: harmonic mean of Precision and Recall

Table 3.5 shows the final results of SM-based earthquake detections. Results show that the detection of earthquakes with a magnitude < 3 is very difficult (F-Measure $< 50\%$). This is because the majority of these earthquakes are only detected by equipment and not by people. For events with a magnitude equal to or greater than 3.5, results show

Magnitude	Earthquakes	TP	FP	FN	Precision	Recall	F-Measure
$> \mathbf{2.0}$	404	17	30	387	36.17%	4.21%	7.54%
$> \mathbf{2.5}$	102	16	30	86	34.78%	15.79%	21.62%
$> \mathbf{3.0}$	26	13	17	13	44.33%	50.00%	46.43%
$> \mathbf{3.5}$	11	9	3	2	75.00%	81.82%	78.26%
$> \mathbf{4.0}$	7	5	0	2	100.00%	71.43%	83.33%
$> \mathbf{4.5}$	2	2	0	0	100.00%	100.00%	100.00%

Table 3.5: *SM-based earthquake detection validation*

an overall good performance of the system (F-Measure $> 75\%$) which seems to suggest the effectiveness of the proposed solution. This is significant given that seismic events of a magnitude around 3 are considered "light" earthquakes and are generally perceived only by a very small number of social sensors. These results are promising, especially considering that the proposed technique is adaptable to other emergency scenarios (flash floods, wildfires, riots, etc.) where automatic detection equipment, playing the role of seismographs for seismic events, might not be available. Furthermore, the textual content of tweets often conveys many other kinds of information, such as the presence/lack of damage in a specific location [49,51]. Mining such content can indeed provide a deep insight into the evolving scenario.

3.7 Mining text to search for damage

The detection of damage in SM messages is a challenging task due to the almost completely unstructured nature of the data to be analyzed [51]. The *DataEnricher* component of Figure 3.1 analyzes the content of tweets with the twofold goal of discarding irrelevant tweets and labeling the relevant ones according to the presence (or lack thereof) of damage mentions. In our system, "damage" refers both to damage to buildings and other structures and to injuries, casualties, and missing people. In other words, damage encompasses all harmful consequences of an emergency on infrastructures and communities.

Here, the damage detection problem is approached as a two-level binary classification task. To our purposes, we are interested in identifying 4 different classes of tweets:

- *Not relevant*: tweets not related to a natural disaster.
- *Relevant*: tweets related to a natural disaster.
- *Without damage*: tweets related to a natural disaster but which do not convey information relevant to damage assessment.
- *With damage*: tweets related to a natural disaster which convey information relevant to damage assessment.

At the first level, the binary classifier (sub-module *ContentFiltering* in Figure 3.1) acts as a filter to discriminate between not relevant and relevant tweets, allowing only the latter ones to pass over to the second level. The classifier at the second level (sub-module *DamageDetector* in Figure 3.1) discriminates between tweets containing relevant information about damage and those not containing information relevant for damage assessment.

		damage			*no damage*			*not relevant*		
Dataset	**Accuracy**	**Precision**	**Recall**	**F-Measure (stdev)**	**Precision**	**Recall**	**F-Measure (stdev)**	**Precision**	**Recall**	**F-Measure (stdev)**
L'Aquila	0.83	0.92	0.87	0.89 (0.025)	0.81	0.87	0.84 (0.032)	0.77	0.71	0.73 (0.078)
Emilia	0.82	0.91	0.88	0.90 (0.039)	0.85	0.89	0.87 (0.016)	0.54	0.46	0.49 (0.060)
Sardinia	0.78	0.86	0.93	0.89 (0.019)	0.50	0.46	0.47 (0.099)	0.31	0.14	0.29 (0.113)

Table 3.6: *Results of the damage-detection task on the datasets using the NLP approach*

3.7.1 Natural Language Processing approach to damage detection

SM messages are processed by the system with a machine learning classifier that performs a multi-level linguistic analysis and operates on texts that are morpho-syntactically tagged and dependency-parsed by the DeSR parser using a multi-layer perceptron as the learning algorithm, a state-of-the-art linear-time shift-reduce dependency parser for the Italian language [51].

Given a set of features and a training corpus, the classifier creates a statistical model using the feature statistics extracted from the training corpus. This model is then employed for the classification of new tweets collected by the data acquisition component. We implemented the damage detection component as a linear Support Vector Machine (SVM) classifier using LIBSVM as the machine learning algorithm. We focused on a wide set of features organized into five main categories: (i) *raw and lexical text features*, including tokens count, n-grams analysis, hashtags number, punctuation, (ii) *morpho-syntactic features*, like part-of-speech n-grams, (iii) *syntactic features*, which cover lexical and type dependencies, (iv) *lexical expansion features* and (v) *sentiment analysis features*, including emoticons analysis, polarity n-grams, polarity modifiers. This partition closely follows the different levels of linguistic analysis automatically carried out on the text being evaluated, (i.e., tokenization, lemmatization, morpho-syntactic tagging and dependency parsing) and the use of external lexical resources [51]. In addition to the 3 feature classes derived from the linguistic analysis of tweets, we also exploited lexical expansion and sentiment polarity features. Lexical expansion features are frequently used to overcome the problem of the lexical sparsity in tweets, due to their short length in terms of words. Sentiment polarity features are used to infer the polarity of a piece of text. Notably, it has recently been demonstrated that these features can actually contribute to damage assessment tasks from SM [51]. This is because the emotional state of eyewitnesses of a disaster is typically reflected in the messages they share [11]

We evaluated the damage detection component on 3 datasets related to different disasters that struck Italy in recent years. Statistics about total collected data per disaster are reported in Table 3.2. Table 3.6 shows the results of our evaluation, carried out with a 10-fold cross validation process, with respect to well-know machine learning evaluation metrics. Results shows that the system achieved a good global accuracy ranging from 0.78 (`Sardinia`) to 0.83 (`L'Aquila`). For our goal, the scores obtained in the classification of the *damage* class are particularly important. The F-Measure score for this class is always higher than 0.89 thus showing that the damage detection component is accurate enough to be integrated in a crisis mapping system.

3.7.2 Word Embeddings

To overcome complexity related to NLP approaches, we also built two binary classifiers using the Support Vector Machines (SVM) algorithm [46] with a simple linear kernel as machine learning method[22]. The set of features used by the single classifier was obtained from tweets by analyzing the textual content of a tweet using an approach based on word embeddings [20].

The NLP (Natural Language Processing) research field has gained a lot of attention in the last years, due to the renewed interest in neural network technologies (e.g., deep learning), the continuous growth of the computational power of the CPUs and GPUs, and the explosion of available data that can be used to train these neural networks in an unsupervised way. In this Section we specifically used the approach proposed by Mikolov *et al.* [132] describing the `word2vec` software, which is currently the most popular model for embeddings used in NLP-related tasks. Word embedding techniques are an elegant solution to the problem of the features sparseness in document vectors created by using classic approaches like bag-of-words, char-N-grams, or word-N-grams [161]. From one side, they aim to create a vector representation with a much lower dense dimensional space. On the other side, they are useful to extract semantic meaning from text, to enable natural language understanding by learning the latent context associated with every specific word extracted from training data. More formally, distribute word representations (word embeddings) learn a function $W : (word) \rightarrow R^n$ that maps a word into a vector where each dimension models the relation of that specific word with a specific latent aspect of the natural language, both in syntactic or in semantic way. The vectors are learned through the training of neural language models together with the parameters of the network from a set of unannotated texts and according to an objective function (e.g., distributional hypothesis[23]) [19]. The resulting word vectors are computed so as to maintain the semantic/syntactic relationship existent between words which in turn allow to (i) visualize the vectors of similar words very close into a given metric space (e.g., visualize the word embeddings space on 2-D space through techniques like t-SNE), (ii) compute algebraic operations on vectors to point out some specific characteristic of the data (e.g., $W(\text{"}queen\text{"}) \cong W(\text{"}king\text{"}) - W(\text{"}man\text{"}) + W(\text{"}woman\text{"})$).

The approach we described in this section to build the damage-detection component is different from the one introduced in Subsection 3.7.1 on the same research topic [6]. Conversely from 3.7.1, we built this component using a multiclass classifier based on SVM with linear kernel but operating with features extracted from training data using classic NLP techniques [51]. The classifier based on classic NLP guarantees a good accuracy at classification time, but it has some significant drawbacks that we aim to mitigate in this Section. In particular, in 3.7.1 we used 5 different classes of features (e.g., lexical text features, morphosyntactic features, sentiment-analysis features, etc.) that are almost language-dependent and extracted with a quality level strongly dependent from the set of NLP tools and resources available to analyze the textual data. The choice of which features to extract (often referred to as the "feature engineering" problem) is a non-trivial task that must be solved in order to provide the classifier with a

[22] As software implementation we used the SVC class available in the `scikit-learn` Python package.

[23] The meaning of this hypothesis is that words appearing in similar contexts often have similar meaning.

	dataset	damage	no damage	not relevant
NLP	L'Aquila	0.89	0.87	0.73
	Emilia	0.90	0.87	0.49
	Sardinia	0.89	0.46	0.29
EMB	L'Aquila	0.85	0.82	0.72
	Emilia	0.85	0.87	0.52
	Sardinia	0.82	0.43	0.33

Table 3.7: *NLP classifier vs. Embeddings classifier in terms of $F1$ effectiveness compared over the three available labeled datasets.*

set of features sufficiently informative for the resolution of the specific problem, given the input domain. Moreover, the total number of features extracted in this way is very high (in the order of several hundred-thousands features) and it has a severe impact on the system in terms of performance both at training and classification time, this being dependent both from the complexity of the SVM algorithm (which is linearly increasing with the number of features) and the time spent to use external NLP tools to enrich data. Conversely, using our new approach based on word embeddings helps handling this type of issues because this technique completely avoids the feature engineering process and makes the system almost independent from any specific language. The only requirement affecting this model is the language coherence on a set of unannotated textual data used for the training of the embeddings, a condition often satisfiable very easily and with no human effort on many application domains (such as the Twitter domain). Another useful implication of word embeddings is the reduced number of features used by classifier, often being in the order of a few hundred, and therefore contributing to speed-up learning new models and classifying new documents.

To validate the goodness of our new proposal based on embeddings, we compared the $F1$ effectiveness [161] obtained with a 10-fold cross evaluation of both approaches using the multiclass single-label configuration proposed in our previous work, as shown in Table 3.7. The NLP classifier (reported as NLP) has been tuned as described in [6] while the embeddings classifier (reported as EMB) has been tuned as described in the following. The word embeddings vectors have been obtained by training the system on a dataset composed of the union of the tweets coming from `L'Aquila`, `Emilia`, `Sardinia`, and `Amatrice` using the CBOW method [132], and the size of the embeddings was set to 100. Given a tweet, to obtain a single vector representing the tweet content, we computed the tweet vector as the average sum of the vectors of the words contained in the text of the tweet[24]. The SVM classifier working over embeddings has been set to use the linear kernel with the parameter $C = 1$ and we have handled unbalanced data distribution among labels by assigning to each label a weight proportional to its popularity[25]. As reported in the Table 3.7, the embeddings classifier obtains very similar results to the NLP classifier in all tested datasets, confirming that our new approach provides all previously discussed advantages without sacrificing too much the accuracy of the damage detection system.

[24] We did not use more sophisticated methods like "Paragraph Vector" [114] because these statistical methods do not work well for small texts like tweets.

[25] We used the 'balanced' value for class weight, see `scikit-learn` documentation at `http://bit.ly/2g5QSqk`. In this way we indicate to SVM to treat the various labels in different ways during training phase, giving more importance to class errors (measured with used loss function) made for skewed classes.

This comparison also suggests that in order to improve the accuracy of the embeddings classifier we can operate at two different levels. Firstly, we can use more training data to train the classifier: a simple and effective solution is to merge the four separated datasets (`L'Aquila`, `Emilia`, `Sardinia`, and `Amatrice`) into one bigger training dataset. Secondly, to simplify the task of classification model's learning, we can change the target problem from a multiclass single-label classification problem into a pair of separated binary classification problems, as described at the beginning of this section. This last modification has also the advantage to clearly separate the filter part that identifies relevant tweets from the damage detection phase. This separation enables parallel execution of the damage classification task and the geoparsing task, decreasing the total time required to entirely process a single tweet.

We tested the system with the proposed improvements using a 10-fold cross evaluation and by optimising the model parameters in order to build a reasonable good set of classifiers. The measure used to drive the choice in the best configuration values is the micro averaged $F1$ [161]. The set of parameters subject to optimisation were the following:

- *Embeddings dataset*: the dataset used to learn word embeddings. We have used 5 possible different datasets: `Amatrice`, `L'Aquila`, `Emilia`, `Sardinia`, and the union of all previous ones (`All`).
- *Embeddings size*: the size used to represent word embeddings vectors and consequently the vector size of a single tweet. The possible set of values are 50, 100, 200, and 400. The default value is 100.
- *Class weight*: the weight of the errors associated with the positive class used in the two binary classifiers. The positive classes were "Not relevant" for filter classifier and "With damage" for damage classifier. The possible set of values were 1.0, 1.5, 2.0, 2.5 and "Balanced" (same meaning as explained in footnote 25). The default value is 'Balanced'.
- *C*: indicates the cost penalty associated with a misclassification. The possible set of values are $\in$ [0, 12] with a step increment of 0.1. The default values is 1.0.

To avoid testing all possible combinations of the above parameters, and in order to choose a reasonable good set of parameter values, we followed a simplified procedure as illustrated in the following. Using the order of the parameters as described above, we optimize one parameter at a time testing the full set of values for that specific parameter and using the default values for the other parameters. In case one of the other parameters has already been optimized, we use instead its best found value. The dataset used to evaluate the system[26] has been generated in two different versions, one for filtering and one for damage detection. In the case of filtering, every tweet in the dataset originally labeled with "With damage" or "Without damage" labels has been relabeled with "Relevant" label, resulting in a final dataset containing 4,351 relevant tweets and 857 not relevant tweets. In the case instead of damage detection, every tweet in the dataset originally labeled with "Not relevant" has been relabeled with "Without damage" label, resulting in a final dataset containing 3,672 tweets marked with "Without damage" label and 1,536 tweets marked with "With damage" label.

[26]The dataset is the union of the `L'Aquila`, `Emilia` and `Sardinia` datasets.

		Not relevant			Relevant			Micro avg results		
	Configuration	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
Emb. dataset	Amatrice	0.28	0.73	0.41	0.92	0.63	0.75	0.82	0.65	0.69
	L'Aquila	0.26	0.89	0.40	0.96	0.50	0.65	0.84	0.56	0.61
	Emilia	0.28	0.82	0.41	0.94	0.58	0.72	0.83	0.62	0.67
	Sardinia	0.21	0.85	0.34	0.93	0.37	0.53	0.81	0.45	0.50
	All **	0.36	0.79	0.49	0.95	0.72	0.82	0.85	0.73	0.76
Emb. size	50	0.35	0.80	0.48	0.95	0.71	0.81	0.85	0.72	0.76
	100 **	0.36	0.79	0.49	0.95	0.72	0.82	0.85	0.73	0.76
	200	0.35	0.79	0.49	0.94	0.72	0.81	0.85	0.73	0.76
	400	0.35	0.79	0.49	0.95	0.71	0.81	0.85	0.73	0.76
Class weight	1.0	1.00	0.00	0.00	0.84	1.00	0.91	0.86	0.84	0.76
	1.5	0.42	0.23	0.30	0.86	0.94	0.90	0.79	0.82	0.80
	2.0 **	0.44	0.54	0.48	0.91	0.86	0.88	0.83	0.81	0.82
	2.5	0.42	0.59	0.49	0.91	0.84	0.88	0.83	0.80	0.81
	Balanced	0.36	0.79	0.49	0.95	0.72	0.82	0.85	0.73	0.76
Best C	9.1	0.43	0.56	**0.49**	0.91	0.86	**0.88**	0.83	0.81	**0.82**

Table 3.8: *Choice of Optimal Parameters for Embeddings Classifier Filtering Relevant/not Relevant Tweets.*

		Without damage			With damage			Micro avg results		
	Configuration	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
Emb. dataset	Amatrice	0.92	0.75	0.82	0.58	0.85	0.69	0.82	0.77	0.78
	L'Aquila	0.87	0.89	0.88	0.72	0.69	0.70	0.83	0.83	0.83
	Emilia	0.95	0.88	0.91	0.76	0.89	0.82	0.89	0.88	0.89
	Sardinia	0.91	0.89	0.90	0.74	0.79	0.76	0.86	0.86	0.86
	All **	0.96	0.87	0.91	0.75	0.92	0.83	0.90	0.89	0.89
Emb. size	50	0.96	0.87	0.91	0.75	0.92	0.82	0.90	0.88	0.89
	100 **	0.96	0.87	0.91	0.75	0.92	0.83	0.90	0.89	0.89
	200	0.96	0.87	0.91	0.75	0.92	0.82	0.90	0.88	0.89
	400	0.96	0.87	0.91	0.75	0.92	0.82	0.90	0.88	0.89
Class weight	1.0	0.93	0.92	0.92	0.81	0.84	0.82	0.89	0.89	0.89
	1.5 **	0.95	0.89	0.92	0.78	0.88	0.83	0.90	0.89	0.89
	2.0	0.96	0.88	0.92	0.76	0.91	0.83	0.90	0.89	0.89
	2.5	0.96	0.87	0.91	0.75	0.92	0.82	0.90	0.88	0.89
	Balanced	0.96	0.87	0.91	0.75	0.92	0.83	0.90	0.89	0.89
Best C	0.4	0.95	0.90	**0.92**	0.78	0.88	**0.83**	0.90	0.89	**0.89**

Table 3.9: *Choice of Optimal Parameters for Embeddings Classifier Identifying With Damage/Without Damage Tweets.*

In Table 3.8 and Table 3.9 we report the experimental results obtained from the optimization process, respectively while building the filter classifier and the damage classifier. Except from the C parameter, which we report only as the best value found, we have marked with ** the best configuration found among all those tested. In the case of parameters "Embeddings dataset" and "Embeddings size" in both type of classifiers the best results are obtained with the full dataset (`All`) and 100 as the dimension of embeddings[27]. The optimal weight class values are different for each classifier reflecting the fact that the data distribution is very different between the two used datasets. In particular, the dataset used for filter classifier is more unbalanced towards relevant tweets, resulting in more variance in the results obtained for each tested configuration and in general providing quite low F1 values for class "Not relevant" (~ 0.5). Anyway, the results also show that the errors made for this last class by the filter classifier are partially recovered by the damage classifier, considering that the F1 accuracy on both damage classes are remarkably high (> 0.8 in both cases), resulting in an effective and useful implementation of the damage detection component.

3.8 Geoparsing

Geoparsing – namely, the resolution of toponyms in a textual document to a set of geographic coordinates – is typically considered the focal point of crisis mapping and has been faced since the diffusion of the Web. This task is typically solved by extracting toponyms from the text and looking for matches in gazetteers containing all the possible matches between a set of place names and their geographic coordinates [130]. This approach requires an offline phase where the geoparsing system is specifically set to work in a geographically-limited region. Although this solution is effective for limited areas and offers a fast response, it is practically infeasible to load associations between toponyms and coordinates for a wide region or for a whole country [6]. Another challenge related to geoparsing is that of toponymic polysemy – that is, the situation in which a toponym might have different meanings, thus possibly referring to different places, according to the context in which it is used (e.g., the word "Washington" may refer to the first US president, to the US capital, to the US state, etc.)[28]. This last problem is particularly relevant for geoparsing systems based on gazetteers lookups, since with this approach, there is no way to perform a disambiguating operation of the toponyms in order to understand their actual meanings [6, 49].

To overcome these limitations, the *GeoParsing* sub-module of the *DataEnricher*, shown in Figure 3.1, adopts semantic annotators in the geoparsing process. Semantic annotation is a process aimed at augmenting a plain-text with pertinent references to resources contained in knowledge-bases such as Wikipedia and DBpedia. The result of this process is an enriched (annotated) text where mentions of knowledge-bases entities have been linked to the corresponding Wikipedia/DBpedia resources. This annotation process is highly informative since it enables the exploitation of the rich information associated to the Wikipedia/DBpedia resources that have been linked to the annotated text. Here, we exploit semantic annotations for our geoparsing task by checking whether knowledge-bases entities, which have been linked to our tweets, are actually

[27] In case of configurations with equal results in terms of F1 we prefer to choose those having more balanced values between precision and recall measures.

[28] `http://en.wikipedia.org/wiki/Washington`

	GPS		DBpedia Spotlight		Dexter		TagMe	
L'Aquila	0	(0%)	271	(25.5%)	578	(54.4%)	721	(67.9%)
Emilia	205	(6.5%)	975	(30.8%)	1,037	(32.7%)	1,671	(52.7%)
Milan	15	(3.8%)	99	(25.3%)	320	(81.8%)	364	(93.1%)
Sardinia	51	(5.2%)	530	(54.3%)	784	(80.3%)	582	(59.6%)

Table 3.10: *Contribution of our GeoParsing sub-module, used in conjunction with the different semantic annotators, on the number of geolocated tweets.*

places or locations. Semantic annotation also has the side effect of alleviating geoparsing mistakes caused by toponymic polysemy. In fact, some terms of a plain-text can potentially be linked to multiple knowledge-bases entities. Semantic annotators automatically perform a disambiguating operation and only return the most likely reference to a knowledge-base entity for every annotated term. Overall, our proposed geoparsing technique overcomes 2 major problems affecting current state-of-the-art crisis mapping systems: (i) it avoids the need to preload geographic data about a specific region by drawing upon the millions of resources of collaborative knowledge-bases such as Wikipedia and DBpedia, (ii) it reduces the geoparsing mistakes caused by toponymic polysemy that are typical of those systems that perform the geoparsing task via lookups in preloaded toponyms tables. Another additional useful characteristic of our geoparsing technique is that it is unsupervised, unlike the one presented in [75].

Because of these reasons, our geoparsing technique is particularly suitable for being employed in a system aimed at producing crisis maps *impromptu*, such as the one that we are proposing. The possibility to link entities mentioned in emergency-related messages to their pages, also allows to exploit the content of their Wikipedia/DBpedia pages in order to extract other useful information about the unfolding emergency. Most commonly used semantic annotators also provide a confidence score for every annotation. Thus, it is possible to leverage this information and only retain the most reliable annotations, discarding the remaining ones.

Among all currently available semantic annotators, `CrisMap` is currently based on `TagMe` [69], `DBpedia Spotlight` [129], and `Dexter` [174], three well-known, state-of-the-art systems [175]. `TagMe` is a service of text annotation and disambiguation developed at the University of Pisa. This tool provides a Web application[29] as well as a RESTful API for programmatic access and can be specifically set to work with tweets. Since `TagMe` is based on the Wikipedia knowledge-base, the annotated portions of the original plain-text are complemented with the ID and the name of the linked Wikipedia page. `TagMe` also returns a confidence score *rho* for every annotation. Higher *rho* values mean annotations that are more likely to be correct. After annotating a tweet with `TagMe`, we resort to a Wikipedia crawler in order to fetch information about all the Wikipedia entities associated to the annotated tweet. In our implementation we sort all the annotations returned by `TagMe` on a tweet in descending order according to their *rho* value, so that annotations that are more likely to be correct are processed first. We then fetch information from Wikipedia for every annotation and check whether it is a place or location. The check for places/locations can be simply

[29] `https://tagme.d4science.org/tagme/`

achieved by checking for the *coordinates* field among Wikipedia entity metadata. We stop processing annotations when we find the first Wikipedia entity that is related to a place or location and we geolocate the tweet with the coordinates of that entity. The very same algorithmic approach is employed for the exploitation of the other semantic annotators: `DBpedia Spotlight` and `Dexter`. Indeed, it is worth noting that our proposed geoparsing technique does not depend on a specific semantic annotator and can be implemented with any annotator currently available or with a combination of them.

We used our *GeoParsing* sub-module to geocode all the tweets of our datasets. Then, following the same approach used in [130] and [74], we manually annotated a random subsample of 1,900 tweets to validate the geoparsing operation. Noticeably, our system achieves results comparable to those of the best-of-breed geoparsers with an $F1 = 0.84$, whether the systems described in [130] and [74] scored in the region of $F1 \sim 0.80$. Furthermore, to better quantify the contribution of our *GeoParsing* sub-module, we report in Table 3.10 the number of natively geolocated tweets (GPS column) and the number of tweets geolocated by our *GeoParsing* sub-module via `TagMe`, `DBpedia Spotlight`, and `Dexter`. As shown, our system geoparses the highest number of tweets based on the annotations of `TagMe`, for all datasets, except for the `Sardinia` one, for which the best results are achieved with `Dexter`'s annotations. In any case, our *GeoParsing* sub-module managed to geoparse from a minimum of 25.3% tweets, to a maximum of 93.1% tweets of the `Milan` dataset, meaning that almost all tweets of that dataset were associated to geographic coordinates, allowing to use such tweets in our crisis maps.

3.9 Mapping data

Results of the damage detection and message geolocation components are jointly exploited by the crisis mapping component. Given a set of tweets with damage and geolocation information, many data visualization techniques can be exploited in order to provide a clear picture of the unfolding emergency. Among the diverse data visualization techniques, *choropleth maps* are commonly employed to represent the geographical distribution of a statistical variable.

A choropleth map is a thematic representation in which subareas of the map are filled with different shades of color, in proportion to the measurement of the given variable being displayed[30]. This visualization technique is usually exploited to depict the spatial distribution of demographic features such as population, land use, crime diffusion, etc. In `CrisMap` we exploit the same visualization technique to show the spatial distribution of damage in the aftermath of an emergency. A clear advantage of exploiting choropleth maps instead of the typical on/off maps used in previous works [130], lies in the possibility to apply different shades of color to the different areas of the map, according to the estimated extent of damage suffered by that area. This complements well with the prioritization needs that arise in the first phases of an emergency response. Most notably, our system can be easily extended to produce different end-results. In other words, we can choose to produce a choropleth crisis map, or any other visualization of the analyzed tweets exploiting the high flexibility of the Kibana interface.

[30] `https://en.wikipedia.org/wiki/Choropleth_map`

task	evaluation metrics					
	Precision	*Recall*	*Specificity*	*Accuracy*	*F-Measure*	*MCC*
Detection of all damaged areas	0.895	0.202	0.992	0.797	0.330	0.365
Detection of areas that suffered significant damage	0.867	0.813	0.992	0.982	0.839	0.830

Table 3.11: *Binary detection of damaged municipalities for the* `Emilia` *earthquake using the NLP damage damage detector*

Notably, `CrisMap` is capable of producing choropleth crisis maps with a spatial resolution at the level of *municipalities*. Anyway, when tweets are accurate enough, it is also possible to precisely identify objects (e.g., a specific building) that suffered damage. It is also worth noting that the choice to produce crisis maps showing the estimated degree of damage among the different municipalities is not due to a region-level only geoparsing. Indeed, the exploitation of semantic annotators potentially allows to geocode every entity that has an associated Wikipedia/DBpedia page. So, in those cases when a tweet contains detailed geographic information, it is possible to geoparse it to building- or even street-level. Our choice to produce crisis maps at the level of municipalities is instead motivated by an effort to rigorously compare our crisis maps to data officially released by the Italian Civil Protection agency, which reports damages at the municipality-level.

Here, we first show the accuracy of our visualizations referring to two case studies, the `Emilia` earthquake and the `Sardinia` flood. We compare the maps realized with SM data against those produced using authoritative data provided by Italian civil protection agency. Finally, we present results obtained by applying `CrisMap` to study the `Amatrice` earthquake. Unfortunately, there is no fine economic loss estimation for the `Amatrice` earthquake, since the same area suffered a second severe shake just few months after the first one, when official damage surveys still had to be completed. Nonetheless, in the case of the `Amatrice` earthquake, we are still able to provide a qualitative case-study of our crisis maps.

3.9.1 Quantitative validation

The authoritative data that we used for the comparison is the economic loss/damage (quantified in millions of euros) suffered by the different municipalities, as assessed by the Italian Civil Protection agencies of Emilia Romagna[31] and Sardinia[32].

It is possible to perform a first quantitative evaluation of our crisis maps following the approach used in [130], that is, performing the evaluation as a classification task. Under this hypothesis, the goal of the system is to detect damaged municipalities disregarding those requiring prioritized intervention – namely, those that suffered the most damage. Thus, we can exploit well-known machine learning evaluation metrics to compare crisis maps generated by our system with those obtained from official data. The comparison is performed by checking whether a municipality with associated damage in authoritative data also appears as damaged in our crisis maps.

Tables 3.11 and 3.12 report the results of this comparison for the `Emilia` earth-

[31] `http://www.openricostruzione.it` - Web site maintained by the Emilia Romagna regional district.

[32] `http://www.regione.sardegna.it/documenti/1_231_20140403083152.pdf` - Italian Civil Protection report about damage to infrastructures and agricolture.

	evaluation metrics					
task	*Precision*	*Recall*	*Specificity*	*Accuracy*	*F-Measure*	*MCC*
Detection of all damaged areas	1.000	0.178	1.000	0.797	0.303	0.375
Detection of areas that suffered significant damage	1.000	0.813	1.000	0.992	0.897	0.898

Table 3.12: *Binary detection of damaged municipalities for the* `Emilia` *earthquake using the WE damage detector*

	evaluation metrics					
task	*Precision*	*Recall*	*Specificity*	*Accuracy*	*F-Measure*	*MCC*
Detection of all damaged areas	0.640	0.410	0.973	0.915	0.500	0.470
Detection of areas that suffered significant damage	0.500	0.643	0.973	0.960	0.563	0.545

Table 3.13: *Binary detection of damaged municipalities for the* `Sardinia` *flood using the NLP damage detector.*

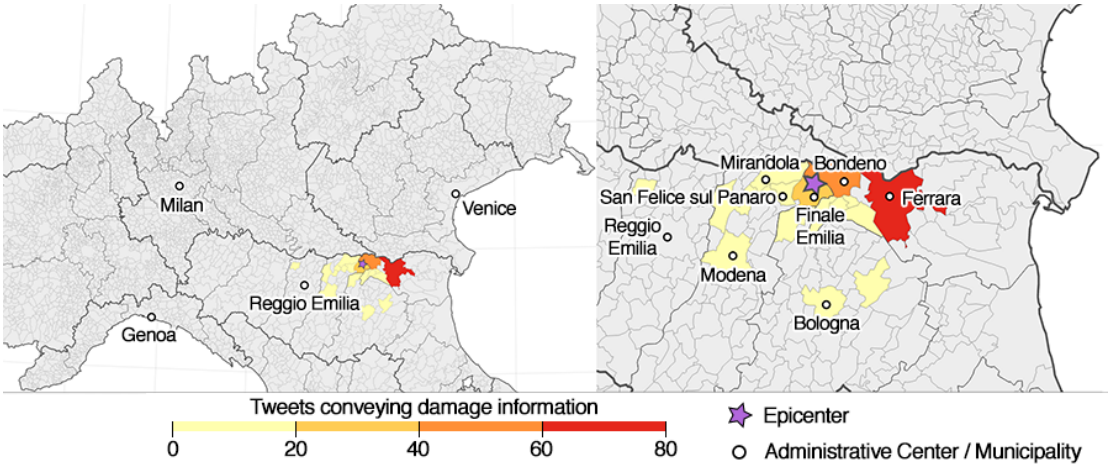
quake. We first consider all the municipalities of the affected region, namely the Emilia Romagna region, and then we repeat the comparison by only considering those municipalities that suffered a significant degree of damage (more than 10% of the damage suffered by the Ferrara municipality, which is the maximum value for the `Emilia` earthquake). As clearly highlighted by Table 3.12, the proposed crisis mapping system is able to accurately identify the areas where damage actually occurred. However, not all the damaged municipalities are identified by the system, as represented by the low Recall value in the first row of the tables. Anyway, when we remove from the comparison those municipalities that suffered the lowest damage, the Recall metric reaches a value of 0.813, showing the system's ability in detecting areas that suffered a significant amount of damage. In other words, the great majority of the mistakes of our system occurred in municipalities that suffered relatively low damage and not on those requiring immediate attention. The same also applies for the `Sardinia` flood, as reported in Tables 3.14 and 3.13, with an improvement of the Recall metric from 0.128 to 1 when considering municipalities that suffered more than 2% of the maximum damage (i.e. the damage suffered by the municipality of Olbia).

Overall, the results obtained by our system with respect to the detection of damaged areas are comparable to those reported in [130]. However, our system operated with a fine geographic resolution on 2 case studies of natural disasters that affected wide, rural, and sparsely populated areas. Conversely, the system presented in [130] has a fine resolution only for an emergency affecting a densely and uniformly populated area (Manhattan, New York) while it shows coarse resolution results for a disaster striking a

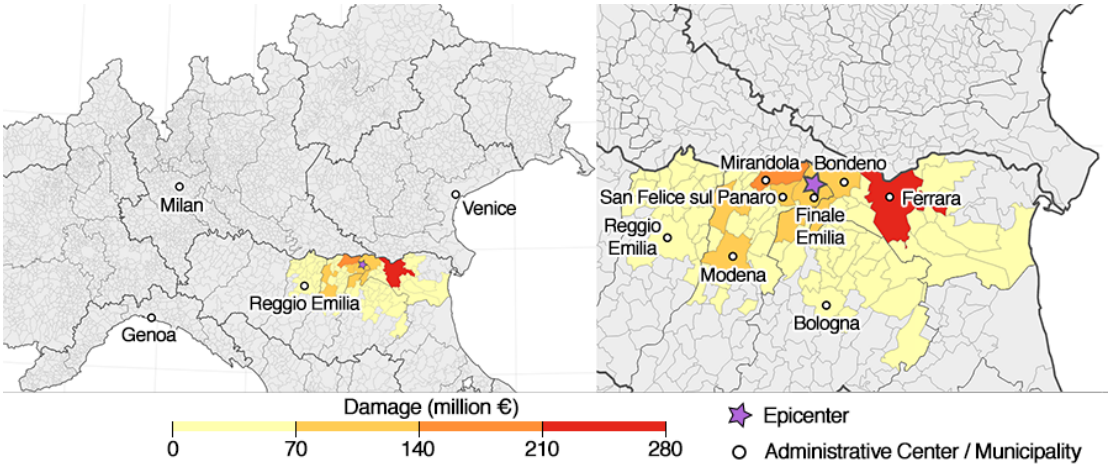
	evaluation metrics					
task	*Precision*	*Recall*	*Specificity*	*Accuracy*	*F-Measure*	*MCC*
Detection of all damaged areas	0.833	0.128	0.993	0.814	0.222	0.280
Detection of areas that suffered significant damage	0.500	1.000	0.995	0.995	0.667	0.705

Table 3.14: *Binary detection of damaged municipalities for the* `Sardinia` *flood using the WE damage detector.*

(a) *Overview and detail of the crisis map derived from SM data.*



(b) *Overview and detail of the authoritative post-event damage assessment map. Authoritative data about the economic losses are provided by Emilia Romagna regional district (*`http://www.openricostruzione.it`*)*



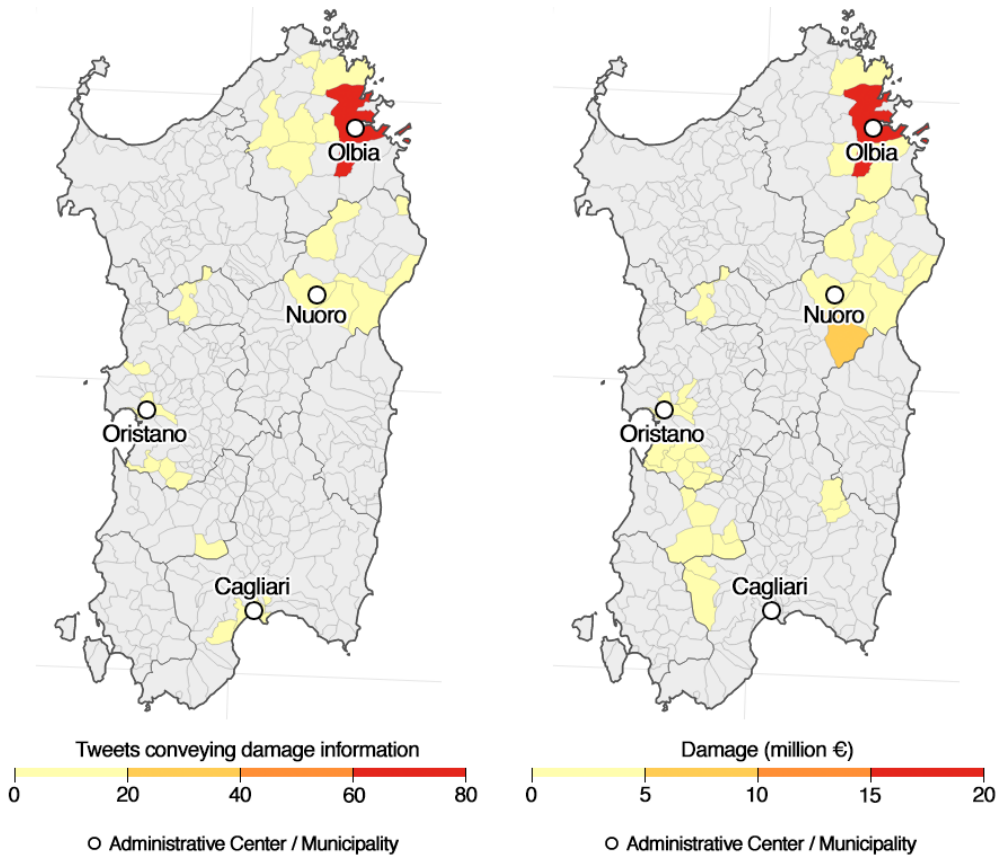
.

Figure 3.4: *`Emilia` 2012 earthquake.*

wide area (the state of Oklahoma).

Figure 3.4a shows the choropleth crisis map generated by our system with regards to the `Emilia` 2012 earthquake. The map assigns a color to a municipality according to the number of tweets of the *damage* class geolocated in that municipality. Areas in which our system did not geolocate any *damage* tweets are grey colored. As shown, despite geolocating tweets in all Northern Italy, our system only highlighted municipalities around the epicenter, clearly pointing to the damaged area. In the Figure, reddish colors are assigned to the municipalities of Bondeno, Ferrara, Finale Emilia and San Felice sul Panaro, thus accurately matching the most damaged locations. A qualitative evaluation of our crisis map can be obtained by comparing Figures 3.4a and 3.4b, the latter depicting authoritative data about the economic losses suffered by the municipalities hit by the earthquake, released by the Emilia Romagna regional district. In the case of the `Sardinia` 2013 flood, a similar evaluation comes through comparing Figures 3.5a and 3.5b.

In addition to detecting damaged areas, `CrisMap` also visually sorts municipali-



(a) *Overview of the crisis map derived from SM data.*

(b) *Overview of the authoritative post-event damage assessment map. Authoritative data about the economic losses are provided by the Civil Protection Agency of Sardinia regional district (*`http://www.regione.sardegna.it/documenti/1_231_20140403083152.pdf`*).*

Figure 3.5: `Sardinia` *2013 flood.*

ties using a color hue that is proportional to the intensity of the damage suffered. In other words, it order ranks municipalities based on the (normalized) number of tweets conveying damage information. This unprecedented feature opens up the possibility to perform a finer evaluation of our crisis maps than that carried out in previous works. Indeed, it is possible to compare the ranking of damaged municipalities as obtained from tweets, with a ranking derived from authoritative sources, such as those provided by civil protection agencies. A crisis mapping system that is able to rapidly identify the most damaged areas would become a valuable tool in the first phases of emergency response, when resource prioritization plays a dominant role. A possible way of performing such evaluation is by employing metrics that are typically used to assess the performance of ranking systems, such as search engines. Search engines are designed to return the most relevant set of results to a given user-submitted query. In our scenario, we can consider `CrisMap` as a basic "search engine" that returns a list of areas and that is specifically designed to answer a single, complex query: "which areas suffered the most damage?". Search engines are evaluated with several metrics and indices aimed at capturing a system's ability to return desired resources (e.g.; Web pages, text documents, etc.) among the first results. We can then evaluate the ability of `CrisMap` to correctly identify areas that suffered a high degree of damage by employing evaluation metrics of search engines. Specifically, among such well-known metrics are the *normalized Discounted Cumulative Gain* ($nDCG$) [100] and the *Spearman's Rho coefficient*. The $nDCG$ measures the performance of a "recommendation" (or ranking) system based on the graded relevance of the recommended entities. It is the normalized version of the *Discounted Cumulative Gain* and ranges from 0 to 1, with 1 representing the ideal ranking of the entities. This metric is commonly used in information retrieval to evaluate the performance of web search engines:

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

$$nDCG_k = \frac{DCG_k}{IDCG_k}$$

where rel_i is the graded relevance of the result at position i and $IDCG_k$ is the maximum possible (ideal) DCG for a given set of entities.

Spearman's Rho instead measures the correlation between two variables described using a monotonic function, and it is evaluated as:

$$\rho = 1 - \frac{6 \sum_i D_i^2}{N(N^2 - 1)}$$

where $D_i = r_i - s_i$ is the difference between actual position (given by the system) and expected position (given by the reports). For instance, it measures the correlation between the ideal output of the system (Civil Protection ordering) with the result of the system and describes how likely one variable is going to change (tweets with damage) given the other (damage amount). Being a correlation coefficient, it ranges from -1 to 1, with values in the region of 0 indicating no correlation. Using these metrics we assessed the ability of our system in detecting the most stricken areas against authoritative data based on the economic damage suffered by the affected municipalities.

	NLP approach		WE approach	
dataset	*nDCG*	*Spearman's Rho*	*nDCG*	*Spearman's Rho*
Emilia	0.894	0.596	0.770	0.698
Sardinia	0.765	0.521	0.647	0.408

Table 3.15: *Ranking evaluation of most damaged municipalities.*

Our experiments confirm that there is a considerable agreement between tweet-derived rankings and those based on authoritative data, as reported in Table 3.15.

A simple test for statistical significance of our ranking results with Spearman's Rho further supports our claims, achieving a confidence score $> 99\%$ for both the Emilia earthquake and the Sardinia flood. Overall, results of our system in detecting all damaged areas, as well as the most damaged ones, demonstrate the applicability and the usefulness of CrisMap also in wide, rural, and sparsely populated regions.

3.9.2 The qualitative Amatrice case-study

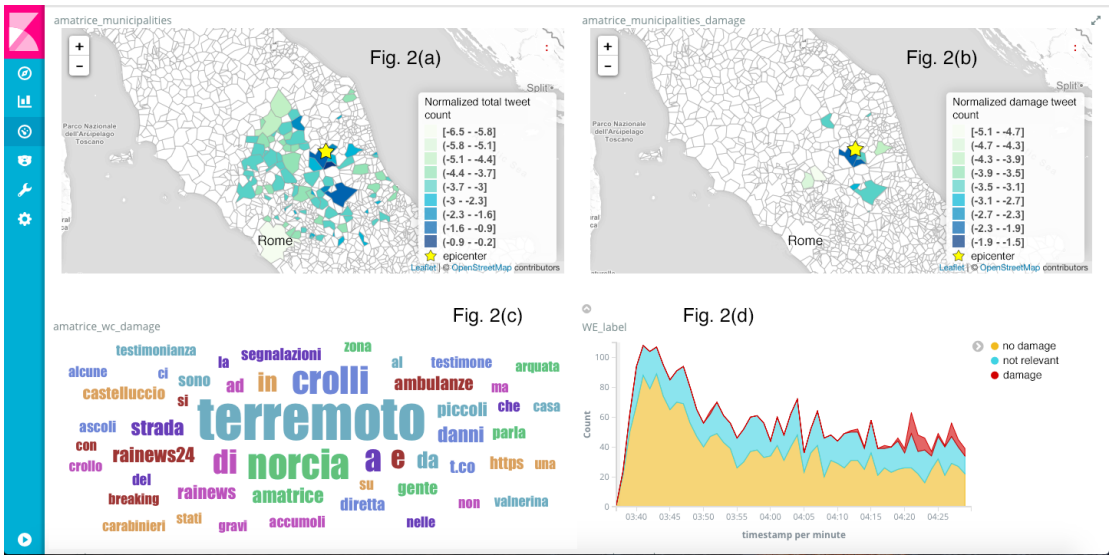


Figure 3.6: *Excerpt of the CrisMap dashboard in the aftermath of the Amatrice earthquake. The dashboard comprehends our ad-hoc choropleth maps showing the spatial distribution of tweets, a temporal view of collected and classified tweets per minute and a word-cloud. The dashboard is easily extensible and customizable, allowing end-users to include more visualizations by choosing from the many ones natively provided by Kibana.*

Figure 3.6 shows an excerpt of the CrisMap dashboard in the aftermath of the Amatrice earthquake. All the visualizations shown in figure are related to the Amatrice dataset collected during the first hour from the earthquake occurrence. On top of Figure 3.6 there are two choropleth maps – visualizations 3.6(a) and 3.6(b) – obtained by embedding the choropleth plugin, that we specifically developed for CrisMap, inside the Kibana interface. Figure 3.6(a) shows the map generated from all the tweets that we collected, while Figure 3.6(b) is obtained only from *damage* tweets. Our choropleth maps clearly highlight the most damaged municipalities, namely Norcia, Amatrice and Accumoli, located in the vicinity of the epicenter. The map in Figure 3.6(b), related

only to damage tweets, is rather sparse, since in the first hour after the earthquake only a very small fraction of tweets conveyed damage reports. This is also clearly visible from Figure 3.6(d), showing the volume of tweets collected and classified by `CrisMap` every minute. As shown, tweets reporting damage (red-colored) have been shared almost only in the last minutes of the first hour. Yet, despite the very few tweets reporting damage, the word-cloud of Figure 3.6(c) highlights the key consequences of the earthquake: (i) the 3 most damaged villages Norcia, Amatrice, Accumoli; (ii) and several mentions of damage, such as "crolli" (collapsed buildings) and "danni" (widespread damage).

Notably, all the visualizations of Figure 3.6 are interactive and are updated in real-time, as new data is collected and analyzed by `CrisMap`. Regarding user interactions, for instance, it is possible to click on a specific municipality in the choropleth maps in order to update all other visualizations by showing only data that is related to the selected municipality. Alternatively, one could click on a word in the word-cloud to visualize the temporal and spatial distribution of tweets containing that word. These functionalities open up the possibility to promptly perform drill-down analyses of the most relevant facts.

3.10 Discussion

This Chapter discussed a methodology and a set of technologies capable of supporting emergency responders (e.g. national Civic Protection agency) during crisis management of natural or man-made disasters. The methodology is able to process incoming SM data from Twitter in order to quickly produce crisis maps useful to prioritise the allocation of available resources, especially in the first phases of the crisis, towards the populations and territories most affected by the specific disaster. The proposed solution is designed and built using Big Data technologies, allowing the system to be scalable, fault-tolerant, and able to process incoming data in near-real-time. To overcome the limits resulting from the unstructured nature of Twitter data and to identify useful information for our purposes, we analyze the data using a two-fold perspective. On the one hand, we introduced a damage detection component exploiting word embeddings and a SVM classifier to detect messages reporting damage to infrastructures or injuries to the population. On the other hand, we proposed a message geolocation component that performs the geoparsing task by exploiting online semantic annotators and collaborative knowledge-bases. The approach using word embeddings has also been compared with a traditional one based on classic NLP techniques, pointing out the potential advantages of the former in relation to complexity and performance of the proposed method. The accuracy and the reliability of the system was validated analytically comparing the experimental results of `CrisMap` against the authoritative data for 2 past disasters. Furthermore, we also performed a qualitative evaluation of the system on a case-study of a recent severe earthquake in Italy for which authoritative data are not available. The proposed system offers some room for further improvements, to increase the readability of the maps and, in general, to acquire a better situational awareness of unfolding events. With regards to geoparsing results, recent developments of semantic annotation tools open up the possibility to provide more implementations of our proposed geoparsing technique. Therefore we envision the possibility to simultaneously exploit multiple semantic annotators in an ensemble or voting system. In the future

this approach could allow to obtain even better results and to overcome the possible limitations of a single annotator.

Although this kind of tool leads to a near-real-time solution for monitoring a territory, there might be an issue related to the privacy concern. In fact, the system is based on the opportunistic crowdsensing approach, which requires to access SM messages with public scope. However, if privacy settings of users are too restrictive, the system might not be able to retrieve enough messages and wouldn't work. Here we need to find a balance between the privacy regarding the private sphere and the common good, which is still a widely debated topic. Moreover, while this tool was built for noble intentions, there's no guarantee that it will not be used also for mass surveillance of the population, shifting its main purpose to build a sort of digital "Big Brother".

The crisis mapping system like `Crismap` [10] analyses the textual content of tweets. However, tweets might also contain other useful content, such as images and URLs. A possible direction of improvement includes the online analysis of images and the content of linked Web pages in order to further contribute to the effectiveness of damage detection and geoparsing.

Other avenues of future experimentation might be related to multi-source mining. Indeed, although Twitter is nowadays one of the preferred SM sources, given its "open" policies on providing data to third parties, data collection from multiple sources could mitigate the bias introduced by the analysis of a single SM.

Results obtained by our system in detecting the most damaged areas are promising, also taking into consideration the rural nature of the case studies, characterised by sparsely populated regions. In fact, when operating in the social mining area, one should always consider the number of potential tweeters in the area under investigation, since the performances of any SM-based system can be impaired by the lack of data resulting from a low message rate.

Despite the promising results, our study also highlighted issues that require further investigation. Among them, damage detection. A deeper linguistic analysis aimed, for instance, at identifying the object that suffered the damage (e.g., a building, a bridge, a person) and the severity of the damage, would undoubtedly be desirable. The damage detection component could then output tuples of <*object, degree of damage*>, thus enabling a more specific prioritized intervention.

The proposed approach to geoparsing behaves effectively, but the validation process showed that there is still space for considerable improvements, especially regarding the Recall metric. In addition, the granularity of the geoparsing results via semantic annotators should be assessed and compared to other approaches. Recent developments of semantic annotators open up the possibility to provide more implementations of our geoparsing technique. For example, we envision the possibility to simultaneously exploit multiple annotators in an ensemble or voting system. The extension of our analysis pipeline to languages other than Italian and the assessment of the resulting system performances also requires further investigation.

Finally, the general system architecture proposed in this Chapter can be easily adapted to perform analysis other than emergency management, such as the monitoring of public events or other broad social phenomenon, by applying similar detection algorithms to those adopted for emergency detection and management. In particular we think that such a system is suitable for non-crisis situations where time-sensitive spatio-temporal

analyses on big streams of data are required.

CHAPTER 4

Violence and Social Media

Most of the work described in this Chapter is focused on detecting hate speech in SM conversations. Although this is not the first time this issue has been addressed in literature, we believe that the Italian language can still benefit from further works aimed at limiting the negative effect of SM misuse. The work [59] describes a possible mitigation to this problem and hereby is described how the implementation works.

4.1 Introduction to hate speech in Social Media: issue and consequences

Social Network Sites (SNSs) are ideal places for Internet users to keep in touch, share information about their daily activities and interests,for publishing and accessing documents, photos and videos, and are adopted by a varied range of users, spanning from teenagers to mid-age people or older. SNSs like Facebook, Twitter, Ask.fm and Google+ give the ability to create profiles, to have a list of peers to interact with and to post and read what others have posted; every SNS has its user base with different sex partitions and age ranges. It comes as no surprise that, overall, SNSs - together with search engines - are among the most visited websites[1].

Unfortunately, SNSs are also the ideal plaza for proliferation of harmful information. Cyberbullying, sexual predation [108], self-harm practices incitement [35] are some of the effective results of the dissemination of malicious information on SNSs. Many of these attacks are often carried out by a single individual, but they can also be managed by groups. The target of the *trolls* are often selected victims but, in some circumstances, the hate can be directed towards wide groups of individuals, discriminated for some features, like race or gender. Such campaigns may involve a very large number of *haters* that are self excited by hateful discussions, and such hate might end up with

[1] `http://www.alexa.com/topsites` - All websites were last accessed on October, 09, 2017.

physical violence or violent actions. To study and monitor the phenomenon of hate in SNSs which will be referred to as hate speech, this Chapter proposes a methodology to prevent the important social consequences of massive online hate campaigns and will compare the approach with other results of other academic works.

Work in [89] characterises the attacker and it provides a definition of *trolls*, i.e., online users pretending to sincerely strive to be part of an online community, but whose real intentions are to cause disruption and exacerbate conflict, for the purposes of their own amusement. Thus, sexists, religious fanatics, political extremists massively use SNSs to foster *hate* against specific individuals/organisations, by causing a sounding board effect, which may critically damage the targets of the hate campaign, by using both psychological and physical violence.

Although more experienced users could face threats and trolls, the great majority of them cannot bear the attacks easily, especially minors and those who might get exposed mediatically to public judgment. Media frequently report evidences about the (unfortunately extreme in some cases) consequences that naive and emotive users have faced to[2].

The issue has been tackled in the past with different approaches, laying somehow in the middle between pre-emption and mending. One approach aims at mitigating chat conversations through ad hoc filters, like in [189], by semantically detecting offensive content and removing it. The second approach operates on published content and tries to remove the offending one, often leveraging the analysis on multiple messages, as in [27, 37, 186].

Interestingly, the connections between the users' profiles on SNSs are often strictly related to the connections in their real life [63]. Using machine learning, it was possible to recognise those users that adopt troll profiles in cyberbullying practices [28, 52, 72]. Similarly, text analysis approaches have been used to link together the contents of anonymous users among different opinion websites [1]. Relationships based on profile connections and behaviours have been exploited to effectively identify fake Facebook profiles [44], while lightweight profile features succeeded in recognising fake Twitter followers [48, 50]. Regarding text classification for automatic hate speech detection, a seminal work is in [168], which is one of the first and successful applications of hate speech detection. In [124], the authors propose a rule-based classifier to distinguish between legitimate and abusive information in texts. PALADIN [107] is a pattern mining tool to mine patterns of language and to detect anti-social behaviours of users. The authors of [186] focus on Twitter and propose a semi-supervised approach and statistical topic modeling for the detection of offensive content, while work in [27] presents a supervised machine learning text classifier, trained and tested to distinguish between hateful and antagonistic responses, with a focus on race, ethnicity and religion. Work in [62] adopts neural language models to learn distributed low-dimensional representations of comments. The approach generates text embeddings that can be used to feed a classifier. The authors of in [78] describe the distinction among flame and hate speech (the latter being more directed to groups, rather than individuals). The same work proposes the three level hate classification adopted in this paper (partially suffering for the low IAA too). Some studies concentrate on the users' behaviour. Authors of [141] pro-

[2] `http://osservatorio-cyberbullismo.blogautore.repubblica.it/` (La Repubblica - Italian newspaper online edition)

pose a reputation system, which tracks reputation of users using positive and negative opinions. A behavioural analysis of banned users is in [37], showing a certain degree of similarity in their texts, containing often irrelevant content too.

Work	Target	Technique
ChatCoder [108]	Study of chat code transcripts to analyse to develop a theory on online predation.	Use J48 of Weka software to classify conversations of predators and victims. Correct classification 60% of times.
Chattopadhyay [35]	Evaluation of suicidal risk.	Retrospective analysis of patients and development of a classifier.
Xu [189]	Automatic filtering of offensive content and proposition of an automatic sentence-level filtering approach able to semantically remove the offensive language.	Grammatical relations among words. The approach leads to 90.94% accuracy on English filtering.
Burnap [27]	Build a supervised machine learning text classifier, trained and tested to distinguish between hateful and/or antagonistic responses with a focus on race, ethnicity or religion; and more general responses.	Combination of probabilistic, rule-based and spatial based classifiers with a voted ensemble meta-classifier. This results in an overall F-measure of 0.95 using features derived from the content of each tweet.
Cheng [37]	Characterisation of antisocial behaviour.	Study of the behaviour of banned users in a community.
Xiang [186]	Detection of profanity-related content in Twitter.	Logistic Regression, with a True Positive rate of 75.1%.
Cambria [28]	Detect trolls and prevent web-users from being emotionally hurt by malicious posts.	Sentic Computing using WordNet-Affect to evaluate the trollness of the users. Results lead to an F-score of 0.78.
Dadvar [52]	Detection of cyberbullying.	Analysis of user context, with textual and meta information. F-score achieved was 0.64.
Galàn-Garcìa [72]	Detect and associate fake profiles on Twitter social network, which are employed for defamatory activities, to a real profile within the same network.	Analysis of the content of the comments generated by both profiles with different classification algorithms. Best accuracy, provided with a PolyKernel, is 68.47%.

Spertus [168]	Build a system for flame recognition.	C4.5 classifier based on a set of rules. Flames are correctly classified on 64% of cases.
Mahmud [124]	Build an automated system to distinguish between information and personal attacks containing insulting or abusive expressions in a given document.	Set of rules to extract the semantic information of a given sentence from the general semantic structure of that sentence to separate information from abusive language.
Djuric [62]	Hate speech detection in online user comments.	Learning of distributed low-dimensional representations of comments using neural language models (word-embeddings), that can then be fed as inputs to a classification algorithm.
Gitari [78]	Build a classifier that can be used to detect the presence of hate speech in web discourses such as web forums and blogs, including the proposition of a hate speech taxonomy.	Using subjectivity and semantic features related to hate speech, authors create a lexicon that is employed to build a classifier for hate speech detection. Best F-score is 70.83%.
Del Vigna [59]	Build a classification model to **detect hate speech on SM for the Italian language on Facebook**.	**NLP features and word embeddings** are used to build a SVM and LSTM classifiers. Best result in hate detection for F-measure is 72.8% and 80.60% for accuracy.

Table 4.1: *Evaluation of the different solution for hate speech detection, filtering and reporting in literature, compared to this work.*

Table 4.1 summarises and compares the major works in the area of hate speech classification that represent the actual state-of-the-art that can be found in the literature. The aim of this study is not to censor online content, but we mostly address its classification for the Italian language to pinpoint anomalous waves of hate and disgust. Using Facebook as a benchmark, the system classifies the content of comments which appeared on a set of public pages.

Given that social issues caused by hate can lead to very unpleasant consequences either in mental discomfort or in public order and security, here this thesis proposes a method to address the issue based on work in [59]. Mainly, what is hereby discussed includes:

- the design and development of a hate speech classifier for the Italian language and

a comparison between two different approaches based on state-of-the-art learning algorithms for sentiment analysis tasks;

- starting from classifying the single comment on a Facebook page, the results proposed in this paper constitute the prelude to the detection of violent discussions as a whole. This with the ultimate goal to promptly detect waves of hate, which several users may take part in, as has happened recently and unfortunately on Facebook pages[3,4];
- the introduction of a taxonomy for a variety of hate categories, expanding the classes proposed in [78] and specifically considering the *subject of the hate*, e.g., hate for religious, racial, socio-economical reasons; while not directly employed here in the classification, the definition of such taxonomy is a step towards more refined classification tasks.

The next section introduces the corpus for hate speech detection, while Section 4.3 presents the classification techniques and reports its performance results. Finally, Section 4.4 draws some conclusions about hate speech detection and monitoring and discusses some implications.

4.2 Hate Speech Corpus

For the building of the hate speech detection tool this work made use of a hate speech Italian corpus, which has been retrieved from Facebook purposely for this task and manually annotated.

4.2.1 Data crawling

Aiming at monitoring the “hate level" across Facebook, we have built a corpus of comments retrieved from the Facebook public pages of Italian newspapers, politicians, artists, and groups. These pages typically host discussions spanning across a variety of topics that trigger threads of comments full of hate.

We have developed a versatile Facebook crawler, which exploits the Graph API[5] to retrieve the content of the comments to Facebook posts. The crawler leverages the flexibility of the PHP Laravel framework to deploy a wide set of features, like flexibility, code reuse, different storage strategies and parallel processing. Implemented as a Web service, it can be controlled through a Web interface or by using a cURL command[6]. The tool requires a set of registered application keys and some target pages to crawl. It is capable of storing data in the filesystem either as JSON[7] files, or in Kafka[8] queues or in Elasticsearch[9] indexes. According to the number of application keys provided to the application, it is able to crawl multiple pages in parallel. Starting from the most recent post, the crawler collects all the information related to the posts, up to comments

[3] https://www.facebook.com/Black-block-430370993643807/ (Facebook page)

[4] http://www.iltempo.it/cronache/2016/10/05/news/tiziana-cantone-suicida-per-i-video-hard-facebook-dice-no-alla-rimozione-delle-immagini-1022376/ (Il Tempo - Italian Newspaper online edition)

[5] https://developers.facebook.com/docs/graph-api

[6] https://curl.haxx.se

[7] http://json.org

[8] http://kafka.apache.org

[9] https://www.elastic.co/products/elasticsearch

Title of Facebook page	Annotated posts	Comments	Annotations
salviniofficial	19	5404	15298
matteorenziufficiale	2	158	584
lazanzarar24	10	307	1253
jenusdinazareth	2	132	460
sinistracazzateliberta2	7	79	234
ilfattoquotidiano	11	126	135
emosocazzi	4	73	75
noiconsalviniufficiale	14	223	270

Table 4.2: *Dataset description and annotations of the dataset.*

to comments. However, for the sake of simplicity, in this work we have limited our analysis to direct comments to the posts.

4.2.2 Data annotation

The crawler was used to collect comments related to a series of web pages and groups, chosen since they were suspected to possibly contain hate content. Those pages, in facts, are related to politicians and satirical pages, which often triggers violent conversations.

Overall, 6,502 posts have been annotated at least once (spanning over 66 posts) and at most received 5 annotations from distinct human annotators. Moreover, a total of 18309 annotations have been produced by human annotators on comments related to those posts. 5 bachelor students were asked to annotate comments, and the majority of comments received more than one annotation. Students annotated 5742, 3870, 4587, 2104 and 2006 comments respectively. In particular, among the annotated comments, 3,685 received at least 3 annotations. On average, each annotator annotated about 3,662 comments. For convienience, we reported in Table 4.2 the annotated dataset.

The annotators were asked to assign one class to each post, where classes span over the followings levels of hate: *No hate*, *Weak hate*, and *Strong hate*.

We then divided hate messages into distinct categories: *Religion*, *Physical and/or mental handicap*, *Socio-economical status*, *Politics*, *Race*, *Sex and Gender issues*, and *Other*.

Given that the majority of comments have been annotated by more than one annotator, we have also computed the Fleiss' kappa κ inter-annotator agreement metric [83], which measures the level of agreement of different annotators on a task. The level of agreement among annotators conveys the level of the difficulty of a task. In our case, considering the 1,687 comments that received annotations from all the 5 annotators, we obtain $\kappa = 0.19$ when discriminating over three hate classes, while $\kappa = 0.26$ over two classes (where *Strong Hate* and *Weak Hate* have been merged together). Such low κ values testify that the annotation task was really hard for our students.

4.3 Text Classification

This section describes the classification approaches and gives their results. From the annotated dataset, a series of features is extracted and computed to build two classifier models. These features are described in detail in the following, whilst a series of lexi-

cons used to derive part of the features is described in Section 4.3.1. The comments in the dataset are then represented as a vector of features, given as input to the classifier, along with the result of the annotation. In the training phase, the classifier learns to classify a comment according to the values of its features and the annotation result. In the test phase, the classifier takes decision and tags comments as expressing hatred or not, according to the learned model.

4.3.1 The classifier

We tested two different classifiers based on different learning algorithms: the first one based on Support Vector Machines (SVM) and the second one on a particular Recurrent Neural Network named Long Short Term Memory (LSTM). While SVM is an extremely strong performer, hardly to be transcended, unfortunately this type of algorithm capture "sparse" and "discrete" features in document classification tasks. This makes the detection of relations in sentences really hard, while this is often the key factor in detecting the overall sentiment polarity of a document [170]. On the contrary, LSTM networks are a specialisation of Recurrent Neural Networks (RNN), which are able to capture long-term dependencies in a sentence. This type of neural network has recently been tested on sentiment analysis tasks [170, 188], reaching outperforming classification performance [136], with even a 3-4 points improvement with respect to commonly used learning algorithms. This work makes use of the Keras [40] deep learning framework and LIBSVM [33] to generate, respectively, the LSTM and the SVM statistical models. Since the approach relies on morpho-syntactically tagged texts, the hate speech corpus was automatically morpho-syntactically tagged by the Part-Of-Speech tagger described in [61].

Lexical resources

To improve the overall accuracy of our system, the system uses both sentiment polarity and word embedding lexicons. Sentiment polarity lexicons[10] were already successfully tested for the classification of positive, negative and neutral sentiment of Italian SM posts [16]. We used the ones described in [41] which include a manually created lexicon for Italian [185], two automatically translated sentiment polarity lexicons originally created for English [94, 185], an automatically created Twitter sentiment polarity lexicon and two word similarity lexicons automatically created using *word2vec*[11] [131], starting from two Italian corpora: (i) PAISÀ [122], a large corpus of authentic contemporary Italian texts; and (ii) a lemmatized corpus of 1,200,000 tweets, automatically collected.

In addition to these resources, we created two Word Embedding lexicons, to overcome the issue that lexical information in a short text can be very sparse. For this purpose, we trained two predict models using *word2vec*. These models learn lower-dimensional word embeddings. Embeddings are represented by a set of latent (hidden) variables and each word is a multidimensional vector that represents a specific instantiation of these variables. The first lexicon was built using a tokenized version of the itWaC corpus[12]. The itWaC corpus is a 2 billion word corpus constructed from the

[10] The lexicons can be downloaded from `www.italianlp.it`

[11] http://code.google.com/p/word2vec/

[12] `http://wacky.sslmit.unibo.it/doku.php?id=corpora`

Web, limiting the crawl to the *.it* domain and using medium-frequency words from *La Repubblica* corpus and basic Italian vocabulary lists as seeds. The second lexicon was built from a tokenized corpus of tweets. This corpus was collected using the Twitter APIs and it is made up of 10,700,781 Italian tweets.

The SVM classifier

The SVM classifier exploits a wide set of features, ranging across different levels of linguistic description. With the exception of the *word embedding combination*, these features have already been already used in sentiment polarity classification tasks [41] showing their effectiveness. The features are organised in three main categories: *raw and lexical text features*, *morpho-syntactic and syntactic features*, and *lexicon features*.

Raw and Lexical Text Features. *Number of tokens*: number of tokens occurring in the analyzed text; *Character n-grams*: presence or absence of contiguous sequences of characters in the analyzed text. *Word n-grams*: presence or absence of contiguous sequences of tokens in the analyzed text. *Lemma n-grams*: presence or absence of contiguous sequences of lemma occurring in the analyzed text. *Repetition of n-grams chars*: presence or absence of contiguous repetition of characters in the analyzed text. *Punctuation*: checks whether the analyzed text finishes with one of the following punctuation characters: "?", "!".

Morpho-syntactic and Syntactic Features. *Coarse grained Part-Of-Speech n-grams*: presence or absence of contiguous sequences of coarse-grained PoS, corresponding to the main grammatical categories (noun, verb, adjective). *Coarse grained Part-Of-Speech distribution*: the distribution of nouns, adjectives, adverbs, numbers in the text. *Fine grained Part-Of-Speech n-grams*: presence or absence of contiguous sequences of fine-grained PoS, which represents subdivisions of the coarse-grained tags (e.g., the class of nouns is subdivided into proper vs common nouns, verbs into main verbs, gerund forms, past particles). *Dependency types n-grams*: presence or absence of sequences of dependency types in the analyzed text. The dependencies are calculated with respect to *i)* the hierarchical parse tree structure and *ii)* the surface linear ordering of words *Lexical Dependency n-grams*: presence or absence of sequences of lemmas calculated with respect to the hierarchical parse tree. *Lexical Dependency Triplet n-grams*: distribution of lexical dependency triplets, where a triplet represents a dependency relation as (ld, lh, t), where ld is the lemma of the dependent, lh is the lemma of the syntactic head and t is the relation type linking the two. *Coarse Grained Part-Of-Speech Dependency n-grams*: presence or absence of sequences of coarse-grained Part-of-Speech, calculated with respect to the hierarchical parse tree. *Coarse Grained Part-Of-Speech Dependency Triplet n-grams*: distribution of coarse-grained Part-of-Speech dependency triplets, where a triplet represents a dependency relation as (cd, ch, t), where cd is the coarse-grained Part-of-Speech of the dependent, h is the coarse-grained Part-of-Speech of the syntactic head and t is the relation type linking the two.

Lexicon features. *Lemma sentiment polarity n-grams*: for each n-gram of lemmas extracted from the analyzed text, the feature checks the polarity of each component lemma in the existing sentiment polarity lexicons. Lemmas that are not present are marked

with the *ABSENT* tag. This is for example the case of the trigram "tutto molto bello" (*all very nice*) that is marked as "*ABSENT-POS-POS*" because *molto* and *bello* are marked as positive in the considered polarity lexicon and *tutto* is absent. The feature is computed for each existing sentiment polarity lexicon. *Emoticons*: presence or absence of positive or negative emoticons in the analyzed text. The lexicon of emoticons was extracted from `http://it.wikipedia.org/wiki/Emoticon` and manually classified. *Polarity modifier*: for each lemma in the text occurring in the sentiment polarity lexicons, the feature checks the presence of adjectives or adverbs in a left context window of size two. If this is the case, the polarity of the lemma is assigned to the modifier. This is for example the case of the bigram "non interessante" (*not interesting*), where "interessante" is a positive word, and "non" is an adverb. Accordingly, the feature "non_POS" is created. The feature is computed 3 times, checking all the existing sentiment polarity lexicons. *PMI score*: for each set of unigrams, bigrams, trigrams, four-grams and five-grams that occur in the analyzed text, the feature computes the score given by $\sum_{\text{i}-gram \in text} score(\text{i}-gram)$ and it returns the minimum and the maximum values of the five values (approximated to the nearest integer). *Distribution of sentiment polarity*: this feature computes the percentage of positive, negative and neutral lemmas that occur in the text. To overcome the sparsity problem, the percentages are rounded to the nearest multiple of 5. The feature is computed for each existing lexicon. *Most frequent sentiment polarity*: the feature returns the most frequent sentiment polarity of the lemmas in the analyzed text. The feature is computed for each existing lexicon. *Sentiment polarity in text sections*: the feature first splits the text into three equal sections. For each section, the most frequent polarity is computed using the available sentiment polarity lexicons. The purpose of this feature is to identify changes of polarity within the same text. *Word embeddings combination*: the feature returns the vectors obtained by computing separately the average of the word embeddings of the nouns, adjectives and verbs of the text. It has been computed once for each word embedding lexicon.

The LSTM classifier

The LSTM unit was initially proposed by Hochreiter and Schmidhuber [93]. LSTM units are able to propagate an important feature that came early in the input sequence over a long distance, thus capturing potential long-distance dependencies. LSTM is a state-of-the-art performer for semantic composition and it allows to compute the representation of a document from the representation of its words, with multiple abstraction levels. Each word is represented by a low dimensional, continuous and real-valued vector.

We employed a bidirectional LSTM architecture since it allows to capture long-range dependencies from both the directions of a document by constructing bidirectional links in the network [159]. In addition, we applied a dropout factor to both the input gates and to the recurrent connections in order to prevent overfitting which is a typical issue of neural networks [71]. As suggested in [71], we have chosen a dropout factor value in the optimum range $[0.3, 0.5]$, more specifically 0.45 for this work. Concerning the optimization process, categorical cross-entropy is used as a loss function and optimization was performed by the rmsprop optimizer [173].

To train the LSTM architecture, each input word in the text is represented by a 262-

dimensional vector which is composed by:
Word embeddings: the concatenation of the two word embeddings extracted by the two available Word Embedding lexicons (128 components for each word embedding, thus resulting in a total of 256 components), and for each word embedding an extra component was added in order to handle the "unknown word". *Word polarity*: the corresponding word polarity obtained by exploiting the Sentiment Polarity lexicons. This feature adds 3 extra components in the resulting vector, one for each possible outcome in the lexicons (negative, neutral, positive). We assumed that a word not found in the lexicons has a neutral polarity. *End of Sentence*: a component indicating whether or not the sentence was totally read.

4.3.2 Experiments and Results

We conducted two different classification experiments: the first considering the three different categories of hate (*Strong hate*, *Weak hate* and *No hate*) the second considering only two categories, *No hate* and *Hate*, where the last category was obtained by merging the *Strong hate* and *Weak hate* classes.

For the experiments, we used only documents that were annotated at least by three different annotators and where the most annotated class exists. This process resulted in two datasets: the three-class dataset, composed by 3,356 documents - divided into 2,816 *No hate*, 410 *Weak hate* and 130 *Strong hate* documents, and the two-class dataset, composed by 3,575 documents - divided into 2,789 *No hate* and 786 *Hate*. To balance the datasets, we selected a subset of the *No hate* texts, which was limited to the double size of the documents belonging to the *Weak hate* class in the three-class experiment and to the double size of the *Hate class* in the two-class one. To evaluate the accuracy of the two hate speech classifiers in the two experiments, we followed a 10-fold cross validation process: each dataset was randomly split into ten different non overlapping training and test sets. The overall *Accuracy*, *Precision*, *Recall* and *F-score* for each class were calculated as the average of these values over all the ten test sets. Accuracy, Precision, Recall and F-score are evaluation metrics employed in standard classification tasks. In our scenario: *Accuracy* measures how many comments are correctly identified in the classes; *Precision* measures how many comments, among those classified as expressing hate, have been correctly identified; *Recall* expresses how many comments, in the whole set, have been correctly recognized: a low recall means that many relevant comments are left unidentified and *F-score* is the harmonic mean of Precision and Recall.

Table 4.3 reports the results for the three-class experiment. Both SVM and LSTM are not able to discriminate between the three classes and this is particularly true for the Strong hate one. These results may be due to the small number of Strong hate documents (that is the class with the lowest number of documents) and the low level of annotator agreement. These results lead us to conduct the two-class experiment, whose accuracies are in Table 4.4. As expected, the results are much higher than those in the previous experiment. This is probably due to the higher number of Hate documents with respect to the Strong and Weak classes and to the higher annotator agreement with respect to the three-class experiments.

To evaluate the impact of the annotator agreement in the classification performances, we performed a last experiment, where we selected the documents for which at least

70% of the annotators were in agreement (321 Hate and 642 No-Hate documents). As Table 4.4 shows, more agreement yields an increasing accuracy for both the classification algorithms. This improvement is particularly significant for the classification of the Hate class, with F-score of about 72%. These results pave the way to the employment of our system in a real-use context. In addition, the outcome shows that this Hate Speech corpus, filtered with respect to the annotator agreement, allows to build automatic hate speech classifiers able to achieve accuracy in line with the ones obtained in mostly investigated sentiment analysis tasks for Italian, such as subjectivity and polarity classification [16].

		Strong hate			**Weak hate**			**No hate**		
Classifier	Accuracy (%)	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
SVM	64.61	.452	.189	.256	.523	.525	.519	.724	.794	.757
LSTM	60.50	.501	.054	.097	.434	.159	.221	.618	.950	.747

Table 4.3: *Ten-fold cross validation results on* Strong hate*,* Weak hate *and* No hate *classes.*

		Hate			**No hate**		
Classifier	Accuracy (%)	Prec.	Rec.	F-score	Prec.	Rec.	F-score
SVM	72.95	.625	.568	.594	.778	.817	.797
LSTM	75.23	.640	.6832	.657	.824	.791	.805
		$\geq$ 70% of Agreement					
SVM	80.60	.757	.689	.718	.833	.872	.851
LSTM	79.81	.706	.758	.728	.859	.822	.838

Table 4.4: *Ten-fold cross validation results on* Hate *and* No hate *classes.*

4.4 Discussion

So far we have described how SM content can be analyzed to spot hateful content. The hate speech classifier for Italian texts has plenty of application in SM contexts. Concerning the performance and considering a binary classification, the classifier achieved results comparable with those obtained in mostly investigated sentiment analysis tasks for Italian. Although there is some room for further improvements, we think that this approach has scored a point for the improvements of the communications on SM. We leave for future work the refinement of the classifier results when considering distinction (i) among hate levels (whereas the current classifier fails to achieve satisfactory results) and (ii) among different types of hate (which we defined in the paper and worked with at the annotation level). We will carry out a thorough analysis of the results of the two classifiers to investigate whether they can be combined in order to increase the performance. In addition, it is interesting to apply the proposed methodology to other SM platforms (e.g., Twitter) for research purposes, in order to verify whether the same assumptions on Facebook posts hold. The quality of the classification could be improved by enlarging the annotation process, both increasing the corpus size and collecting more annotations for a single comment. Besides, the effect of sarcasm affects the classifier performance, but this reveals to be hard to avoid since even humans have difficulties in understanding it. From the classification of single comments, the hate classifier may evolve to detect bursts of hate, thus preventing virtual discussions giving

rise to severe injuries to people and assets. Given that human moderators cannot monitor the huge user generated texts on SNSs, we believe this work represents the basis to track divergent states of Italian texts in online conversations.

CHAPTER 5

De-anonymization of Social Media content

With billions of users, social media probably represent the most privileged channel for publishing, sharing, and commenting information. In particular, social networks are often adopted to spread news content [119]. According to a Pew Research study, Americans often gets news online, with a double share of them preferring social media rather than print magazines[1]. As a matter of fact, popular newspapers have an official account on social platforms. Through their pages, news stories - or their previews - are published - often under the form of a short post, with a further link to the complete text. The readers' community can like, share, and re-post news stories. Users can also comment and discuss issues in the news themselves. Still, a Pew Research survey highlights that a share of 37% of social media news consumers comments on news stories, while the 31% "discuss news on the news in the site"[2]. Undeniably, users' comments and discussions may help to increase the awareness and value of the published information, thanks to the addition of informative details. Examples of support are, for example, the depiction of the context in which the news facts took place, or to track down mistakes, draw rectifications, and even unveil fake information.

In this chapter, we focus on online news articles and, in particular, on those news portions that organisations choose not to make public. The approach described here is derived from [180].

News censorship may occur for different reasons. Organisations, be them military, commercial, governmental, or judicial, may decide to veil part of the information to protect sensitive data from, e.g., competitors, customers, or hostile entities. Standard examples of censored data are identities: from a business point of view, a press agency

[1]The Modern News Consumers, a Pew Research study: http://www.journalism.org/2016/07/07/pathways-to-news/, July, 7, 2016 – ; All URLs in this Chapter have been lastly accessed on February, 20, 2018.

[2]10 facts about the changing of the digital landscape: http://www.pewresearch.org/fact-tank/2016/09/14/facts-about-the-changing-digital-news-landscape/, September 14, 2016.

may veil the identity of the buyer of a huge amount of fighters. Also, the names of the victims of particularly hateful offences, like rapes and abuses on minors, are typically obfuscated, as for regulations dictated by law. Finally, a peculiar practice when publishing Israeli military-related news on social media is the veiling of the identities of public officers (e.g., Corporal S., rather than the explicit identity of such officer, see, e.g., [31]). However, as highlighted by recent literature [160], given the essential nature of social networking, the "non identification alone is ineffective in protecting sensitive information". This is due to the fact that, featuring a *commented post* structure of the published news, a specific information, withheld in the news, is compromised through the effects of users' comments, where specific content may reveal, either *explicitly* or *implicitly*, that information.

This work places itself amongst a few ones, like, e.g., [25, 31] that investigate to which extent the connections among news articles, comments, and social media influence the effectiveness of identities censorship procedures. In particular, we present a novel approach to unveil a censored identity in a news post, by exploiting the fact that, on the social Web, it is not unusual to find the same content, or a very similar one, published elsewhere, e.g., by another publisher with different censorship policies. Also and noticeably, as discussed above, the amount of user generated content on social networks may lead to the very unexpected phenomenon according to which the hidden information may emerge in the users' comments.

Differently from prior work in the area, which exploits the friendship network of the commenters to some censored news, here we inherit from the field of text analysis. In particular, we exploit techniques often used to address co-reference resolution [42], based on recognising the context in which certain names tend to appear, to successfully address the task of unveiling censored names. To the best of our knowledge, this is the first attempt that addresses the task by means of a semi-supervised approach, which only makes use of texts, without relying on metadata about the commenters, and trying to reconstruct missing information exploiting similar contexts. For running and validating our analysis, we make use of Facebook data, which we *purposely* censored for the sake of the experiments. Even if we rely on an experimental setting that is built *ad hoc*, our synthesised scenario is easily connectable to real use cases, as described in subsequent sections.

Our extensive experimental campaign explores a dataset of almost 40,000 posts published on the Facebook pages of the top 25 US newspapers (by weekday circulation). By exploiting an algorithm based on context categorisation, we train a classifier on the posts, and related comments, in the dataset, to demonstrate the capability to reveal the censored term. The system performances are benchmarked against two baselines, obtaining a more than significant improvement.

Summarising, the Chapter contributes along the following dimensions:

- the design and development of a methodology based on text-analysis, here applied for the first time to spot identities that have been censored in social media content;

- the proposal of an approach that is solely based on very loosely structured data, in contrast to other proposed techniques that leverage the social network structure. The latter have the issues that 1. the association between names and social network nodes needs to be addressed, and 2. the structure of the social network constitutes

significant a-priori knowledge. Instead, we simply use raw data, by only assuming a “commented post" structure of the data;

- starting from revealing censored popular identities, our results constitute the prelude to the detection of other kind of censored terms, such as, e.g., brands and even identities of common people, whose veiling is a usual practice often applied by publishers for privacy issues, be them driven by legal, military, or business motivations.

Here we consider SNs as a relevant case study due to their coverage. However, our methodology is general enough to be applied to various data sources, provided there is a sufficient number of training examples. In the next section, we first introduce real identity censorship procedures, discussing the role of comments - and commenters - in bypassing their effectiveness. In Section 5.2, we discuss related work in the area of investigation. Section 5.3 presents the data corpus for our analyses, also highlighting similarities of such corpus with the real scenarios presented in Section 5.1. Section 5.4 presents the methodology, and Section 5.5 describes the experiments and comments the results. Finally, Section 5.6 presents a discussion about the approach and draws a conclusion about the potential of this solution.

5.1 Identities censorship in online news and its circumvention

Moving from motivations for identities censorship in news, this Section discusses the effectiveness of such censorship when news are published online [160].

Traditionally, the censorship of identities in news occurs for three main reasons: 1) business, since, e.g., it may be not advantageous to disclose the real identity of a participant in a commercial transaction, such as a large quantities of weapons; 2) legal (e.g., do not become known minors abused); and 3) military, to protect the individuals, and their relatives, from being identified by adversaries. As an example, in Israeli “policy dictates many situations in which the identity of officers must not be released to the public [160]. In the above circumstances, the censorship usually takes place either by putting an initial or using a fancy name.

With the advent of the social media era, the publication of news on social networks sites became a usual pratice. Thus, when the news is published on social networks sites, such as on the Facebook pages of newspapers, the identities are still blurred as written above, directly from the organisation that chooses not to publish that information (therefore, either the government, or the news agency that publishes the news, or some other military or commercial stakeholder).

However, when the news post is on the social network, and a “commented post" structure of the data is followed, the comments are freely posted by users other than the news publisher. Also, comments are generally not moderated by the platform administrators, unless they are reported as offensive content, inciting hate campaigns, or some judicial authority required the cancellation of specific comments.

The fact that there are a number of uncensored comments leads to the phenomenon that, although in the post the information is withheld, that information is compromised by one, or more, comments. In fact, it has been proven that, although the organisation that posted the news has censored an identity in the news itself, and so published it, those people who know the censored name and who make comments tend to talk about

the name, indirectly or even directly. This is the case featured, e.g., by the Facebook dataset analysed in [160], where 325,527 press items from 37 Facebook news organisation pages were collected. A total of 48 censored articles were identified by a pattern matching algorithm first, and then manually checked. On the whole amount of comments tied to those articles, the 19% of them were classified as comments presenting an explicit identification of the name or the use of a pseudonym. A de-censorship analysis based on the social graph of the commenters has been carried out to recognise the censored names [31]. In the rest of the Chapter, we will propose a methodology based instead of recognising the textual context in which certain terms tend to appear.

To test the methodology, we rely on a Facebook dataset that we intentionally censored, by however resembling the real scenarios depicted above. We deliberately censored identities in a Facebook dataset of US newspaper posts, leaving comments to posts unchanged. The dataset is described in the Section 5.3.

5.2 Other works on censorship

Social Media provide Internet users with the opportunity to discuss, get informed, express themselves and interact for a myriads of goals, such as planning events and engaging in commercial transactions. In a word, users rely on online services to say to the world what they are, think, do; and, viceversa, they learn the same about the other subscribers. For a decade, scientists have been evaluating and assessing the attitude of users to disclose their personal information to receive higher exposure within the network community [111, 121].

Both sociologists and computer scientists have investigated the amount and kind of personal information available on social networks. As an example, work in [193] represents one of the first studies on identity construction on Facebook (comparing the difference in the identities narration on the popular social network with those on anonymous online environments). Two years later, the authors of [138] studied the amount of personal information exposed by Facebook, characterising it according to the account age (the younger the user, the more the personal information exposed) and the inclination to set up new relationships. Thus, despite the enormous volume of daily communications, which leads to levels of obfuscation and good resistance to analysis techniques [45], social networks naturally offer a huge amount of public information – even redundant – with its fast diffusion supported by influencers and even weak ties among users [14, 81]. Recently, researchers also concentrated on misinformation spread, considering the dynamics and motivations for the large number of followers of fake news [22]. The demonstrated tendency of people to disclose their data, despite privacy issues [133], has let researchers argue that, where the structure of data is under the form of a commented post, the content of comments may reveal a lot about the post itself, in those cases in which parts of them have been obfuscated. Indeed, several approaches have been tested in the literature to automatically classifying which comments lead to leakage of information. Some of these approaches exploit discourse analysis, to “examine how language construct phenomena" [144] or semiotic analysis, which concerns the study of signs to infer the “deeper meaning" of the data [80]. In [160], the authors showed how to leverage discourse and semiotic analysis, in conjunction with standard text classification approaches, to automatically categorise leakage and

non-leakage comments.

The issue of censorship on the Web has been faced from different perspectives: in the early 2000's, some information leakage was already possible to circumvent national censorship [67, 68]. In this work, we specifically consider censored texts. Work in [162] proposes a method to make textual documents resilient to censorship sharing them over a P2P network, one of the most frequently used decentralised sharing mechanism. However, when considering centralised systems, like Facebook, censorship might occur on document instances (i.e., posts and comments). Thus, strategic information may be altered - or censored- by malicious users of P2P networks, as well as by authors of single posts on social media.

Regarding news, work in [126, 160] characterises those comments exploitable for revealing censored data. In [31], the authors consider comments to Facebook posts about Israeli military news. While specific identities in the post contents are censored, through the analysis of the social graph of the commenters [25], the authors were able to spot the identity of the mentioned people. In particular, the approach exploits accounts' public information to infer the ego network of an individual and tries to reconstruct it when the access to user data is restricted from Facebook API, assuming that the account of the censored identity is linked to one of the commenters [30]. Thus, the approach is effective when the commenters are friends of the target, even in the case that comments are few. However, it might be less effective in general scenarios, in which commenters are not friends of the target of the censorship (like, e.g., when the censored identity is a popular one). Also, leveraging the social network structure, a significant a-priori knowledge is needed. The work in [164] describes an effective approach to de-anonymize a social network using a random forest classifier and the number of friends of each node in the network as a feature.

In a past work, the authors showed how to leverage a semi-supervised analysis approach to detects drugs and effects in large, domain-specific textual corpora [60]. Here, we inherit that snippets and contexts classification approach, to propose a methodology solely based on very loosely structured data. As clarified in the rest of the Chapter, we reveal identities in censored Facebook posts, only relying on a corpus made of very short texts, some of them even irrelevant to the domain of the post where the censored name is. This approach, although conceptually not far from what is proposed in [85, 86] for author disambiguation, adopts textual contents rather than other metadata to link information. Disambiguation of identities on social networks is the matter of investigation in [152, 153], which exploits semantic social graphs to disambiguate among a set of possible person references. The last approach is somehow in the middle between the de-anonymization technique proposed in [25] and the semantic analysis performed in [60], although the knowledge extraction is mainly performed for different purposes.

Compared with previous work with similar goals, the current proposal differentiates because it does not rely on the social graph of the commenters to recognise the censored term. Instead, the Chapter proposes an adaptation - and application - of a text-analysis approach to the issue of unveiling censored identities. The approach is tested on a synthesised scenario, which however resembles a real use case.

Furthermore, it is worth noting that, although a series of work consider automatic classification of comments to detect those ones leading to leakage of information, we decided here to bypass such a classification, and to consider the whole set of identi-

ties in the comments dataset. Thus, we ran a Name Entity Identifier to recognise the terms representing identities, and we directly pass to launch our methodology: this to distinguish, amongst the set of identified candidates, the censored term.

For the sake of completeness, we acknowledge the occurrence of different kinds of censorship, from DNS to router level ones [176]. Differently from veiling single terms, an entire domain might not be accessible, thus requiring different approaches to circumvent the block [36]. Work in [77, 115, 116] propose a survey on different topics related to censorship, either on detection or possible countermeasures. Also, monitoring tools exist, to trace the diffusion of the phenomenon [135, 163]. Finally, emails or other communication channels different from social media might be affected by censorship [139].

Work	Target	Technique
Lindamood [121]	Explore how to launch inference attacks using released social networking data to predict undisclosed private information about individuals.	Naïve Bayes classifier applied to profile traits and friendship links.
Lam [111]	Study the involuntary information leakage in social network services with regards to Wretch.	Analysis of real names, age, and gender of profiles from friend annotations.
Nosko [138]	Analysis of disclosed information in Facebook profiles.	Study of categories of features and information extracted from profiles.
Schwartz [160]	Study of the interplay between online news, reader comments, and social networks to detect and characterize comments leading to the revelation of censored information.	Manual analysis of comments related to censored news on SM.
Burattin [25]	Show Facebook privacy vulnerabilities about personal information and friends through the application SocialSpy.	Exploitation of information leakage from Facebook profiles through friends' profiles. In this way, it is possible to detect up to 70% of victim's friends.

Feamster [67]	Build of Infranet, a system that enables clients to surreptitiously retrieve sensitive content via cooperating Web servers distributed across the global Internet when countries and companies routinely block or monitor access to parts of the Internet.	Infranet uses a tunneling protocol that provides a covert communication channel between its clients and servers. Infranet clients send covert messages to Infranet servers by associating meaning to the sequence of HTTP requests being made. Infranet servers return content by hiding censored data in uncensored images using steganographic techniques.
Feamster [68]	Face the proxy discovery problem by censors.	Authors propose to separate the proxy into two distinct components: the messenger, which the client discovers using keyspace hopping and which simply acts as a gateway to the Internet; and the portal, whose identity is widely-published and whose responsibility it is to interpret and serve the client's requests for censored content.
Serjantov [162]	Proposition of a new Peer-to-Peer architecture for a censorship-resistant system with user, server and active-server document anonymity as well as efficient document retrieval.	Separate the role of document storers from the machines visible to the users, which makes each individual part of the system less prone to attacks, and therefore to censorship.
Cascavilla [31]	Study of information leakage through discussions in online social networks. The work focuses on articles published by news pages, in which a person's name is censored, and examines whether the person is identifiable.	Analysis of comments and social network graphs of commenters.
Sharad [164]	Automated approach to re-identifying nodes in anonymised social networks.	Machine learning (decision forests) to matching pairs of nodes in disparate anonymised sub-graphs.

Verkamp [176]	Explore the mechanics of Web censorship in 11 countries around the world, including China.	Analysis of triggers of the censorship the different effects in many countries.
Nabi [135]	Study of the cause, effect, and mechanism of web censorship in Pakistan.	Use of a publicly available list of blocked websites and check their accessibility from multiple networks within the country.
Sfakianakis [163]	Design and implementation of a web censorship monitor, called CensMon.	CensMon conducts extensive accessibility tests. It consists of two basic building blocks: the central server and the network of sensing nodes.
Del Vigna [180]	Text analysis approach to **unveil censored identities** in news and other documents.	**Redundant and leakage information exploitation** to build a context classifier for censored names, with no knowledge of the SN structure.

Table 5.1: *List of main works in the field of de-censorship and information leakage on the Internet and SM compared to this Thesis.*

For convenience we summarised all major approaches to de-censorship and de-anonymization in Table 5.1, highlighting authors and the techniques adopted.

5.3 Dataset of US newspapers Facebook pages

In this work, we leverage the fact that many newspapers have a public social profile, to check whether an identity - that was censored in some news - can be spotted by analysing comments to that news - as well as other news published on different online newspapers. We consider a set of posts and comments from the Facebook pages of the top 25 newspapers, by weekday circulation, in US[3]. Table 5.2 shows the names of the newspapers, the corresponding Facebook page (if any), and the number of collected posts and comments, crawled from the pages.

We developed a specific crawler to collect the data. The crawler is written in the PHP scripting language, using the Laravel[4] framework, to make it scalable and easy manageable. In particular, the crawler exploits the Facebook Graph API[5] to collect all the public posts, comments, and comments to comments from the Facebook pages of the newspapers. The collection requires as input the URL of the Facebook page and a set of tokens required to authenticate the application on the SN. The crawler supports parallel downloads, thanks to its multi-process architecture. It recursively downloads

[3] https://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_States
[4] https://laravel.com
[5] https://developers.facebook.com/docs/graph-api

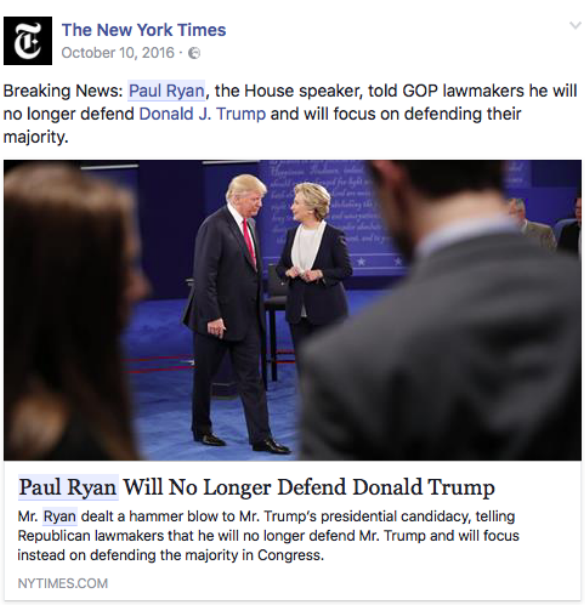


Figure 5.1: *Example of a post published on the NYT Facebook page.*

data, until it fully covers the time span specified by the operator. It stores data in the JSON format, since it is the most natural format for SM data. Also, such format can be easily employed to feed storage systems like Elasticsearch[6] or MongoDB[7].

We collected posts and comments from August 1, 2016 to October 31, 2016. Overall, we obtained 39,817 posts and 2,125,038 comments.

Figure 5.1 reports an example of a typical post, published on the Facebook page of one of the most influential American newspapers, The New York Times. As shown in the figure 5.1, the text of the post is short and not very informative since the actual content of the news is in the link to the journal website. In that very short piece of text, two identities are mentioned, Paul Ryan and Donald J. Trump. Figure 5.2 shows an example list of some comments related to the same post. Comments are usually copious, with discussions that are focused on several aspects. Notably, not all the comments are strictly related to the topics of the post. As we can see from the figure, more than one identity is mentioned in the comments set (i.e., Donald Trump, Hillary Clinton, Ben Carson, Paul Ryan, Abraham Lincoln, and McConnell) among which the two mentioned in the related post. Also, references to some of the identities are with different variants (e.g., Ryan, Paul Ryan, Trump, Donald Trump).

It is worth noting that comments are generally short, thus comparable to microblog messages, e.g., tweets or Tumblr posts.

For the sake of the experiments illustrated in the next section, we purposely censored the identities of some people in part of the crawled Facebook posts, to simulate a censorship behaviour.

In the following, we will propose a methodology to recognise the censored name among a set of candidate names as they appear in comments to the censored post. Among the candidate names, there is the same name as the one in the censored post

[6] https://www.elastic.co
[7] https://www.mongodb.com/it

Newspaper	Facebook profile	Posts	Comments
The Wall Street Journal	wsj	1577	136969
The New York Times	nytimes	265	98099
USA Today	usatoday	560	155893
Los Angeles Times	latimes	532	124477
San Jose Mercury News	mercurynews	0	0
New York Daily News	NYDailyNews	1637	124948
New York Post	NYPost	479	132715
The Washington Post	washingtonpost	232	101260
Chicago Sun-Times	thechicagosuntimes	2215	64675
The Denver Post	denverpost	1376	113621
Chicago Tribune	chicagotribune	2401	141361
The Dallas Morning News	dallasmorningnews	2458	148154
Newsday	newsday	2432	60549
Houston Chronicle	houstonchronicle	1350	920
Orange County Register	ocregister	1123	37153
The Star-Ledger	Star.Ledger	284	3142
Tampa Bay Times	tampabaycom	1539	76388
The Plain Dealer	ThePlainDealerCLE	4	33
The Philadelphia Inquirer	phillyinquirer	2124	10491
Star Tribune	startribune	2820	106357
The Arizona Republic	azcentral	2073	151590
Honolulu Star-Advertiser	staradvertiser	3487	52447
Las Vegas Review-Journal	reviewjournal	3588	108614
San Diego Union-Tribune	SanDiegoUnionTribune	2163	45530
The Boston Globe	globe	3098	129652

Table 5.2: *Top 25 US newspapers by weekday circulation (as of March, 2013).*

(e.g., Trump in the censored post, Trump in the comments). Remarkably, since our analysis is based on similarities of the *surroundings* of a name, rather than on the name itself, the proposed approach is still valid also when the censored identity is referred in comments with a pseudonym (e.g., Trump in the censored post, Mr. President in the comments). This last case often happens in real scenarios, as introduced in Section 5.1.

5.4 Methodology

The text analysis approach proposed in this Section to unveil censored identities is motivated by the hypothesis that, in the collaborative web, a “perfect" censorship is impossible. Indeed, the redundancy of information (e.g., the same fact reported by multiple sources, many of which do not apply the same restriction policy for publishing that fact) gives us a means to figure out who the subject of a censored post is. Obviously, as shown in the post and comments of Figures 5.1 and 5.2, the same news reported and/or commented by various sources will not use the same exact wording (see, for example, Donald J. Trump, Donald Trump, Trump), so that the problem remains a non trivial one.

5.4.1 Overall methodology

The proposed approach makes use of two distinct Named Entity Recognizers (NER). Named Entity Recognition (NER) is the process of identifying and classifying entities within a text. Common entities to identify are, e.g., persons, locations, organizations,


[redacted] PLEASE READ:
Mr. Ryan,
When you give up on Donald Trump, you give up on all of us people of the Republican party. How fair is that? Yes there are some skeletons in his closet, but OMG the Clinton's, there are more then skeletons on theirs. Nobody is perfect and if Hilary Clinton wins, we are doomed. Maybe if you have a conversation with Dr. Ben Carson, he will help you understand, you of all people, we are Republicans. If we don't still keep together we won't win and our Country will not what it once was. Please get back on the Trump train and help us win the election. Please don't let Hilary win. We need your help!!! Thank you.
Like · Reply · 6 · October 11, 2016 at 6:36am
2 Replies
[redacted] Paul Ryan you should be ashamed of yourself you're supposed to be standing up for the Republican Party the people of this United States have you never ever in your life said something inappropriately maybe just didn't get televised he hasn't raped anybody hasn't murdered anybody he was inappropriate with his language is that make it right now but you should not turn your back on the candidate that the Republican people of this country has nominated you should be ashamed of yourself
Like · Reply · 7 · October 10, 2016 at 9:34pm
[redacted] Your non-support of Mr. Trump is like looking for any excuse not to do what "we,the people" have clearly supported. A far better choice than Hilary Clinton. And Trump is right, his words are not actions. Her actions make her deplorable. We have lost respect for Speaker Ryan. He buys into the crazy left-wing media frenzy instead of bringing a reasoned dialog to the world of politics.
Like · Reply · 3 · October 11, 2016 at 6:36am
[redacted] Don't know why these Republicans all of a sudden are jumping ship. Trump just exposed the true ugliness of their policies. The Republican attitude towards women's rights and reproductive choice is just one huge [redacted] grab. Why bother. There is no dignity to salvage for this party. Clinton's debate performance was pretty lame (Abraham Lincoln??!!!) but Feel have no choice but her in this dismal spectacle. Neither of them take any responsibility for what they say and do.
Like · Reply · 4 · October 10, 2016 at 7:23pm
2 Replies
[redacted] He surely must know that if the GOP doesn't fund Trump's campaign and Hillary wins, they will probably lose the house and Senate as well. People are not going to vote for the RINOs who have helped the liberals stay in control.
Like · Reply · 2 · October 11, 2016 at 7:22am
[redacted] I'm exhausted from supporting McConnell and you in fact all the GOP for NOT going after the Dems , the way the Dems go after us. I will not be breaking the chain of registered Republicans in my family, and to think my grandmother who was the GOP city clerk of our home town , won 9 elections , most of the time with Dem mayors BUT I am now to old to deal with your establishment games
Like · Reply · 3 · October 10, 2016 at 9:47pm
[redacted] I hope Paul Ryan has a week to read all of these comments. It would be to his benefit to read every word. He needs to know how conservatives feel about his decision to throw Donald Trump under the bus along with our party.
Like · Reply · 2 · October 11, 2016 at 8:08am

Figure 5.2: *Excerpt of comments tied to the NYT post in Fig. 5.1.*

dates, times, and so on. NER state-of-the-art systems use statistical models (i.e., machine learning) and typically require a set of manually annotated training data, in combination with a classifier. Popular NERs frequently adopted in the literature are the Stanford NER tagger[8], also available through the NLTK Python library[9], and the spaCy NER (the one adopted in this Chapter, introduced in Section 5.5.1). Mostly based on machine learning techniques, NERs exploit features such as strong indicators for names (e.g., titles like "Ph.D.", "Mr.", "Dr.", etc.) to determine whether a small chunk of text (say, a window of 50-200 characters around the name) indicates a person, an organization, a date, and so on. NERs are to this day standard pieces of an NLP pipeline; we plan here on showing how to build on top of one to address our specific task. We first

[8] https://nlp.stanford.edu/software/CRF-NER.shtml
[9] https://www.nltk.org

apply a generic NER to the whole collection of data depicted in Table 5.2, to detect the identities within. Then, by extending the NER scope to a wider context, we exploit a second, more specific entity recognizer, hereafter called Candidate Entity Recognizer (CER), to recognize specific people (out of a smallish set of candidates).

Let's suppose that, in a piece of text (like a Facebook post) the name "Mark Smith" has been censored by replacing it with "John Doe". The text might contain sentences such as: "during his career as a drug dealer, John Doe was often in contact with the Medellín Cartel". In order to reveal the original name hidden by John Doe, we will proceed as follows:

1. in the text, we will identify names to "resolve" (John Doe in the example), using the generic NER;

2. still relying on the generic NER, we will then identify a set of candidate names for the actual identity of John Doe, among which the correct one ("Mark Smith"), by searching in a textual data collection wider than the single piece of censored text. For the sake of this simple example, let us assume that we have "Mark Smith" and "Mary Jones" as possible candidates;

3. we will obtain a set of sentences that include the candidates, e.g., "The notorious drug baron Mark Smith used to spend much time in Medellín" and "Mary Jones's career as police officer was characterized by her fights with the drug dealers";

4. using the sentences retrieved at the previous step, we will train a customized version of a NER, i.e., the Candidate Entity Recognizer (CER), to identify instances of Mark Smith and Mary Jones, discarding the actual name from the text. In practice, we will train the CER with sentences like: "The notorious drug baron <MarkSmith> XXXX XXXX </MarkSmith> used to spend much time in Medellín" or "<MaryJones> XXXX XXXX </MaryJones>'s career as police officer was characterized by her fights with the drug dealers";

5. finally, we will apply the CER model thus obtained to the censored sentences in the original text, to find out who they refer to. In practice, by feed the sentence: "during his career as a drug dealer, XXXX XXXX was often in contact with the Medellín Cartel", we expect the trained CER to return "Mark Smith" rather than "Mary Jones".

For the sake of a more compact representation, we synthesise the algorithm as follows:

Data: $dataset, target$
Result: $candidate$
$identities \leftarrow \emptyset$;
$namedEntities \leftarrow \emptyset$;
$contexts \leftarrow \emptyset$;
$CER \leftarrow \emptyset$;
for $text \in dataset$ **do**
  Find all names in $text$ with the NER and put them in $namedEntities$;
  for *each* $namedEntity$ *found* **do**
    Put all sentences from $text$ that contains the $namedEntity$ in $contexts[namedEntity]$, substituting $namedEntity$ with XXXX XXXX in the sentences;
  end
end
for $namedEntity \in namedEntities$ **do**
  Train $CER[namedEntity]$ using $contexts[namedEntity]$;
end
Apply $CERs$ to the $target$ and check which named entity better fits.

Algorithm 1: How to train a *CER* to spot censored identities.

It is worth noting that the application of the algorithm to the whole collection would be costly, due to the large amount of text. To shorten the procedure, we restrict its application to only those posts having a meaningful number of occurrences of the candidate name in the comments, as detailed in Section 5.5.

5.5 Experiments and Results

Here, we actualize the methodology in 5.4.1 for the Facebook dataset described in Section 5.3. In that dataset, the data reflect our hypothesized structure of "post & comments": a post published onto a platform that accepts un-moderated (or, at least, not heavily moderated) comments.

Clearly, the original posts in the dataset are not censored: the names have not been altered, it being a condition that we need in order to check for the correctness of our decensorship system. However, in order for us to work in a simulated censored condition, we need to first pre-process the dataset, by removing specific target names occurring in the source post. In order to do this, we rely on a generic Name Entity Recognizer to detect the names, which we then replace by unique random strings, to give the following processing steps no hints as to who the removed names were.

Therefore, to run the experiments according to the methodology defined in Subsection 5.4.1, we implement the steps listed below:

1. selecting a target name that is the object of censorship in a Facebook post. We choose the name among a set of *popular* ones, specifically selected for the experiment, for which we expect to find enough data (Subsection 5.5.1);

2. retrieving a sub-set of the posts containing the target name, which will constitute the corpus that will be subject of censorship (Subsection 5.5.2);

3. censoring such posts: the target name is removed from that sub-set of posts and it is replaced by a random string (pre-processing phase, Subsection 5.5.2);

4. applying a NER to the comments tied to the sub-set of posts, to extract candidate names (Subsection 5.5.3);

5. filtering the candidate names, so that only k candidates remain (Subsection 5.5.4);

6. searching for the candidate names in the whole collection – all the posts and comments in the dataset described in Section 5.3 – except for the set of posts from which the target name has been removed (Subsection 5.5.5);

7. training a specific Candidate Entity Recognizer (CER), so that, instead of a generic "person" class, it is able to distinguish occurrences of the various candidates. The CER is trained on the data retrieved at step 6 (Subsection 5.5.6);

8. applying the CER for the k candidates to the censored name in the original set of posts, to see whether the classifier is able to correctly recognize the name that was censored (Subsection 5.5.7);

9. evaluating the results in terms of standard measures, by comparing the decisions taken by the system *vs* the name actually removed from the posts (Subsection 5.5.8).

Below, we provide details of the implementation of each step.

5.5.1 Selecting names

Names to test the system against were selected relying on a *popularity* criterion. In fact, we need names that are 1) present in the collection and 2) popular enough to appear in a certain amount of posts and comments. The latter requirement might appear as a strict constraint, but on SM data is often redundant.

To obtain the list of such names, we run a pre-trained NER over the whole data collection, retrieving all occurrences of person names. For this task, we use the spaCy[10] Named Entity Recognizer in its pre-trained form. We make this particular choice because we need a NER that comes in pre-trained form, to ease our task, and features the capability of being re-trained easily over new examples, for our purposes. SpaCy is an open source software library for the Python language, entirely devoted to natural language processing. It includes several pre-trained models. The spaCy website reports an F-score for the used English NER of 85.30[11].

The result of the NER application is a list of more than 1,000 terms, with their occurrences in the whole data collection. We first polish the list, by removing terms erroneously recognized by the NER as person names (examples are "Facebook", "Wikileaks", "Twitter", and "Google"). Then, we keep only those names occurring more than 100 times. The result, consisting of 149 names, are listed in Table 5.8 and Table 5.9. For readability, we show here just a short excerpt of the two tables (Table 5.3) and move the whole collection in 5.7.1. Politicians turn out to represent the most present category

[10] https://spacy.io/docs/usage/entity-recognition
[11] https://spacy.io/models/en

in the collection (64 names). The other names, which mostly include journalists, entrepreneurs, activists, lawyers, athletes, TV stars and actors, are grouped together under the generic label “Celebrities" (85 names).

Politicians	Freq	Celebrities	Freq
Hillary Clinton	23615	George Soros	1173
Donald Trump	17913	Margaret Sanger	600
Bill Clinton	6668	James Comey	585
Gary Johnson	3153	Paula Jones	566
Michelle Obama	1079	Billy Bush	554
John McCain	1079	Monica Levinsky	438
Bernie Sanders	890	Colin Kaepernick	401
Paul Ryan	863	Julian Assange	373
Mike Pence	859	Melania Trump	362
Barack Obama	782	Saul Alinsky	361
Mitt Romney	328	Ryan Lochte	320
John Podesta	323	Steve Bannon	321
Huma Abedin	320	Bill Cosby	314
Sarah Palin	303	Seymour Hersh	304
Rudy Giuliani	281	Ben Carson	283
Rick Scott	248	John Kass	280

Table 5.3: *Excerpt of politicians and celebrities with more than 100 occurrences.*

5.5.2 Retrieving and censoring the posts with the target names

To retrieve all posts containing a certain name, we indexed the whole collection using Apache SOLR indexing and search engine[12]. SOLR is a very popular keyword-based search engine system. Its core functionality is to index a text, and retrieve it by keywords (though more sophisticated means of retrieval are available). Its task is essentially defined as “retrieve all and only the documents matching the query". The way we used in this Chapter, SOLR is responsible for:

1. retrieving all documents containing a string (it is reasonable to expect that any bugless retrieval system performs this task with a 100% accuracy);

2. returning the window of text around the search string (same consideration as above).

This was an enabling means for us to effectively retrieve all documents containing a name and then apply our methodology – SOLR has no notion of the meaning of that name, so searching for “John Wick" might return documents referring to the movie, or documents referring to a person by that name. Should SOLR be replaced by any alternative (ElasticSearch[13], or MySQL[14], or even `grep`'ping files), no changes to the algorithm or its accuracy would be observed.

Clearly, searching for the occurrence of a name (e.g., “Donald J. Trump”) does not guarantee that all the references to the same person are returned (a person might be referenced by aliases, nicknames, etc.), but, in its simplicity, it provides the highest possible precision, at the expense of recall, and it makes little difference for our purposes. Indeed, we will only consider those texts that are, in fact, returned. Furthermore,

[12] `http://lucene.apache.org/solr/`
[13] `https://www.elastic.co/`
[14] `https://www.mysql.com/`

using SOLR actually accomplishes a second feat to us. Since we will build a "custom" NER (i.e., our CER – Candidate Entity Recognizer) on the immediate surrounding of names, we do not need the posts with the name in their whole form – but just a window of text surrounding the occurrences of the name. Therefore, we can use SOLR's *snippet* feature to immediately retrieve a chunk of a few words surrounding the name instance. We asked for snippets of 200 characters and ignored all the snippets shorter than 50 characters. The choice of this particular length is due to the fact that it is coherent, e.g., with the length of a tweet or a Facebook comment. From these snippets, we removed the actual name, replacing it with a token string composed of random letters (all unique to the snippet), with the first letter capitalized. This simulates our censorship of the snippets. These censored snippets are those bits of texts out of which we will try to reconstruct the original name.

It is worth noting that the target names are not replaced in the whole corpus of posts. Instead, we set as 20 the maximum number of posts where the name is replaced. In many cases, the name was present in less than 20 posts. The threshold has been chosen after diverse attempts, and 20 is the one that guarantees the best performances of our approach.

5.5.3 Retrieving candidates from comments

For each SOLR resulting document censored so far, we retrieve the relative comments by exploiting the association between posts and their comments. In the comments, as per our assumption, we expect to find many names, among which the one we removed earlier from the snippet. To obtain a list of all names in the comments, we run the spaCy NER on them. The application of the NER to the comments produces, as a result, a list of names that we consider suitable candidates to fill the spot of the original name previously removed from the text snippet.

5.5.4 Filtering candidates

An initial test showed that the candidates retrieved in 5.5.3 were often too many. Since we are going to train a specific Candidate Entity Recognizer (CER) to recognize them, we need to produce training examples for each of them, possibly a lengthy task. Therefore, we select only k candidates from the list of all those returned at the previous step. The selection criteria is simple: we select the k most frequent candidates found in the comments. It is worth noting that considering the k most frequent candidates might not include the actual name we are looking for. Thus, we always include the actual name within the k filtered candidates. We remark that, even if we include the actual name in the k candidates regardless of its frequency among the comments, we verified that the name actually appears in the same comments. This preserves the fairness of the approach.

This also gives us a convenient baseline to compare our system against: what if the actual name is always the first one in the list of candidates? We would have solved the problem at this stage without further ado. In Section 5.5.8, we will compare the performance of our system against this naïve baseline solution.

5.5.5 Fetching examples for each candidate

After filtering the list of candidates, we need to figure out the typical context in which each of the candidates occurs. Clearly, the mentions in the comments tied to the censored post are a starting point, but we can use more examples. Still using SOLR, we tap into our dataset, searching for each candidate name and retrieving all the snippets in which it appears, them being relative to both posts and comments of the whole collection (excluding the original posts that we retrieved and censored in Section 5.5.2). It is worth noting that we make no attempt at reconciling the names. Thus, there is the possibility to obtain different sets of examples for 2 different names that might actually refer to the same person (e.g., "Donald J Trump" and "The Donald"). This might have the disadvantage of spreading our training set that becomes too thin. In fact, still considering "Donald J Trump" and "The Donald", we could have two sets, one for "Donald J Trump" and the other for "The Donald"; furthermore, obviously, the two sets would be smaller than their union. However, it is of paramount importance to act in this way, to avoid the infusion in the system of a-priori knowledge beyond the data. It could also be the case that, when 2 (or more) names refer to the same person, the corresponding CER models will be very similar. The fetched snippets of text constitute the training set for training our CER, to recognize the name based on the surrounding text. For the purposes of the training, we only keep longish snippets (> 50 chars).

5.5.6 Training the Candidate Entity Recognizer

A Named Entity Recognizer usually works on broad entity categories, such as people, organizations, locations, etc. Being based on machine learning techniques, however, nothing keeps us from training a recognizer for a more specific task: the identification of not just any person, but, specifically, one of the k candidates. To do that, we censor the snippets retrieved as in Section 5.5.5 the same way we censored the original post, so that the CER is forced to build its model without relying on the specific occurrence of a name, as described by Algorithm 1. In fact, usually, a NER would leverage features of the name itself (e.g., recognizing the first given name). By removing the candidates' names, we force the model to rely on other features, i.e., the characteristics of the surrounding words. In order to be able to pinpoint a specific name, we annotate the censored names by one of the following k classes: ANON for the occurrences of the target name, and DUMBO1, ..., DUMBO$k-1$ for the other candidates. Table 5.4 shows the way a NER is usually trained and the specific training we decided to implement here.

5.5.7 Resolving the target name

The last step for finding out who the target name is applies the trained CER classifier to the censored snippets of Section 5.5.2. Given the way the CER was built, we can check whether, in correspondence to a censored snippet, the CER returns the class ANON. This is indeed the class label that we assigned to the occurrences of the actual name removed from the snippet, and therefore, the correct classifier answer. All DUMBOx's and, possibly, empty class assignments are to be considered wrong answers.

"Standard" NER annotation.	As anyone with access to the internet can attest to, there hasn't exactly been a lack of news to pick from: President <PERSON>Donald Trump</PERSON>'s possible collusion with the <ORGANIZATION>Russian Kremlin</ORGANIZATION> during the 2016 election and Secretary of State <PERSON>Rex Tillerson</PERSON>'s decision to break 18 years of tradition by declining to host a <EVENT>Ramadan</EVENT> event at the state department are two prime examples.
The NER annotation we used: in the example, Donald Trump is our name to find, Rex Tillerson is one of the candidates. Both have been replaced by random letters. For our purposes, organizations and events are ignored.	As anyone with access to the internet can attest to, there hasn't exactly been a lack of news to pick from: President <ANON>Xhyclertd</ANON>'s possible collusion with the Russian Kremlin during the 2016 election and Secretary of State <DUMBO1>Vlargdiun</DUMBO1>'s decision to break 18 years of tradition by declining to host a Ramadan event at the state department are two prime examples.

Table 5.4: *How a NER is normally trained vs how we train the CER for our purposes.*

5.5.8 Measuring the performance

In addition to this, we can measure how hard the task is, by comparing our performances with the ones of two simple baselines: 1) the first baseline assigns to the censored snippet the most frequent candidate that appears in the related comments; 2) the second one assigns a choice at random among our k candidates. Intuition might suggest that the first baseline could perform well, whereas the second one effectively represents a performance lower bound. All the experiments were conducted with $k = 5$, 10, and 20.

The CER is tested over the names in Tables 5.8 and 5.9. In particular, over a number of 149 names, occurring at least 100 times in the whole Facebook dataset, we consider only those with at least 50 occurrences in the comments related to the post where the name was censored, resulting in 95 names. The 95 names are reported in Table 5.10 in the Appendix.

For the evaluation of the system performances, we consider the following metrics. Given that the censored name is always inserted among the k candidates (with k = 5,

10, 20):

- The *CER* accuracy is defined as the joint probability that 1) the system successfully recognises the censored name, and 2) the name is really one of the k most frequent names in the comments associated to the censored posts.
- The *Global* accuracy is the probability of the system to successfully recognise the censored name, regardless of the fact that the name is among the k most frequent candidates (we remind the reader that the name is however present in the comments associated to the censored posts, thus guaranteeing the fairness of the approach).
- The *Most frequent selection* accuracy is the probability that the most frequent candidate is the censored name (first baseline).
- The *Random among top* k accuracy is the probability to recognise the censored name by randomly choosing from the top k candidates (and being the name among such candidates, second baseline).

Remarkably, the CER accuracy gives the accuracy of the system when the analysis is performed choosing among the actual top k most frequent names, over the associated comments per censored post. Thus, the CER accuracy and the Global accuracy match when the name is actually among the k most frequent names in all the comments associated to the censored posts. Since the CER accuracy is computed as a joint probability, it holds that CER accuracy $\leq$ Global accuracy ≤ 1.

Target name	Post	CER accuracy	Global accuracy	Most freq. selection	Random among top 10
Mitt Romney	10	0.20	0.40	0.00	0.07
Rudy Giuliani	18	0.44	0.44	0.22	0.10
Bernie Sanders	20	0.45	0.45	0.05	0.09
Gary Johnson	20	0.45	0.45	0.35	0.10
Mike Pence	20	0.50	0.50	0.30	0.10
Rahm Emanuel	20	0.75	0.75	0.20	0.10
Ryan Lochte	14	0.79	0.79	0.36	0.10
Colin Kaepernick	20	0.75	0.85	0.45	0.09
Paul Ryan	14	1.00	1.00	0.00	0.10
Rick Scott	15	1.00	1.00	0.27	0.10
μavg (all 49 candidates)	10.94	0.43	0.55	0.14	0.07

Table 5.5: *System performances: The worst 5 and top 5 results, considering target names censored in at least 10 posts. Settings: k = 10, $nocc \geq 200$.*

Table 5.5 and Table 5.6 report the results under the following settings: k = 10 and $nocc \geq 200$ (where $nocc$ is the number of occurrences of the name in the whole collection). Considering only the names that occur at least 200 times in the whole data collection and, among them, the ones that appear at least 50 times in the comments related to the post where those names was censored, we get a total of 49 names. Obviously, the 49 names are included in the 95 mentioned at the beginning of this Subsection 5.5.8. The complete outcome over the 49 names is in the Appendix, where Table 5.11 shows

the average results, both in terms of the single names, considering the number of posts in which the name has been censored, and in terms of the μaverage along all the candidates, considering the scenario with k = 10. The μaverage is computed as the average of the single averages, per name.

Table 5.5 shows an excerpt of Table 5.11 in the Appendix. In particular, it reports the worst and best 5 results (in terms of Global accuracy) considering those target names censored in at least 10 posts. As an example, let the reader consider the case of "Colin Kaepernick". Over the 20 posts in which the name was censored, the classifier correctly recognised the term in 75% of the time, if the analysis is run considering the real 10 most frequent candidates per post. The accuracy rises to 85% if we force to 1 the probability that the target name is among the 10 most frequent candidates, for the whole set of comments associated to each censored post.

Table 5.6 gives the overall statistics of the system, still evaluated over 49 names and k=10. It does not consider the single identities and it reports the flat average accuracy. Over a total number of 525 analysed posts, 49 censored names, and 10 possible candidates to choose from, the Candidate Entity Recogniser was able to correctly recognise the name 54% of the time. When not considering the actual most frequent candidates per post, the average system accuracy rises to 0.62. Such results outperform the outcome of the two baselines. Notably, and probably not so intuitively, picking up the most frequent candidate mentioned in the comments as the censored name is successful only in 19% of the cases. Even worse, choosing randomly among the 10 most frequent candidates leads to a success rate of about 10%, as is to be expected. This is an indication of how hard the task is, and whereas 60% performance might seem low for a classifier, the complexity of the task, and the simplifying assumptions must be taken into account. Regarding possible steps to improve the actual performances, we argue that the most direct direction to look into is widening the text window taken into account: this was not done in this Chapter, mainly because it further raises the issue of determining whether the enlarged window is relevant with respect to the censored name. Naively, here we assumed as relevant a short window around the occurrence.

Metric	Value
Total posts	525
Total names	49
Posts/names	10.93
CER accuracy	0.54
Global accuracy	0.62
Most freq. selection accuracy	0.19
Random among top ten	0.09

Table 5.6: *Overall statistics (k = 10, $nocc \geq 200$).*

Closing the discussion with settings k = 10 and $nocc \geq 200$, the classifier succeeds, on average, largely more than 50% of the time, choosing among 10 different candidate names. We argue that the performance of our snippet classification approach is very promising. Indeed, it is worth noting how we heavily constrained our operational setting, by considering concise snippets (50-200 characters each), both for posts and for comments. Furthermore, not all the comments related to a post are strictly related to the content of that post. Remarkably, the Facebook pages in which the posts are pub-

k	*nocc*	Most freq. selection	Most freq. selection μ avg	Random among top k	Random among top k μ avg
5	> 100	0.17	0.15	0.13	0.11
10	> 100	0.19	0.14	0.08	0.07
20	> 100	0.15	0.13	0.04	0.04
5	> 200	0.16	0.13	0.13	0.11
10	> 200	0.19	0.14	0.09	0.07
20	> 200	0.14	0.11	0.04	0.04

k	*nocc*	Global acc.	Global acc. μ avg	CER acc.	CER acc. μ avg
5	> 100	0.61	0.49	0.39	0.26
10	> 100	0.57	0.47	0.48	0.35
20	> 100	0.48	0.39	0.39	0.29
5	> 200	0.65	0.57	0.42	0.31
10	> 200	0.62	0.56	0.54	0.44
20	> 200	0.53	0.49	0.43	0.34

Table 5.7: *System performances varying the number of candidates k and the number of occurrences of the names $nocc$.*

lished contain the link to the complete news: we expect that considering the complete news text will lead to a sensitive improvement of the results. Finally, we notice that the length of our snippets is comparable to that of tweets, leading to the feasibility of applying the proposed approach over social platforms other than Facebook (e.g., Twitter and Tumblr).

5.5.9 Performances of the classifier under different settings

We conclude the presentation of the results by considering all the different settings in which we ran the experiments. Table 5.7 shows the system performances varying the values of k and $nocc$. Focusing on the CER accuracy, the best average performances are achieved when considering the pair (k=10, $nocc \geq 200$): such results have been already presented and discussed in the previous section. Turning to the global accuracy, we achieve a slightly better result still considering only the names that appear at least 200 times in the whole collection, but having only 5 names as possible candidates (Global Accuracy = 0.65).

Considering the number of occurrences of the target name in the whole collection, from the analysis of the values in the table, we can see that the worst performances are achieved with $nocc \geq 100$ (see, e.g., the column that reports the CER accuracy values, with 0.39, 0.48, and 0.39 $\leq$ 0.42, 0.54, and 0.43, respectively).

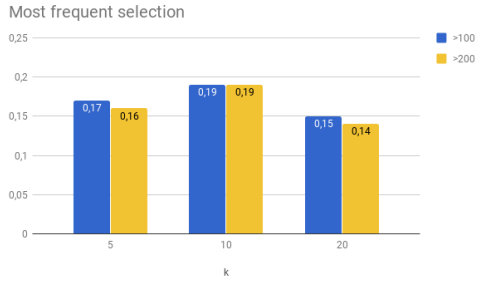
Instead, considering the number k of candidates to search within, taking into accounts all the 95 names ($nocc \geq 100$, first three lines in the table), the CER accuracy is noticeably higher when searching among 10 possible candidates (0.48) than that obtained with k = 5 and k = 20 (where the classifier achieves the same CER accuracy of 0.39). A different result is obtained for the Global Accuracy: the less the value of k, the better the accuracy.

The above considerations still hold considering the last three lines of the table (the ones with $nocc \geq 200$).

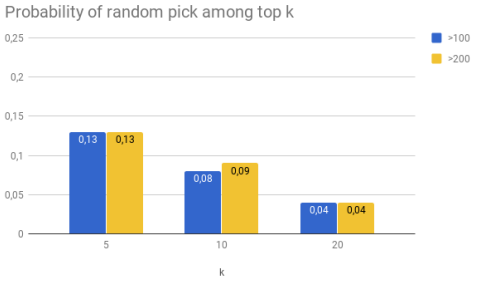
For all the possible pairs of k and $nocc$, there is an evident degradation of the performances, both when choosing the most frequent name as the censored name and when randomly guessing the name among the k candidates.

Finally, Table 5.12 shows the complete results over all the tested names, with k = 10.

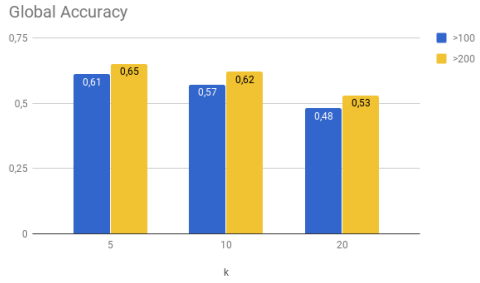
Figure 5.3 shows in a pictorial way the classifier performances and those of the



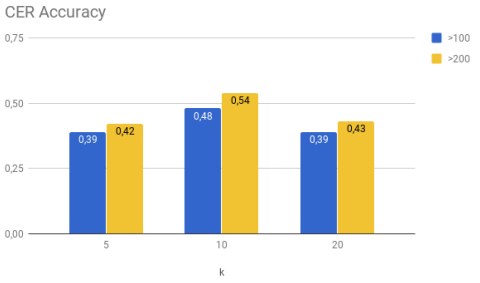
(a) *Probability that the censored name is the most frequent one*



(b) *Probability of randomly selecting the censored name among the top k candidates*



(c) *Global Accuracy*



(d) *CER Accuracy*

Figure 5.3: *System performances at a glance: bar charts.*

baseline techniques. Noticeably, the accuracy obtained when applying the baseline techniques (most frequent selection and random guessing among top k) are extremely poor, for all the tested combinations of k and $nocc$ (Figures 5.3a and 5.3b). When applying the proposed classifier, considering only names that occur in the whole collection at least 200 times leads to better results than considering the names with at least 100 occurrences. This assumption holds independently from the value of k for both Global and CER Accuracy (see the yellow bars in Figures 5.3c and 5.3d, with respect to the blue bars in the same figures). Finally, as already noticed, the best results, in terms of CER Accuracy, are achieved with the configuration k=10, $nocc \geq 200$ (Figure 5.3d). Overall, the Global Accuracy values are higher than the CER Accuracy values, since the former disregard the probability that the censored name is indeed in the top k most frequent candidates (see the comparison between the values in Figure 5.3c and Figure 5.3d).

5.6 Discussion

In this Chapter, we applied a text analysis technique based on snippet classification to unveil censored identities in texts. As a running scenario, we considered the Facebook pages of the major US newspapers, considering posts to news and their related comments. The approach outperforms baseline techniques such as choosing randomly from a set of possible candidates or picking up the most frequent name mentioned in the comments.

A limitation of the approach is given by the number of occurrences of the censored name that we need to find: in our experiments, we considered only names 1) cited more

than 200 times in the whole data collection, and 2) with at least 50 occurrencies in the associated comments to the posts where the names appear.

Considering the average outcome of our classifier, the accuracy of the system is largely above 50%, meaning that we were able to identify the correct name in more than half of the tested cases (choosing among 10 possible candidates). This is an encouraging result. Indeed, we considered very short snippets (similar to tweets): on the one hand, this demonstrates the capability to apply the same technique to microblogging platforms like Twitter; on the other hand, this leaves room for augmenting the performances, when considering longer texts (as an example for our scenario, the full version of the news, which is typically linked after the text of the post on the Facebook page). Finally, it is worth noting that, due to its coverage, we considered Facebook as a relevant case study. However, our methodology is general enough to be applied to various data sources, provided there is a sufficient number of training examples.

5.7 Appendix

5.7.1 Further details on candidate names

Politicians	Freq	Politicians	Freq
Hillary Clinton	23615	Rahm Emanuel	230
Donald Trump	17913	Debbie Wassermann	228
Bill Clinton	6668	Dick Cheney	222
Gary Johnson	3153	Chris Christie	222
Jill Stein	1210	Newt Gingrich	216
Michelle Obama	1079	Marco Rubio	215
John McCain	1079	Joe Biden	204
Bernie Sanders	890	Ken Starr	192
Paul Ryan	863	John Kerry	188
Mike Pence	859	Sheriff Joe	175
Barack Obama	782	Donna Brazile	172
Tim Kaine	707	Roger Stone	169
Vladimir Putin	686	Kellyanne Conway	169
Al Gore	529	Rand Paul	168
Harry Reid	509	Richard Nixon	166
George W. Bush	419	Jimmy Carter	162
Ronald Reagan	386	Kim Yong-un	162
David Duck	379	Mitch McConnell	158
Bill Weld	369	Pam Bondi	151
Vince Foster	361	Bashar al-Assad	149
Colin Powell	348	Jeb Bush	147
Loretta Lynch	335	Martin Luther King	135
Antony Weiner	333	John Brennan	134
Elizabeth Warren	333	Janet Reno	133
Mitt Romney	328	Jesse Jackson	124
John Podesta	323	Kelly Ayotte	119
Huma Abedin	320	George Duncan	118
Sarah Palin	303	Mark Kirk	116
Rudy Giuliani	281	Rob Portman	116
Robert Byrd	251	Thomas Jefferson	113
Paul Manafort	249	Bruce Rauner	102
Rick Scott	248	Pat Toomey	100

Table 5.8: *Politicians with more than 100 occurrences – whole collection.*

Celebrities	Freq
George Soros	1173
Margaret Sanger	600
James Comey	585
Paula Jones	566
Billy Bush	554
Monica Levinsky	438
Colin Kaepernick	401
Julian Assange	373
Melania Trump	362
Saul Alinsky	361
Ryan Lochte	320
Steve Bannon	321
Bill Cosby	314
Seymour Hersh	304
Ben Carson	283
John Kass	280
Tom Brady	278
Juanita Broaddick	272
Andrew McCarthy	260
Michael Phelps	259
Shaun King	257
Lester Holt	257
Isabel Kilian	253
Mark Cuban	235
Frank Gaffney	220
Tony Shaffer	218
Rosie o'Donnell	208
Sean Hannity	208
Clair Lopez	207
Alex Jones	205
Megyn Kelly	200
Amy Schumer	188
Roger Ailes	186
Rupert Murdoch	183
Mark Westcott	172
Beyonce	170
Bill Murray	168
Jeffrey Epstein	164
Al Sharpton	163
Lillie Rose Fox	162
Alec Baldwin	160
Jerry Jones	160

Celebrities	Freq
Sheldon Adelson	157
Madonna	152
Howard Stern	152
David Koresh	151
Ann Coulter	149
Anderson Cooper	149
Clint Eastwood	148
Matt Lauer	144
John Hinckley	143
Gennifer Flowers	143
Ilene Jacobs	142
Warren Buffett	139
Gloria Allred	137
Bob Dylan	135
Rachel Flowers	135
Brandon Marshall	133
Zoe Baird	132
Chris Wallace	131
David Hinckley	129
Mike Evans	128
Branch Davidian	125
Garrison Keillor	127
William Kennedy	126
John Oliver	126
Billy Dale	124
Tony Romo	123
Brock Turner	121
Alicia Machado	118
Rachel Flowers	135
Khizr Khan	115
Hope Solo	114
Michael Moore	112
Kim Kardashian	110
Michael Jackson	108
Rush Limbaugh	107
Brian Johnston	107
Scott Baio	106
Chelsea Clinton	105
Pope Francis	105
Dan Rather	104
Ted Nugent	103
Kevin Hart	101
Patrick Grant	100

Table 5.9: *Celebrities with more than 100 occurrences – whole collection.*

Politicians	Freq
Hillary Clinton	23615
Donald Trump	17913
Bill Clinton	6668
Gary Johnson	3153
Jill Stein	1210
Michelle Obama	1079
Bernie Sanders	890
Paul Ryan	863
Mike Pence	859
Barack Obama	782
Tim Kaine	707
Vladimir Putin	686
Al Gore	529
Harry Reid	509
George W. Bush	419
Ronald Reagan	386
Anthony Weiner	333
Elizabeth Warren	333
Mitt Romney	328
John Podesta	323
Huma Abedin	320
Rudy Giuliani	281
Paul Manafort	249
Rick Scott	248
Rahm Emanuel	230
Chris Christie	222
Newt Gingrich	216
Marco Rubio	215
Joe Biden	204
John Kerry	188
Sheriff Joe	175
Donna Brazile	172
Roger Stone	169
Kellyanne Conway	169
Richard Nixon	166
Jimmy Carter	162
Kim Jong-un	162
Pam Bondi	151
Jeb Bush	147
Martin Luther King	135
Janet Reno	133
Jesse Jackson	124
Kelly Ayotte	119
Mark Kirk	116
Bruce Rauner	102

Celebrities	Freq
George Soros	1173
James Comey	585
Paula Jones	566
Billy Bush	554
Monica Levinsky	438
Colin Kaepernick	401
Julian Assange	373
Melania Trump	362
Ryan Lochte	320
Steve Bannon	321
John Kass	280
Tom Brady	278
Juanita Broaddick	272
Michael Phelps	259
Shaun King	257
Lester Holt	257
Mark Cuban	235
Sean Hannity	208
Alex Jones	205
Megyn Kelly	200
Amy Schumer	188
Roger Ailes	186
Beyonce	170
Bill Murray	168
Alec Baldwin	160
Jerry Jones	160
Sheldon Adelson	157
Madonna	152
Howard Stern	152
Anderson Cooper	149
Clint Eastwood	148
Matt Lauer	144
John Hinckley	143
Gennifer Flowers	143
Warren Buffett	139
Brandon Marshall	133
Chris Wallace	131
Mike Evans	128
Garrison Keillor	127
Branch Davidian	125
John Oliver	126
Tony Romo	123
Brock Turner	121
Alicia Machado	118
Khizr Khan	115
Hope Solo	114
Kim Kardashian	110
Chelsea Clinton	105
Pope Francis	105
Kevin Hart	101

Table 5.10: *Candidates with at least 100 occurrences in the whole collection and at least 50 occurrences in the comments associated to the post where the name is censored*

5.7.2 Classifier Performances – k = 10, $nocc \geq$ 200

Target name	Post	CER accuracy	Global accuracy	Most freq. selection accuracy	Random among top 10 accuracy
Al Gore	3	0.00	0.00	0.00	0.07
Alex Jones	1	1.00	1.00	0.00	0.10
Anthony Weiner	18	0.50	0.56	0.11	0.09
Barack Obama	20	0.60	0.75	0.00	0.09
Bernie Sanders	20	0.45	0.45	0.05	0.09
Bill Clinton	20	0.55	0.55	0.00	0.10
Billy Bush	15	0.47	0.67	0.00	0.07
Chris Christie	7	0.14	0.43	0.00	0.01
Colin Kaepernick	20	0.75	0.85	0.45	0.09
Donald Trump	20	0.65	0.65	1.00	0.10
Elizabeth Warren	11	0.36	0.55	0.09	0.07
Gary Johnson	20	0.45	0.45	0.35	0.10
George Bush	15	0.40	0.60	0.13	0.07
George Soros	2	0.00	0.50	0.00	0.00
Harry Reid	12	0.58	0.75	0.17	0.08
Hillary Clinton	20	0.60	0.60	0.35	0.10
Huma Abedin	12	0.42	0.75	0.08	0.06
James Comey	20	0.60	0.75	0.00	0.08
Jill Stein	11	0.73	0.73	0.00	0.10
Joe Biden	7	0.57	1.00	0.14	0.06
John Kass	16	0.75	0.81	0.06	0.09
John Podesta	13	0.23	0.69	0.00	0.03
Juanita Broaddrick	1	0.00	0.00	0.00	0.00
Julian Assange	1	0.00	1.00	0.00	0.00
Lester Holt	2	0.00	0.00	0.00	0.05
Marco Rubio	9	0.22	0.33	0.44	0.07
Mark Cuban	6	0.00	0.17	0.00	0.05
Megyn Kelly	9	0.56	0.67	0.00	0.08
Melania Trump	20	0.60	0.60	0.80	0.10
Michael Phelps	1	0.00	1.00	0.00	0.00
Michelle Obama	20	0.50	0.50	0.05	0.10
Mike Pence	20	0.50	0.50	0.30	0.10
Mitt Romney	10	0.20	0.40	0.00	0.07
Monica Levinski	1	0.00	0.00	0.00	0.00
Newt Gingrich	8	0.25	0.38	0.25	0.06
Paul Manafort	1	0.00	0.00	0.00	0.10
Paul Ryan	14	1.00	1.00	0.00	0.10
Paula Jones	2	0.50	0.50	0.00	0.10
Rahm Emanuel	20	0.75	0.75	0.20	0.10
Rick Scott	15	1.00	1.00	0.27	0.10
Ronald Regan	4	0.50	0.50	0.25	0.10
Rudy Giuliani	18	0.44	0.44	0.22	0.10
Ryan Lochte	14	0.79	0.79	0.36	0.10
Sean Hannity	2	0.00	0.00	0.00	0.05
Shaun King	5	1.00	1.00	0.00	0.10
Steve Bannon	2	0.00	0.00	0.50	0.10
Tim Kaine	17	0.53	0.53	0.00	0.10
Tom Brady	7	0.71	0.71	0.14	0.10
Vladimir Putin	9	0.56	0.56	0.00	0.10
μ**avg**	11.04	0.54	0.56	0.14	0.07

Table 5.11: *Performances for names with at least 200 occurrences*

5.7.3 Classifier Performances – k = 10, $nocc \geq 100$

Target name	Post	CER accuracy	Global accuracy	Most freq. selection accuracy	Random among top 10 accuracy
Al Gore	3	0.00	0.00	0.00	0.07
Alec Baldwin	5	0.00	0.00	0.00	0.08
Alex Jones	1	1.00	1.00	0.00	0.10
Alicia Machado	10	0.20	0.30	0.00	0.03
Amy Schumer	14	0.29	0.29	0.07	0.10
Anderson Cooper	3	0.00	0.67	0.00	0.00
Anthony Weiner	18	0.50	0.56	0.11	0.09
Barack Obama	20	0.60	0.75	0.00	0.09
Bernie Sanders	20	0.45	0.45	0.05	0.09
Beyonce	1	0.00	0.00	0.00	0.10
Bill Clinton	20	0.55	0.55	0.00	0.10
Bill Murray	4	0.25	0.25	0.50	0.10
Billy Bush	15	0.47	0.67	0.00	0.07
Branch Davidian	1	0.00	0.00	0.00	0.00
Brandon Marshall	6	0.17	0.17	0.33	0.08
Brock Turner	4	0.00	0.50	0.00	0.05
Bruce Rauner	8	0.75	0.75	0.00	0.10
Chelsea Clinton	7	0.00	0.43	0.00	0.03
Chris Christie	7	0.14	0.43	0.00	0.01
Chris Wallace	1	1.00	1.00	0.00	0.10
Clint Eastwood	1	0.00	0.00	0.00	0.10
Colin Kaepernick	20	0.75	0.85	0.45	0.09
Donald Trump	20	0.65	0.65	1.00	0.10
Donna Brazile	2	0.50	0.50	0.00	0.05
Elizabeth Warren	11	0.36	0.55	0.09	0.07
Garrison Keillor	3	0.33	0.33	0.00	0.10
Gary Johnson	20	0.45	0.45	0.35	0.10
Gennifer Flowers	2	0.50	0.50	0.00	0.10
George Bush	15	0.40	0.60	0.13	0.07
George Soros	2	0.00	0.50	0.00	0.00
Harry Reid	12	0.58	0.75	0.17	0.08
Hillary Clinton	20	0.60	0.60	0.35	0.10
Hope Solo	1	1.00	1.00	1.00	0.10
Howard Stern	3	0.00	0.67	0.00	0.03
Huma Abedin	12	0.42	0.75	0.08	0.06
James Comey	20	0.60	0.75	0.00	0.08
Janet Reno	1	1.00	1.00	0.00	0.10
Jeb Bush	3	0.00	0.33	0.00	0.00
Jerry Jones	3	0.00	0.00	0.00	0.10
Jesse Jackson	3	0.67	0.67	0.00	0.10
Jill Stein	11	0.73	0.73	0.00	0.10
Jimmy Carter	1	0.00	0.00	0.00	0.00
Joe Biden	7	0.57	1.00	0.14	0.06
John Hinckley	1	0.00	0.00	0.00	0.10
John Kass	16	0.75	0.81	0.06	0.09
John Kerry	2	0.50	0.50	0.50	0.10
John Oliver	1	0.00	0.00	0.00	0.10
John Podesta	13	0.23	0.69	0.00	0.03
Juanita Broaddrick	1	0.00	0.00	0.00	0.00
Julian Assange	1	0.00	1.00	0.00	0.00
Kelly Ayotte	7	0.29	0.29	0.14	0.09
Kellyanne Conway	1	0.00	0.00	0.00	0.00
Kevin Hart	1	0.00	0.00	0.00	0.10
Khizr Khan	6	0.17	0.17	0.00	0.10
Kim Jong-Un	1	1.00	1.00	0.00	0.10
Kim Kardashian	9	0.56	0.56	0.33	0.10
Lester Holt	2	0.00	0.00	0.00	0.05
Madonna	1	0.00	0.00	0.00	0.00
Marco Rubio	9	0.22	0.33	0.44	0.07
Mark Cuban	6	0.00	0.17	0.00	0.05
Mark Kirk	8	0.75	0.88	0.63	0.09
Martin Luther King	2	0.00	1.00	0.00	0.00
Matt Lauer	2	0.00	0.00	0.00	0.10
Megyn Kelly	9	0.56	0.67	0.00	0.08
Melania Trump	20	0.60	0.60	0.80	0.10
Michael Phelps	1	0.00	1.00	0.00	0.00
Michelle Obama	20	0.50	0.50	0.05	0.10
Mike Evans	6	0.33	0.33	0.67	0.10
Mike Pence	20	0.50	0.50	0.30	0.10
Mitt Romney	10	0.20	0.40	0.00	0.07
Monica Levinski	1	0.00	0.00	0.00	0.00
Newt Gingrich	8	0.25	0.38	0.25	0.06
Pam Bondi	8	0.75	0.75	0.00	0.10
Paul Manafort	1	0.00	0.00	0.00	0.10
Paul Ryan	14	1.00	1.00	0.00	0.10
Paula Jones	2	0.50	0.50	0.00	0.10

Pope Francis	3	0.33	0.67	0.00	0.03
Rahm Emanuel	20	0.75	0.75	0.20	0.10
Richard Nixon	5	0.20	0.20	0.00	0.02
Rick Scott	15	1.00	1.00	0.27	0.10
Roger Ailes	2	0.00	0.00	0.00	0.05
Roger Stone	2	0.00	0.50	0.00	0.00
Ronald Regan	4	0.50	0.50	0.25	0.10
Rudy Giuliani	18	0.44	0.44	0.22	0.10
Ryan Lochte	14	0.79	0.79	0.36	0.10
Sean Hannity	2	0.00	0.00	0.00	0.05
Shaun King	5	1.00	1.00	0.00	0.10
Sheldon Adelson	1	0.00	0.00	1.00	0.10
Sheriff Joe	9	0.33	0.33	0.56	0.10
Steve Bannon	2	0.00	0.00	0.50	0.10
Tim Kaine	17	0.53	0.53	0.00	0.10
Tom Brady	7	0.71	0.71	0.14	0.10
Tony Romo	5	0.40	0.40	0.80	0.10
Vladimir Putin	9	0.56	0.56	0.00	0.10
Warren Buffett	3	0.00	0.00	0.00	0.03
μ**avg**	7.52	0.35	0.47	0.14	0.07

Table 5.12: *Performances over names with at least 100 occurrences*

CHAPTER 6

Final remarks

As highlighted in our discussion, which was focused on the exploitation of SM for improving citizens' life and health, the growing importance of such user-generated data poses serious questions about the role and the centrality of humans (and the data they produce) in modern intelligent systems. A first dichotomy arises by considering the context in which such data is produced. Indeed, access to private information for the common good could be desirable in some situations but it may also represent a threat to citizen privacy if used for mass surveillance. For instance, during natural disasters users might be willing to disclose more information about themselves and to loosen their privacy requirements since sensitive information is likely to be used for beneficial purposes (e.g., to track last known position of missing people). Anyway, this might not be the case during electoral campaigns when citizen personal information and opinions can be exploited to infer vote intentions or to assess the dissent towards a political party. Even more worrying is the possibility to mass surveil online user activities and private interactions (e.g., chat or phone conversations) in the name of purported intelligence purposes, such as crime or terrorism prevention. With regards to the analyzed content in this work, we could agree that mining drug forums to understand NPS diffusion is positive for health departments which can provide a better response to patients as well as law enforcers that can react promptly to the novel substances. On the other hand, such "invasive" monitoring could not be perceived positively by all drug consumers who could wish to remain anonymous. Moreover, as discussed in Chapter 5, SM identities can be spotted with little effort.

It is likely that we will observe an increasing number of applications focused on citizens mood monitoring and opinion mining, to tackle disappointment about products and brands in business intelligence (as it has already started) and prevent criminals from organizing using SM. Chapter 5 presented some of the aspects related to the fact

that privacy is not trivial to maintain and that censored identities in SM, even with nicknames, can be spotted provided there is a certain number of messages. Therefore, user perception of the common good and of the purpose for which their data is being collected and analyzed might be crucial to motivate SM adopters to disclose such data, thus enabling "social" intelligent systems to perform their tasks.

A possible solution for putting back citizens at the core of modern intelligent systems, instead of being relegated only to the borders of such systems as mere data sources, is to allow them to benefit from the results of those systems directly. Anyway, although appealing, the idea of feeding results back to the citizens has never been implemented in already deployed systems. To fill this gap, designers of the next-generation of "social" intelligent systems should be imaginative to this regard and could take inspiration from the research and applications already developed in the fields of e-democracy and e-government [43]. However, publicly opening the results of modern intelligent systems to the population at large also requires solving some challenges, such as ensuring the trustworthiness and credibility of both the collected and disclosed information, managing the decentralized coordination of information-empowered citizens, and more.

We showed how Web systems could be used to gather data from pervasive social sensors and we provided extensive experimental results derived from the employment of the proposed techniques, mostly in the field of earthquake emergency management and illicit drug trading. Overall results are promising and seem to encourage the adoption of such techniques.

All the techniques introduced to date rely on text analysis and on metadata that complements SM messages. We believe that analysis of such data is critical to get insights into unfolding emergencies or to monitor Web phenomenon or public events in real-time. However, extending such analyses to multimedia content (such as photos and videos) shared in the aftermath of disasters may further improve current SM-based emergency management systems. Indeed, the importance of images towards the assessment of the consequences of disasters has long been asserted, as they can improve situational awareness especially when such imagery data can be coupled with geographic and temporal information [53]. Commonly adopted procedures rely on very high resolution (VHR) satellite images. However, the multimedia content of SM data could be exploited as a complementary source of images in the aftermath of crisis and disasters. In fact, we observe a growing number of disaster-related messages carrying multimedia content, such as pictures of damaged buildings, wildfires, flooded regions, etc. Image classification and clustering techniques should be used for detecting and grouping images carrying sensitive information. Image classification techniques can help selecting only the most informative images, thus reducing the amount of data to be analyzed. Furthermore, being able to automatically group similar images, such as the ones showing the same damaged building, can significantly contribute to the understanding of the unfolding scenario.

Intelligent systems employed in the emergency management field have been originally designed to support decision makers and, usually, are not publicly accessible. Therefore, those social sensors that contributed data to the systems with their volunteered observations may feel the frustration of not directly benefiting from their efforts, nor seeing the results of the analysis performed by such systems. Indeed, social sen-

sors are not even part of the analysis process, as these intelligent systems are entirely data-driven and automated, and their output is the result of complex algorithms and data analysis techniques. The recent research proposed instead to increment the involvement of people, to try and merge human- and machine-computation, thus going in the direction of reducing the gap between volunteer citizens and intelligent systems and technologies [97, 178].

Unfortunately, there is still a rather big divide between volunteer citizens, analysis systems, and emergency stakeholders. A more in-depth and more effective synergy between these worlds might be the key to develop more resilient and human-centered systems [76, 103]. Conversely, following the current paradigm of seeing humans only as ubiquitous data sources, might lead to the opposite direction of mass surveilled societies [76]. This issue is particularly relevant within the context of terrorism and crime fight [15]. Indeed, the recent rise of terrorist attacks and global violence pushed an increasing number of people, from decision-makers to normal citizens, into considering mass surveillance as a possible solution to these problems[1].

The majority of crowdsourced emergency management systems, such as [6, 11, 64, 130, 154, 191], were designed to be of support for decision makers, keeping in mind that final users would have been members of a civil protection agency or experts in responding to emergencies. While developing and fine-tuning the `CrisMap` system, we debated whether a more transparent handling of the extracted information could also be of direct benefit for the population at large, well aware that several concerns may arise. Such concerns are mainly related to the possible misinterpretation of the information made available by the systems and to possible malicious behaviors [12].

Monitoring authorities. SM are becoming monitoring tools that encourage collective intelligence. In fact, an efficient bi-directional communication between population and institutions is fundamental to improve emergency response. This cooperation can be enhanced by exploiting SM to allow citizens to forward public requests for help and to notify dangers and critical situations. Publicly opening SM-based emergency management systems, on the one hand, would allow the population at large to understand better the kind and the volume of information made available to decision makers. On the other hand, this would stimulate civil protection agencies to exploit better information extracted from SM (e.g., ensuring that crowd generated reports were timely taken into account). As a consequence, institutions would be forced to improve the service quality, being aware that the effectiveness of their actions is verifiable, while crowds would be better motivated to share information that can be potentially directly available and verifiable by everyone.

Citizen empowerment. An open SM-based emergency management system makes digital volunteers aware that the information they share can directly help emergency responders and guide their decisions. This self-consciousness fosters the participation of the crowd in discussions and promotes proactivity in documenting the unfolding scenario posting tweets, photos, and videos. This form of citizen empowerment is likely to lead to a *virtuous circle* in which digital volunteers[2] are encouraged to share

[1] `https://www.theguardian.com/us-news/2015/apr/22/mass-surveillance-needed-isis-attack-mike-rogers`

[2] Examples of organizations that currently employ digital volunteers and that define their role in modern emergency management are the Humanitarian Response (`https://www.humanitarianresponse.info/en/applications/tools/category/digital-volunteers`) and the International Red Cross (`http://redcrosschat.org/disaster-digital-volunteer-training/#sthash.yVdEzS21.dpbs`)

more and better information for decision-makers to act. The active participation of people in emergency response is beneficial also in case of terrorist attacks, to help to identify attackers, or in massive events, to monitor the situation and maintain sufficient available services. Only recently, a few works envisioned this possibility and started experimenting with humans carrying out part of the analyses or being actively involved in the dissemination of the results [97, 178]. It is still too early to tell whether these approaches will ultimately succeed or fail, but the growing trend of increasing citizen involvement is undoubtedly fostering the development of increasingly human-centered societies.

False alarms. It is the case, however, that opening this kind of systems to the public at large also poses some serious questions about their trustworthiness and credibility [32]. SM systems are usually evaluated on their recall performance (i.e., the ability to recognize relevant events), while their precision (i.e., the ability to minimise false detections) is often overlooked [9]. Frequent false detections will inevitably lead people to ignore the system's outcome, making it unreliable and, at some level, useless. In some circumstances, false detection may even create alarms that induce panic and fear in populations. Errors are often caused by a poor quality of data source, as SM outputs are subject to people operating in extreme conditions: messages shared during emergencies are fragmented and lack a defined structure and means to assess information trustworthiness and credibility [104]. We argue that system quality may be improved by putting the crowd-in-the-loop, i.e., allowing people to apply corrections to the system autonomously. In fact, statistical results show that the aforementioned "wisdom of the crowds" ensures the absence of errors in data, or contributes to its reduction. Collaborative projects like Wikipedia[3], HarassMap[4], Humanitarian Tracker[5], and Ushahidi[6] are just but a few examples of "open" platforms that benefit from an active citizen participation. As such, those systems are particularly suitable to carry out verification and corroboration of user-contributed reports and might serve as compelling examples of future developments of SM-based emergency management systems.

Attack resilience. False detection might also be caused by malicious users willing to purposely disrupt the service. Robustness to attacks in critical systems, such as those operating in emergency management, is mandatory and should become a primary guideline in system design. Data filtering, a cleaning process carried out in many existing systems [64, 130, 154, 191], helps in validating single messages, but it fails to safeguard the system from bursts of bogus messages purposely shared. Although this issue is well-known to SM researchers, to date, the vast majority of SM-based emergency management systems does not employ security mechanisms [9]. In the `CrisMap` project, we experimented with fake accounts detection algorithms to reduce security concerns and mitigate the problem of bursts of fictitious (i.e., fake) messages [50]. Fake accounts detection and removal represents just one among the possible strategies for protecting SM-based systems from malicious attacks. Indeed, more research and experimentation is needed in this direction to increase the resilience and the reliability of these critical systems.

Centralized vs. distributed emergency management. One of the critical points

[3] `https://en.wikipedia.org`
[4] `http://harassmap.org/en/`
[5] `http://www.humanitariantracker.org/`
[6] `https://www.ushahidi.com/`

of discussion concerns who should manage and organized the emergency response and communications. A centralized approach has the drawback of a lack of fine-grained presence on struck areas. Indeed, it is unfeasible for civil protection agencies to continuously and accurately monitor the territory. Moreover, emergency responders are slow in the adoption of systems that differ from those traditionally used due to authoritative and responsibility issues. This position potentially hinders the amount of information available to decision-makers and makes them not completely aware of the situation when reacting to a crisis. While relying on the crowds can allow task parallelization, the lack of a central authority may conversely deteriorate decision quality. Furthermore, citizens often lack technical skills, and thus efforts might be inhibited by a lack of competences. Nonetheless, in recent emergencies, volunteer citizens converging to the disaster zone played a fundamental role in starting and maintaining grassroots initiatives, as happened in Genoa after a recent flash flood[7], where volunteers, called "mud angels", helped to remove mud from the streets, without any external coordination[8]. In this light, opening emergency management systems to these volunteers would give them tools to improve their coordination and efforts.

These concerns should be taken into account in the design of systems to support the population in critical situations as well as assisting research and law enforcers in maintaining and improving the life quality of society. Different cultural perspectives, especially those concerning privacy and freedom of expression deeply affect the diffusion and success of human-centric sensing and SM based tools. The evolution of this research is strongly tied to political issues, first of all, censorship, and to the will of people to continuously share opinions and information regarding their lives, experiences, and surroundings. Finally, an aspect hard to control is the SM platform policies, which regulate access to data by third parties. As long as SM platforms will make data accessible to everyone, these systems will continue to work (either with opportunistic or participatory approaches) and improve citizens' lives, but policy changes might in future abruptly decree the death of these techniques.

[7] https://it.wikipedia.org/wiki/Alluvione_di_Genova_del_9_e_10_ottobre_2014 (in Italian)
[8] https://en.wikipedia.org/wiki/1966_flood_of_the_Arno_River

Bibliography

[1] Mishari Almishari and Gene Tsudik. Exploring linkability of user reviews. In *ESORICS*, pages 307–324. Springer Berlin Heidelberg, 2012.

[2] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLOS ONE*, 9(1):1–11, 01 2014.

[3] Giuseppe Attardi, Antonio Gullì, and Fabrizio Sebastiani. Theseus: Categorization by context. *Universal Computer Science*, 1998.

[4] M. Avvenuti, S. Cresci, A. Marchetti, C. Meletti, and M. Tesconi. Predictability or early warning: Using social media in modern emergency response. *IEEE Internet Computing*, 20(6):4–6, Nov 2016.

[5] M. Avvenuti, S. Cresci, M. N. La Polla, A. Marchetti, and M. Tesconi. Earthquake emergency management by social sensing. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, pages 587–592, March 2014.

[6] M. Avvenuti, S. Cresci, F. Del Vigna, and M. Tesconi. Impromptu crisis mapping to prioritize emergency response. *Computer*, 49(5):28–37, May 2016.

[7] M. Avvenuti, F. Del Vigna, S. Cresci, A. Marchetti, and M. Tesconi. Pulling information from social media in the aftermath of unpredictable disasters. In *2015 2nd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 258–264, Nov 2015.

[8] Marco Avvenuti, Salvatore Bellomo, Stefano Cresci, Mariantonietta Noemi La Polla, and Maurizio Tesconi. Hybrid crowdsensing: A novel paradigm to combine the strengths of opportunistic and participatory crowdsensing. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1413–1421, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[9] Marco Avvenuti, Mario G. C. A. Cimino, Stefano Cresci, Andrea Marchetti, and Maurizio Tesconi. A framework for detecting unfolding emergencies using humans as sensors. *SpringerPlus*, 5(1):43, Jan 2016.

[10] Marco Avvenuti, Stefano Cresci, Fabio Del Vigna, Tiziano Fagni, and Maurizio Tesconi. Crismap: a big data crisis mapping system based on damage detection and geoparsing. *Information Systems Frontiers*, Mar 2018.

[11] Marco Avvenuti, Stefano Cresci, Andrea Marchetti, Carlo Meletti, and Maurizio Tesconi. Ears (earthquake alert and report system): A real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1749–1758, New York, NY, USA, 2014. ACM.

[12] Marco Avvenuti, Stefano Cresci, Fabio Del Vigna, and Maurizio Tesconi. On the need of opening up crowdsourced emergency management systems. *AI & SOCIETY*, 33(1):55–60, Feb 2018.

[13] James P. Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PLOS ONE*, 6(3):1–8, 03 2011.

[14] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 519–528, New York, NY, USA, 2012. ACM.

[15] Kirstie Ball and Frank Webster. *The intensification of surveillance: Crime, terrorism and warfare in the information age*. Pluto Press, 2003.

[16] Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa, Italy, 2014.

[17] Jennifer Bauduy. Mapping a crisis, one text message at a time. *Social Education*, 74(3):142–143, 2010.

[18] A. Bellandi, S. Nasoni, A. Tommasi, and C. Zavattari. Ontology-driven relation extraction by pattern discovery. In *2010 Second International Conference on Information, Process, and Knowledge Management*, pages 1–6, Feb 2010.

[19] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013.

[20] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.

[21] L. I. Besaleva and A. C. Weaver. Applications of social networks and crowdsourcing for disaster management improvement. In *2013 International Conference on Social Computing*, pages 213–219, Sept 2013.

[22] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science *vs* conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2):e0118093, 2015.

[23] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA, 1998. ACM.

[24] Cody Buntain and Jennifer Golbeck. This is your twitter on drugs: Any questions? In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 777–782, New York, NY, USA, 2015. ACM.

[25] Andrea Burattin, Giuseppe Cascavilla, and Mauro Conti. *SocialSpy: Browsing (Supposedly) Hidden Information in Online Social Networks*, pages 83–99. Springer International Publishing, Cham, 2015.

[26] L Burks, M Miller, and R Zadeh. Rapid estimate of ground shaking intensity by combining simple earthquake characteristics with tweets. In *10th US National Conference on Earthquake Engineering*, 2014.

[27] Peter Burnap and Matthew Leighton Williams. Hate speech, machine classification and statistical modelling of information flows on Twitter. In *Internet, Policy and Politics*, 2014.

[28] Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. Do not feel the trolls. *ISWC, Shanghai*, 2010.

[29] Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka, Jr., and Tom M. Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 101–110, New York, NY, USA, 2010. ACM.

[30] Giuseppe Cascavilla, Filipe Beato, Andrea Burattin, Mauro Conti, and Luigi V. Mancini. OSSINT-open source social network intelligence: An efficient and effective way to uncover private information in OSN profiles, 2016. arXiv preprint arXiv:1611.06737.

[31] Giuseppe Cascavilla, Mauro Conti, David G. Schwartz, and Inbal Yahav. Revealing censored information through comments and commenters in online social networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 675–680, New York, NY, USA, 2015. ACM.

[32] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM.

[33] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[34] Ming-Wei Chang, Lev Ratinov, and Dan Roth. Guiding semi-supervision with constraint-driven learning. In *Annual Meeting - Association for Computational Linguistics*, pages 280–287, 2007.

[35] S. Chattopadhyay, P. Ray, H. S. Chen, M. B. Lee, and H. C. Chiang. *Suicidal Risk Evaluation Using a Similarity-Based Classifier*, pages 51–61. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[36] T. M. Chen and V. Wang. Web filtering and censoring. *Computer*, 43(3):94–97, 2010.

[37] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. In *ICWSM*, pages 61–70, 2015.

[38] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, New York, NY, USA, 2010. ACM.

[39] Cindy Chiu, Chris Ip, and Ari Silverman. Understanding social media in china. *McKinsey Quarterly*, 2(2012):78–81, 2012.

[40] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[41] Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, and Maurizio Tesconi. Linguistically-motivated and lexicon features for sentiment analysis of italian tweets. *4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2014)*, pages 81–86, 2014.

[42] Jonathan H. Clark and Jose P. Gonzalez-brenes. Coreference resolution: Current trends and future directions. *Language and Statistics II Literature Review*, pages 1–14, 2008.

[43] Stephen Coleman and Jay G Blumler. *The Internet and democratic citizenship: Theory, practice and policy*, volume 1. Cambridge University Press, 2009.

[44] M. Conti, R. Poovendran, and M. Secchiero. Fakebook: Detecting fake profiles in on-line social networks. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1071–1078, Aug 2012.

[45] Mauro Conti, Fabio De Gaspari, and Luigi Vincenzo Mancini. *Anonymity in an Electronic Society: A Survey*, pages 283–312. Springer International Publishing, Cham, 2016.

[46] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995.

[47] S. Cresci, M. Avvenuti, M. La Polla, C. Meletti, and M. Tesconi. Nowcasting of earthquake consequences using big social data. *IEEE Internet Computing*, PP(99):1–1, 2016.

[48] S. Cresci, M. Petrocchi, A. Spognardi, M. Tesconi, and R. D. Pietro. A criticism to society (as seen by twitter analytics). In *2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 194–200, June 2014.

[49] Stefano Cresci, Andrea Cimino, Felice Dell'Orletta, and Maurizio Tesconi. *Crisis Mapping During Natural Disasters via Text Analysis of Social Media Messages*, pages 250–258. Springer International Publishing, Cham, 2015.

[50] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Fame for sale: Efficient detection of fake twitter followers. *Decision Support Systems*, 80:56 – 71, 2015.

[51] Stefano Cresci, Maurizio Tesconi, Andrea Cimino, and Felice Dell'Orletta. A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 1195–1200, New York, NY, USA, 2015. ACM.

[52] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *ECIR*, pages 693–696. Springer, 2013.

[53] Shideh Dashti, Leysia Palen, Mehdi P Heris, Kenneth M Anderson, Scott Anderson, and Scott Anderson. Supporting disaster reconnaissance with social media data: A design-oriented case study of the 2013 colorado floods. In *The 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 2014.

[54] Zoe Davey, Fabrizio Schifano, Ornella Corazza, and Paolo Deluca. e-Psychonauts: Conducting research in online drug forum communities. *Mental Health*, 21(4):386–394, 2012.

[55] Susannah Davies et al. Purchasing legal highs on the Internet-is there consistency in what you get? *QJM*, 103(7):489–493, 2010.

[56] Bernard de Bono, Pierre Grenon, Michiel Helvensteijn, Joost Kok, and Natallia Kokash. *ApiNATOMY: Towards Multiscale Views of Human Anatomy*, pages 72–83. Springer International Publishing, Cham, 2014.

[57] Maxwell Guimarães de Oliveira, Cláudio de Souza Baptista, Cláudio EC Campelo, José Amilton Moura Acioli Filho, and Ana Gabrielle Ramos Falcão. Producing Volunteered Geographic Information from social media for LBSN improvement. *Journal of Information and Data Management*, 6(1):81, 2015.

[58] Fabio Del Vigna, Marco Avvenuti, Clara Bacciu, Paolo Deluca, Marinella Petrocchi, Andrea Marchetti, and Maurizio Tesconi. *Spotting the Diffusion of New Psychoactive Substances over the Internet*, pages 86–97. Springer International Publishing, Cham, 2016.

[59] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy*, 2017.

[60] Fabio Del Vigna, Marinella Petrocchi, Alessandro Tommasi, Cesare Zavattari, and Maurizio Tesconi. *Semi-supervised Knowledge Extraction for Detection of Drugs and Their Effects*, pages 494–509. Springer International Publishing, Cham, 2016.

[61] Felice Dell'Orletta. Ensemble system for part-of-speech tagging. *Proceedings of EVALITA*, 9:1–8, 2009.

[62] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 29–30, New York, NY, USA, 2015. ACM.

[63] William H. Dutton, Nicole B. Ellison, and Danah M. Boyd. Sociality through social network sites, 2013.

[64] Paul S Earle, Daniel C Bowden, and Michelle Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), 2012.

[65] R. Ebina, K. Nakamura, and S. Oyanagi. A real-time burst detection method. In *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, pages 1040–1046, Nov 2011.

[66] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91 – 134, 2005.

[67] Nick Feamster, Magdalena Balazinska, Greg Harfst, Hari Balakrishnan, and David Karger. Infranet: Circumventing web censorship and surveillance. In *Proceedings of the 11th USENIX Security Symposium*, pages 247–262, Berkeley, CA, USA, 2002. USENIX Association.

[68] Nick Feamster, Magdalena Balazinska, Winston Wang, Hari Balakrishnan, and David Karger. *Thwarting Web Censorship with Untrusted Messenger Discovery*, pages 125–140. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[69] Paolo Ferragina and Ugo Scaiella. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1625–1628, New York, NY, USA, 2010. ACM.

[70] Clark C Freifeld, John S Brownstein, Christopher M Menone, Wenjie Bao, Ross Filice, Taha Kass-Hout, and Nabarun Dasgupta. Digital drug safety surveillance: monitoring pharmaceutical products in Twitter. *Drug Safety*, 37(5):343–350, 2014.

[71] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1019–1027. Curran Associates, Inc., 2016.

[72] Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53, 2016.

[73] H. Gao, G. Barbier, and R. Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14, May 2011.

[74] Judith Gelernter and Shilpa Balaji. An algorithm for local geoparsing of microtext. *GeoInformatica*, 17(4):635–667, Oct 2013.

[75] Judith Gelernter and Nikolai Mushegian. Geo-parsing messages from microtext. *Transactions in GIS*, 15(6):753–773, 2011.

[76] Karamjit S. Gill. Citizens and netizens: a contemplation on ubiquitous technology. *AI & SOCIETY*, 28(2):131–132, May 2013.

[77] Phillipa Gill, Masashi Crete-Nishihata, Jakub Dalek, Sharon Goldberg, Adam Senft, and Greg Wiseman. Characterizing web censorship worldwide: Another look at the opennet initiative data. *ACM Trans. Web*, 9(1):4:1–4:29, January 2015.

[78] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[79] Rebecca Goolsby. Social media as crisis platform: The future of community maps/crisis maps. *ACM Trans. Intell. Syst. Technol.*, 1(1):7:1–7:11, October 2010.

[80] Carol Grbich. *Qualitative data analysis: An introduction*. Sage, 2012.

[81] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2):17–28, 2013.

[82] Michelle Guy, Paul Earle, Scott Horvatch, Jessica Turner, Douglas Bausch, and Greg Smoczyk. Social media based earthquake detection and characterization. In *KDD-LESI 2014: Proceedings of the 1st KDD Workshop on Learning about Emergencies from Social Information at KDD'14*, pages 9–10, 2014.

[83] Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[84] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[85] Hui Han, Wei Xu, Hongyuan Zha, and C. Lee Giles. A hierarchical naive bayes mixture model for name disambiguation in author citations. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, SAC '05, pages 1065–1069, New York, NY, USA, 2005. ACM.

[86] Hui Han, Hongyuan Zha, and C Lee Giles. A model-based k-means algorithm for name disambiguation. In *Proceedings of the 2nd International Semantic Web Conference (ISWC-03) Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, Cham, 2003. Springer.

[87] Carl L Hanson, Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. Tweaking and tweeting: exploring Twitter for non medical use of a psycho-stimulant drug (Adderall) among college students. *Journal of Medical Internet Research*, 15(4):e62, 2013.

[88] Carl Lee Hanson et al. An exploration of social circles and prescription drug abuse through Twitter. *Medical Internet Research*, 15(9), 2013.

[89] Claire Hardaker. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Politeness Research*, 6(2):215–242, 2010.

[90] Robert Augustus Hardy and Julia R. Norgaard. Reputation in the Internet black market: an empirical and theoretical analysis of the Deep Web. *Journal of Institutional Economics*, FirstView Article:1–25, 2015.

[91] Tommy Hielscher, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. *Mining Longitudinal Epidemiological Data to Understand a Reversible Disorder*, pages 120–130. Springer International Publishing, Cham, 2014.

[92] Jennifer Hillebrand, Deborah Olszewski, and Roumen Sedefov. Legal highs on the Internet. *Substance Use & Misuse*, 45(3):330–340, 2010.

[93] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[94] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[95] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3-4):248–260, 2009.

[96] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.*, 47(4):67:1–67:38, June 2015.

[97] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 159–162, New York, NY, USA, 2014. ACM.

[98] Muhammad Imran, Shady Mamoon Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. In *Proceedings of the 10th International ISCRAM Conference*, pages 791–801, 2013.

[99] James A Inciardi, Hilary L Surratt, Theodore J Cicero, Andrew Rosenblum, Candice Ahwah, J Elise Bailey, Richard C Dart, and John J Burke. Prescription drugs purchased through the Internet: who are the end users? *Drug and Alcohol Dependence*, 110(1):21–29, 2010.

[100] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[101] Andreas M. Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59 – 68, 2010.

[102] Takeo Katsuki, Tim Ken Mackey, and Raphael Cuomo. Establishing a link between prescription drug abuse and illicit online pharmacies: Analysis of Twitter data. *Journal of Medical Internet Research*, 17(12), 2015.

[103] Andrea Kavanaugh, Ankit Ahuja, Manuel Pérez-Quiñones, John Tedesco, and Kumbirai Madondo. Encouraging civic participation through local news aggregation. In *Proceedings of the 14th Annual International Conference on Digital Government Research*, pages 172–179. ACM, 2013.

[104] Andrea L. Kavanaugh, Edward A. Fox, Steven D. Sheetz, Seungwon Yang, Lin Tzy Li, Donald J. Shoemaker, Apostol Natsev, and Lexing Xie. Social media use by government: From the routine to the critical. *Government Information Quarterly*, 29(4):480 – 491, 2012. Social Media in Government - Selections from the 12th Annual International Conference on Digital Government Research (dg.o2011).

[105] Jan H. Kietzmann, Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, 54(3):241 – 251, 2011. SPECIAL ISSUE: SOCIAL MEDIA.

[106] Gary King, Jennifer Pan, and Margaret E. Roberts. How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, 107(2):326–343, 2013.

[107] Ralf Klamma, Marc Spaniol, and Dimitar Denev. Paladin: A pattern based approach to knowledge discovery in digital social networks. In *I-KNOW*, volume 6, pages 6–8, 2006.

[108] April Kontostathis. Chatcoder: Toward the tracking and categorization of internet predators. In *Proc. Text Mining Workshop 2009 held in conjunction with the Ninth SIAM International Conference on Data Mining (SDM 2009). Sparks, NV. May 2009.*, 2009.

[109] Yelena Kropivnitskaya, Kristy F Tiampo, Jinhui Qin, and Michael A Bauer. The predictive relationship between earthquake intensity and tweets rate for real-time ground-motion estimation. *Seismological Research Letters*, 2017.

[110] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science advances*, 2(3):e1500779, 2016.

[111] Ieng-Fat Lam, Kuan-Ta Chen, and Ling-Jyh Chen. Involuntary information leakage in social network services. In *Proceedings of the 3rd International Workshop on Security: Advances in Information and Computer Security*, IWSEC '08, pages 167–183, Berlin, Heidelberg, 2008. Springer-Verlag.

[112] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6:70, 2001.

[113] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[114] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196, 2014.

[115] Christopher S. Leberknight, Mung Chiang, and Felix Ming Fai Wong. A taxonomy of censors and anti-censors part i: Impacts of internet censorship. *Int. J. E-Polit.*, 3(2):52–64, October 2012.

[116] Christopher S. Leberknight, Mung Chiang, and Felix Ming Fai Wong. A taxonomy of censors and anti-censors part ii: Anti-censorship technologies. *Int. J. E-Polit.*, 3(4):20–35, October 2012.

[117] Anders Ledberg. The interest in eight new psychoactive substances before and after scheduling. *Drug and Alcohol Dependence*, 152:73 – 78, 2015.

[118] Jan Marco Leimeister. Collective intelligence. *Business & Information Systems Engineering*, 2(4):245–248, Aug 2010.

[119] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

[120] Yuan Liang, James Caverlee, and John Mander. Text vs. images: On the viability of social media to assess earthquake damage. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 1003–1006, New York, NY, USA, 2013. ACM.

[121] Jack Lindamood, Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham. Inferring private information using social network data. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1145–1146, New York, NY, USA, 2009. ACM.

[122] Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. The paisa corpus of italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, 2014.

[123] Tim K Mackey, Bryan A Liang, and Steffanie A Strathdee. Digital social media, youth, and nonmedical use of prescription drugs: the need for reform. *Journal of Medical Internet Research*, 15(7):e143, 2013.

[124] Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan. Detecting flames and insults in text. In *Conference Papers (Centre for Research on Bangla Language Processing)*. BRAC University, 2008.

[125] Elaine Marsh and Dennis Perzanowski. Muc-7 evaluation of ie technology: Overview of results. In *Proceedings of the seventh message understanding conference (MUC-7)*, volume 20, 1998.

[126] C. M. Mascaro, A. Novak, and S. Goggins. Shepherding and censorship: Discourse management in the tea party patriots facebook group. In *2012 45th Hawaii International Conference on System Sciences*, pages 2563–2572, USA, Jan 2012. IEEE.

[127] Patrick Meier. New information technologies and their impact on the humanitarian sector. *International Review of the Red Cross*, 93(884):1239–1263, 2011.

[128] Patrick Meier. Crisis mapping in action: How open source software and global volunteer networks are changing the world, one map at a time. *Journal of Map & Geography Libraries*, 8(2):89–100, 2012.

[129] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA, 2011. ACM.

[130] S. E. Middleton, L. Middleton, and S. Modafferi. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2):9–17, Mar 2014.

[131] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[132] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[133] Jinyoung Min and Byoungsoo Kim. How are people enticed to disclose personal information despite privacy concerns in social network sites? the calculus between benefit and cost. *J. Assoc. Inf. Sci. Technol.*, 66(4):839–857, April 2015.

[134] Lev Muchnik, Sen Pei, Lucas C. Parra, Saulo D. S. Reis, José S. Andrade Jr, Shlomo Havlin, and Hernán A. Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, 3:15932, 05 2013.

[135] Zubair Nabi. The anatomy of web censorship in pakistan. *CoRR*, abs/1307.1144, 2013.

[136] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. In *SemEval@ NAACL-HLT*, pages 1–18, 2016.

[137] Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 2015.

[138] Amanda Nosko, Eileen Wood, and Seija Molema. All about me: Disclosure in online social networking profiles: The case of Facebook. *Computers in Human Behavior*, 26(3):406 – 418, 2010.

[139] Dawn C Nunziato. *Virtual freedom: net neutrality and free speech in the Internet age*. Stanford University Press, Redwood, CA, USA, 2009.

[140] Karen O'Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. Pharmacovigilance on Twitter? Mining tweets for adverse drug reactions. In *AMIA Annual Symposium*, page 924. American Medical Informatics Association, 2014.

[141] F Javier Ortega. Detection of dishonest behaviors in on-line networks using graph-based ranking techniques. *AI Communications*, 26(3):327–329, 2013.

[142] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing Twitter for public health. *ICWSM*, 20:265–272, 2011.

[143] Anselmo Peñas, Felisa Verdejo, Julio Gonzalo, et al. Corpus-based terminology extraction applied to information access. In *Proceedings of Corpus Linguistics*, volume 2001, pages 458–465, 2001.

[144] Angela Phillips. Sociability, speed and quality in the changing news environment. *Journalism Practice*, 6(5-6):669–679, 2012.

[145] Pranoti Pimpalkhute, Apurv Patki, Azadeh Nikfarjam, and Graciela Gonzalez. Phonetic spelling filter for keyword selection in drug mention mining from social media. *AMIA Summits on Translational Science*, page 90, 2014.

[146] P. Pirolli, J. Preece, and B. Shneiderman. Cyberinfrastructure for social action on national priorities. *Computer*, 43(11):20–21, Nov 2010.

[147] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[148] Kevin R. Scott, Lewis Nelson, Zachary Meisel, and Jeanmarie Perrone. Opportunities for exploring and reducing prescription drug abuse through social media. *Journal of Addictive Diseases*, 34(2-3):178–184, 2015.

[149] Ellen Riloff, Rosie Jones, et al. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pages 474–479, 1999.

[150] Stephane Roche, Eliane Propeck-Zimmermann, and Boris Mericskay. Geoweb and crisis management: issues and perspectives of volunteered geographic information. *GeoJournal*, 78(1):21–40, Feb 2013.

[151] Benjamin Rosenfeld and Ronen Feldman. Using corpus statistics on entities to improve semi-supervised relation extraction from the web. In *Annual Meeting - Association for Computational Linguistics*, pages 600–607, 2007.

[152] Matthew Rowe. Applying semantic social graphs to disambiguate identity references. In Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Simperl, editors, *The Semantic Web: Research and Applications*, pages 461–475, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[153] Matthew Rowe. The credibility of digital identity information on the social web: A user study. In *Proceedings of the 4th Workshop on Information Credibility*, WICOW '10, pages 35–42, New York, NY, USA, 2010. ACM.

[154] T. Sakaki, M. Okazaki, and Y. Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):919–931, April 2013.

[155] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[156] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523, 1988.

[157] Fabrizio Schifano, Ornella Corazza, Paolo Deluca, Zoe Davey, Lucia Di Furia, Magi' Farre', Liv Flesland, Miia Mannonen, Stefania Pagani, Teuvo Peltoniemi, Cinzia Pezzolesi, Norbert Scherbaum, Holger Siemann, Arvid Skutle, Marta Torrens, and Peer Van Der Kreeft. Psychoactive drug or mystical incense? Overview of the online available information on Spice products. *International Journal of Culture and Mental Health*, 2(2):137–144, 2009.

[158] Martin M Schmidt, Akhilesh Sharma, Fabrizio Schifano, and Charlotte Feinmann. Legal highs on the net-Evaluation of UK-based websites, products and product information. *Forensic Science International*, 206(1):92–97, 2011.

[159] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, Nov 1997.

[160] David G. Schwartz, Inbal Yahav, and Gahl Silverman. News censorship in online social networks: A study of circumvention in the commentsphere. *Journal of the Association for Information Science and Technology*, 68:569 – 582, 2017.

[161] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, March 2002.

[162] Andrei Serjantov. Anonymizing censorship resistant systems. In *International Workshop on Peer-to-Peer Systems*, pages 111–120. Springer, 2002.

[163] Andreas Sfakianakis, Elias Athanasopoulos, and Sotiris Ioannidis. Censmon: A web censorship monitor. In *USENIX Workshop on Free and Open Communication on the Internet (FOCI)*, 2011.

[164] Kumar Sharad and George Danezis. An automated social graph de-anonymization technique. In *13th Privacy in the Electronic Society*, pages 47–58. ACM, 2014.

[165] A. Sheth. Citizen sensing, social signals, and enriching human experience. *IEEE Internet Computing*, 13(4):87–92, July 2009.

[166] Noah A. Smith and Jason Eisner. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 354–362, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[167] Christophe Soussan and Anette Kjellgren. Harm reduction and knowledge exchange—a qualitative analysis of drug-related Internet discussion forums. *Harm Reduction Journal*, 11(1):1–9, 2014.

[168] Ellen Spertus. Smokey: Automatic recognition of hostile messages. In *AAAI/IAAI*, pages 1058–1065, 1997.

[169] Mani Srivastava, Tarek Abdelzaher, and Boleslaw Szymanski. Human-centric sensing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 370(1958):176–197, 2011.

[170] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, pages 1422–1432, 2015.

[171] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37(12):1936–1948, Dec 1992.

[172] B. Theilen-Willige, F. E. Bchari, H. A. Malek, M. Chaibi, A. Charif, C. Nakhcha, M. A. Ougougdal, M. Ridaoui, and E. Boumaggard. Remote sensing and gis contribution to the detection of areas susceptible to natural hazards in the safi area, w-morocco. In *2014 1st International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 1–5, March 2014.

[173] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[174] Salvatore Trani, Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Dexter 2.0: An open source tool for semantically enriching data. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, ISWC-PD'14, pages 417–420, Aachen, Germany, Germany, 2014. CEUR-WS.org.

[175] Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. Gerbil: General entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1133–1143, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[176] John-Paul Verkamp and Minaxi Gupta. Inferring mechanics of web censorship around the world. In *Workshop on Free and Open Communications on the Internet*, Bellevue, WA, 2012. USENIX.

[177] Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM)*. AAAI, 2011.

[178] S. Vieweg and A. Hodges. Rethinking context: Leveraging human and machine computation in disaster response. *Computer*, 47(4):22–27, Apr 2014.

[179] Sarah Vieweg, Leysia Palen, Sophia B Liu, Amanda L Hughes, and Jeannette Sutton. Collective intelligence in disaster: An examination of the phenomenon in the aftermath of the 2007 virginia tech shootings. In *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM)*, 2008.

[180] F. Del Vigna, M. Petrocchi, A. Tommasi, C. Zavattari, and M. Tesconi. Who framed roger reindeer? decensorship of facebook posts by snippet classification. Accepted in OSNEM, 2018.

[181] L. Wang and K. Kant. Special issue on computational sustainability. *IEEE Transactions on Emerging Topics in Computing*, 2(2):119–121, June 2014.

[182] Paul A. Watters and Nigel Phair. *Detecting Illicit Drugs on Social Media Using Automated Social Media Intelligence Analysis (ASMIA)*, pages 66–76. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[183] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 06 1998.

[184] Ingmar Weber and Venkata Rama Kiran Garimella. Visualizing user-defined, discriminative geo-temporal twitter activity. In *ICWSM*, 2014.

[185] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[186] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1980–1984, New York, NY, USA, 2012. ACM.

[187] J. Xie and T. Xiong. Stochastic semi-supervised learning on partially labeled imbalanced data. In Isabelle Guyon, Gavin Cawley, Gideon Dror, Vincent Lemaire, and Alexander Statnikov, editors, *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pages 85–98, Sardinia, Italy, 16 May 2011. PMLR.

[188] Steven Xu, HuiZhi Liang, and Timothy Baldwin. Unimelb at semeval-2016 tasks 4a and 4b: An ensemble of neural networks and a word2vec based model for sentiment classification. In *SemEval@ NAACL-HLT*, pages 183–189, 2016.

[189] Zhi Xu and Sencun Zhu. Filtering offensive language in online communities using grammatical relations. In *Collaboration, Electronic Messaging, Anti-Abuse and Spam*, pages 1–10, 2010.

[190] Christopher C. Yang, Haodong Yang, and Ling Jiang. Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media. *ACM Trans. Manage. Inf. Syst.*, 5(1):2:1–2:21, April 2014.

[191] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27(6):52–59, Nov 2012.

[192] Jolanta B. Zawilska and Dariusz Andrzejczak. Next generation of novel psychoactive substances on the horizon – a complex problem to face. *Drug and Alcohol Dependence*, 157:1 – 17, 2015.

[193] Shanyang Zhao et al. Identity construction on Facebook: Digital empowerment in anchored relationships. *Computers in Human Behavior*, 24(5):1816–1836, 2008.

[194] Aoying Zhou, Weining Qian, and Haixin Ma. Social media data analysis for revealing collective behaviors. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1402–1402, New York, NY, USA, 2012. ACM.