Information Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

CrowdPulse: A framework for real-time semantic analysis of social streams

Q1 Cataldo Musto*, Giovanni Semeraro, Pasquale Lops, Marco de Gemmis

Q3 *Universita degli Studi di Bari Aldo Moro, Department of Computer Science, Italy*

ARTICLE INFO

Article history:
Received 30 November 2014
Received in revised form
14 June 2015
Accepted 16 June 2015
Recommended by D. Shasha

Keywords:
Smart cities
Social networks
Text analytics
Sentiment analysis
Semantics

ABSTRACT

The recent huge availability of data coming from mobile phones, social networks and *urban sensors* leads research scientists to new opportunities and challenges. For example, mining micro-blogs content to unveil latent information about people sentiment and opinions is drawing more and more attention, since it can improve the understanding of complex phenomena and paves the way to the development of new innovative and intelligent services.

In this paper we present CrowdPulse, a domain-agnostic framework for text analytics of social streams. The framework extracts textual data from social networks and implements algorithms for semantic processing, sentiment analysis and classification of gathered data. The framework has been deployed in two real-world scenarios in order to identify the most at-risk areas of the Italian territory according to the content posted on social networks and to monitor the recovering state of the *social capital* of L'Aquila's city after the dreadful earthquake of April 2009[1], respectively.

In both scenarios, the framework showed its effectiveness and confirmed the insight that the combination of technologies specifically designed for Big Data processing with state-of-the-art methodologies for semantic analysis of textual content can provide very interesting findings and permits the analysis of such phenomena in a totally new way.

1. Introduction

According to a recent claim by IBM[2], 90% of data available today have been created in the last two years. This uncontrolled and exponential growth of the online information, which typically falls under the name of Big Data [37], led to two different trends: first, new technologies to store and process in an effective way these data are more and more required, as proved by the recent spread of Hadoop [58], BigTable [15] and MongoDB [16]. Second, a big variety of platforms and applications trying to extract some *value* from this *plethora* of information recently arise [25].

Indeed, it is a common viewpoint [27,17] that the combination of data coming from social media, smartphones and especially from *urban sensors* can actually enable the *smart cities model*, paving the way to the development of several innovative services and applications. By following this research line, the recent paradigm of *Social Sensing* [2,4] further emphasized this vision, since it proposed an integrated model in which users themselves are turned into *sensors*, entities that produce simple rough information which is processed and aggregated in order to generate some valuable *human-based* findings

* Corresponding author.
E-mail addresses: cataldo.musto@uniba.it (C. Musto), giovanni.semeraro@uniba.it (G. Semeraro), pasquale.lops@uniba.it (P. Lops), marco.degemmis@uniba.it (M.d. Gemmis).

[1] http://en.wikipedia.org/wiki/2009_L'Aquila_earthquake

[2] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

http://dx.doi.org/10.1016/j.is.2015.06.007
0306-4379/

obtained through the combination and merge of *individual-based* data.

Beyond the typical sensing applications, as those focusing on tracking vehicles [22] to avoid traffic congestions or healthcare tracking and predicting people's lifestyle [26], a big research effort has been made to analyze *text-based signals*, such as those coming from social networks like Twitter or Facebook. The reason is twofold: first, methodologies for Natural Language Processing (NLP) rely on very consolidated and effective algorithms, thus it is relatively simpler to process textual data rather than audio, video or especially environmental-based ones. Second, despite its size grows more slowly than video or audio data[3], textual content represents a very rich, interesting and valuable information source. As an example, 255 million Twitter active users broadcast everyday more than 500 million Tweets to their 208 followers (on average)[4], thus techniques for semantic analysis of textual content coming from social networks can provide very interesting findings and improve the understanding of psycho-social dynamics in a totally new way.

2. Motivations and scenarios

The spread of social networks radically changed and renewed many consolidated behavioral paradigms, since people today exploit these platforms for decision-making related tasks, to support causes, to provide their circles with recommendations or even to express opinions and discuss about the city or the place where they live[5]. Thanks to the heterogeneous nature of the discussions that take place on social networks, several interesting applications relying on the analysis of textual streams, ranging from online brand monitoring [67] and instant polls [45] to event and incident detection [55], recently emerged.

Generally speaking, the development of a framework for semantic analytics of textual content is a non-trivial task, since it requires the combination of several tools and techniques. Given that the required processing is strictly related to the nature of the data as well as to the analysis the user wants to perform, a NLP pipeline is often not enough to extract by itself valuable findings from data, thus it is necessary to couple it with advanced content processing methodologies such as opinion mining [48], content classification [56], semantic processing [62], and network analysis [64].

To this aim, in this paper we present a domain-agnostic framework for semantic analysis of social streams, the so-called CrowdPulse. The framework is able to extract, analyze and aggregate textual content produced by people on social platforms in order to provide users with some interesting findings and information which are hidden and latent in *human-generated* data. Our framework can perform massive extraction and mining of social streams and implements state-of-the-art algorithms for semantic processing and sentiment analysis of content. Moreover, it makes the output available through an interactive analytics console based on widespread and effective data visualization formalisms such as maps, charts and tag clouds.

One of the distinguishing aspects of this work lies in the originality of the scenarios in which the framework has already been deployed: *the Italian Hate Map* and *L'Aquila Social Urban Network*. In both cases, our platform has been exploited to develop novel intelligent services based on the analysis of social streams. In the first case, the aggregation and the semantic analysis of micro-blogs posts has been performed to build a map of the most at-risk areas in Italy, while in the latter semantic processing has been coupled with sentiment analysis and text classification to obtain a snapshot of people feelings and opinions about the state of the city of L'Aquila after the earthquake of 2009. In the following, an overview of both scenarios is provided.

2.1. The Italian Hate Map

This project aims to analyze the content produced on social networks in order to *measure* the level of intolerance of the Italian country. The analysis was performed by analyzing five different facets, called *intolerance dimensions*: homophobia, racism, violence against women, anti-semitism and disability.

The main goal of the project, inspired by the Hate Map built by Humboldt University[6], was to *localize* the areas where intolerant behaviors more frequently occur, in order to guide the definition of specific interventions (recovery and prevention, for example) on the territory. However, different from the American Hate Map, the project aimed at automatically labeling intolerant content and analyzing the Tweets in order to filter out ambiguous or polysemous terms.

In this scenario, CrowdPulse acted as real backbone since it was exploited to identify and extract the intolerant content from social networks, to filter out from the analysis ambiguous Tweets, to calculate the sentiment conveyed by the extracted content and to localize the Tweets in order to produce as final output an heat map[7] as that showed in Fig. 1. The insight behind the adoption of a heat map is to graphically emphasize the areas with a higher ratio of intolerant Tweets. In this specific example, northern Italy immediately emerges as the area where the users more frequently tweet intolerant content. It is worth to state that, by following the guidelines about the release of Open Data[8], all the content (filtered of all the information about both the author of the Tweet and the users who are mentioned) has been made available in CSV format as well, in order to help public administrations to analyze and study the enormous amount of content posted on social network and to proactively plan prevention and awareness activities in specific areas. More details about the

[3] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html

[4] https://about.twitter.com/company

[5] http://www.go-gulf.ae/blog/what-people-share-on-social-networks/

[6] http://users.humboldt.edu/mstephens/hate/hate_map.html

[7] http://http://en.wikipedia.org/wiki/Heat_map/

[8] Publishing Open Data: http://www.w3.org/TR/gov-data/

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

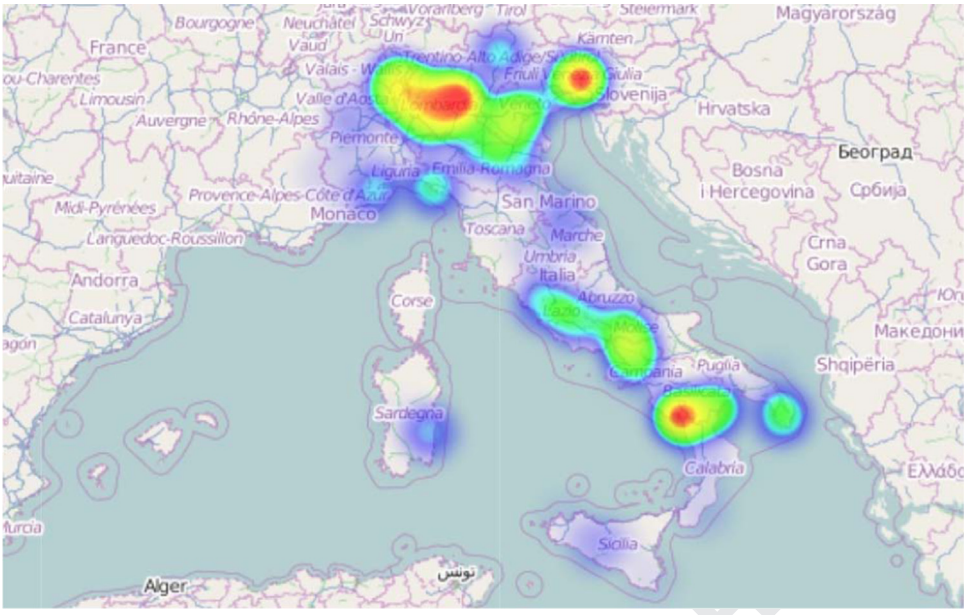

Fig. 1. Italian Hate Map – an example output.

methodology adopted to build the maps will be presented in next section.

2.2. L'Aquila Social Urban Network

L'Aquila hit the headlines in April 2009 because of a tremendous earthquake which killed 297 people. Nowadays, the severe trauma to physical and psycho-social structures is still in the phase of recovery. In this scenario, ENEA[9] (Italian National Agency for New Technologies, Energy and Sustainable Economic Development) proposed a smart cities-related project called *City 2.0*, aiming at empowering and revitalizing the urban heritage and the social capital of the city after the dreadful earthquake. At this end, an interdisciplinary team (researchers, architects and engineers) jointly worked with ENEA on the design of a Social Urban Network (SUN).

The SUN relies on the insight that the analysis of the content produced by the citizens on social networks can produce a reliable *snapshot* of the current state of the recovering process. The multidisciplinary facet of the project lies in the fact that typical Artificial Intelligence and NLP techniques have been coupled with psychological research.

Indeed, in the first part of the project a set of social indicators to be monitored (as *trust* or *sense of community*, see Fig. 2), defined by exploiting standard procedures of psycho-social research [46], has been set. Next, in CrowdPulse we implemented a methodology which automatically mapped all the content posted on social networks by L'Aquila citizens to those social indicators. Finally, by adopting sentiment analysis techniques, each social indicator was provided with a positive or negative synthetic aggregated *score*, defined on the ground of the sentiment conveyed by all the posts which refer to that social indicator. The underlying idea is to appoint some *community promoter* who can monitor in real-time through a visual dashboard the aggregated score of each social indicator, and can tackle the situation by identifying activities or specific interventions aimed at empowering some facets of the social capital when some negative trends emerge.

2.3. Objective and contributions

As previously introduced, the objective of this paper is to propose a domain-agnostic framework for semantic analysis of social streams. According to the nature of the above-described scenarios, it is possible to define a coarse-grained set of requirements which our framework has to implement:

1. Extract textual information from social networks.
2. Associate a richer semantics to each piece of content (e.g. the general topic a textual piece of information is about).
3. Associate an opinion (positive, negative, neutral) to each piece of content.
4. Aggregate and present the information stream in a way which is easy to be understandable for the users.

In the following sections, we will show how each of the features has been implemented, and how our framework has been designed to be easily adapted to any domain. The contributions of this paper can be summarized as follows:

- We propose a domain-agnostic modular framework for real-time processing of social streams of human-

[9] http://www.enea.it

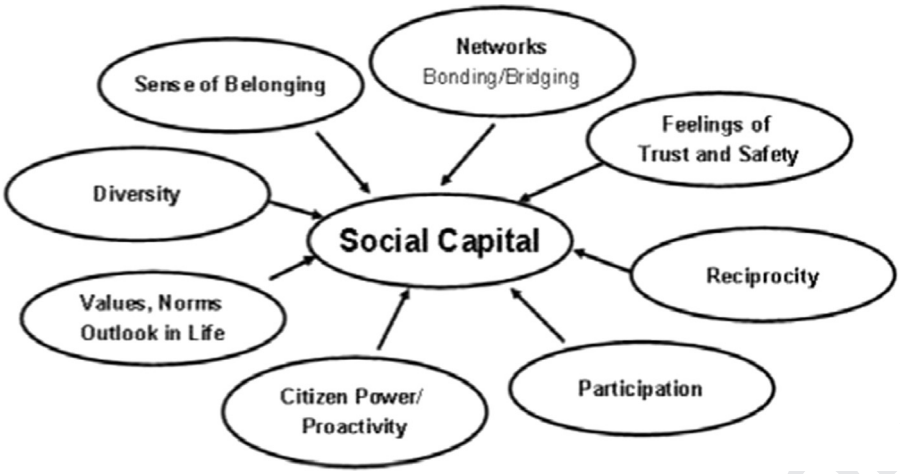

Fig. 2. Social capital indicators.

generated data.

- We introduce a pipeline based on state-of-the-art methodologies for semantic processing and sentiment analysis of textual content.
- We show the application of our framework in two real-world use case scenarios of our platform.
- We evaluate the effectiveness of the framework with two experiments based on real-world data.

The rest of the article is organized as follows: firstly, Section 3 presents Related Work in the area, while Section 4 describes the architecture of the framework: we will show the general workflow as well as the modules which compose it. For each module a detailed description of the functions along with the design choices will be provided. Next, in Section 5 we focus on the experiments performed on the above-mentioned scenarios, and finally Section 6 reports future research directions and concludes our work. In the Appendix we show some of the output produced by CrowdPulse for both scenarios.

3. Related work

The research line of social (or participatory) sensing [2,4,33] is based on the insight that the merge and the combination of crowd-based data can lead to the development of novel services and applications. The first work in the area date back to 2006, due to Campbell et al. [14], who introduced the concept of *people-based urban sensing*. This work paved the way to the definition of the first frameworks designed for collection of mobile-based data, as that proposed by Joki et al. [32]. The design of such platforms has been progressively refined and improved, in order to effectively deal with the recent large availability of data. As showed by Rachuri et al. [52], recent social sensing platforms take into account mobility pattern and energy consumption to gather as much data as possible from mobile sources.

Many recent work investigated the effectiveness of this paradigm in urban scenarios: Shin et al., for example, developed in [57] a model to identify the best transportation mode by exploiting smartphone data. Another innovative service is proposed by Albakour et al. [3], who proposed a novel framework for the retrieval of novel events based on the analysis of microblogging activities. The framework measures unusual microblogging activities in a certain area and uses that as an indication of the occurrence of an event, and exploits this information in a ranking function. Another typical application is represented by disaster management. Slavkovik et al. [59] show the state-of-the-art in the area of platforms for incident identification based on social sensing. The importance of crowdsourcing to improve the awareness of emergency and the usefulness of urban sensing platforms for such domains is also underlined by several research work [65,30]. As an example, Prasetyo et al. [51] propose a framework for the analysis of social data for preventing and monitoring urban disasters. They merge content-based information with emotions expressed, activity performed and network-based information. The research in the area has been fostered by the development of platforms for disaster detection based on text mining algorithms. A popular work in this research line is due to Abel et al. [1], who developed Twitcident, a platform for an incident or crises detection based on the combination of algorithms for real-time extraction of Twitter data streams with techniques for semantic analysis of content.

Regardless of the specific application domain, the research area concerning the application of text analytics algorithms to social media data (as micro-blog ones) falls under the name of Social Media Analytics [34]. According to Zeng et al. [66], social media analytics is supposed to provide tools to collect, monitor, analyze, and visualize social media data in an automated way. As already introduced, the typical application of such methodologies regards the marketing area [40,24]. In [67], Ziegler et al. show the application of social media analytics techniques for online brand monitoring. In [6], Armentano et al. analyze microblog posts to provide users with recommendations of people to follow.

More recent attempts focused on the application of such techniques in different domains: In [10], the authors

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

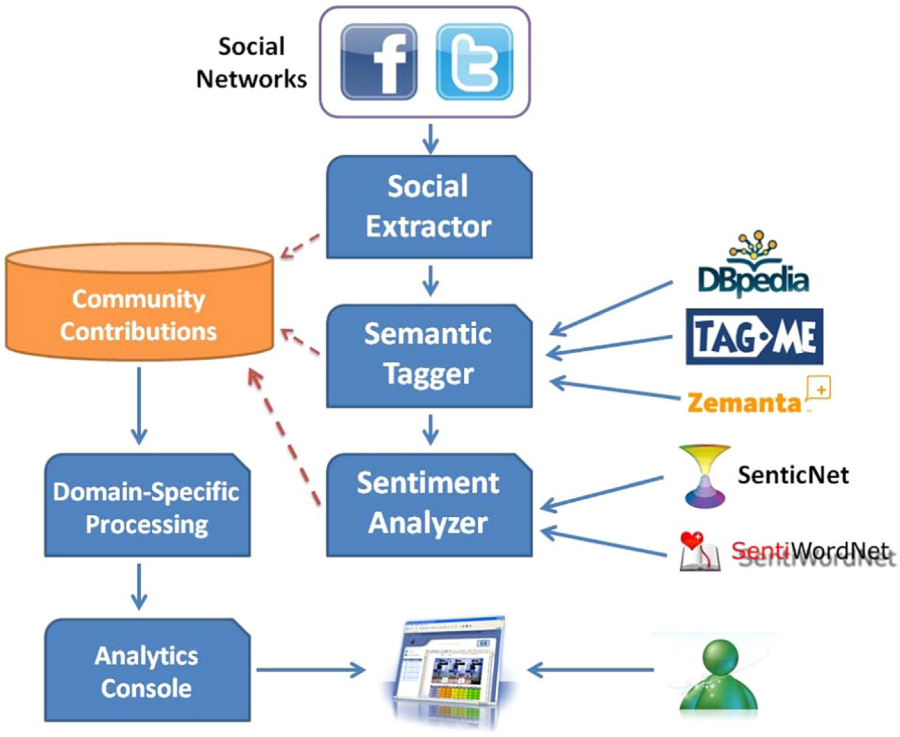

Fig. 3. The architecture of the semantic content analysis framework.

investigate the relationship between the mood expressed by Twitter users and the fluctuations of stock markets. To this aim, they combined algorithms for the extraction of textual content with external resource to obtain the sentiment conveyed by *social content*. The same strategy has been evaluated by the same authors also to analyze the dynamics of socio-economic phenomena [9]. The analysis of the sentiment expressed by people on social networks is also the focus of Felicittá, proposed by Allisio et al. [5]. In this paper, the authors propose a lexicon-based algorithm for sentiment analysis of geolocalized Tweets aiming at estimating the level of happiness of different areas in Italy. Differently, Schedl proposed in [49] a model to learn the relationship between Twitter posting behavior and music listening habits. On another work the authors applied text mining techniques to analyze microblogs discussion in the health domain: in [50], the authors apply statistical-based processing and natural language processing techniques to identify co-occurrences between concepts in health-related Tweets. Such analysis provided several interesting findings as localizing illnesses by geographic region, analyzing symptoms or medication usage and tracking illnesses over time.

Moreover, Paris et al. [49] show that through the analysis of social media political institutions and government one can easily track the opinion of the people about recent measures. The analysis of Twitter posts about politics is also investigated by Stieglitz et al. [60]. This framework represents the most similar attempt to develop a domain-agnostic framework for the extraction and the analysis of textual streams from social networks. Indeed, in this work the authors implement algorithms for the extraction of posts from Twitter, Facebook and weblogs and provide users with several heuristics and several visualization widgets.

But, different from CrowdPulse, this framework does not implement neither algorithms for semantic content representation nor techniques for sentiment analysis. This is a very important issue, since, as showed by our use cases, semantics and sentiment analysis play a key role for most of the potential scenarios. Furthermore, different from most of the applications currently available, the main distinguishing aspect of our framework by the originality of the scenarios in which it has already been deployed. Indeed, up to our knowledge, psycho-social analysis of social data is a never investigated research line.

4. CrowdPulse

CrowdPulse is a framework for real-time semantic analysis of social streams. The framework is based on the concept of *analysis*. Each analysis is run by defining a set of *extraction heuristics* and some *processing steps*. In a typical pipeline, a user interacts with the framework by defining the social networks she wants to analyze and the heuristics based on which will then extract content from those platforms. Next, the user defines the type of processes he wants to perform on the content previously extracted and the kind of data visualization he needs. The goal of the platform is to extract, analyze, aggregate and organize very large amount of rough data, in order to produce some analytics or some data visualization which is valuable for

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

final users. It is important to underline that the framework is totally domain-independent, thus it can aggregate and extract every kind of content the user wants to analyze.

The general architecture of the framework is provided in Fig. 3. Hereafter, a brief description of each component is provided.

- *Social extractor* feeds a database of contributions by exploiting the official APIs of the most popular social networks. This database is updated in real-time and it is fed according to specific heuristics (e.g. to extract all the Tweets containing a specific hashtag, all the posts or the Tweets coming from a specific location, and all the posts crawled from specific Facebook pages).
- *Semantic Tagger* associates to each piece of content the topic it is about. For this step we implemented a technique relying on a pipeline of *entity linking* algorithms, such as Tag.me [21] and DBpedia Spotlight [41].
- *Sentiment Analyzer* associates a polarity to each piece of content. In this case we implemented a Lexicon-based approach that exploits annotated vocabularies which associate a polarity (positive, negative or neutral) or a numerical sentiment score to all the terms of a language (e.g. SentiWordNet [19]).
- *Domain-specific Processing* further processes the output of the extraction and analysis pipeline, in order to produce the outcomes required by each specific scenario. To this aim, it integrates a broad range of Data Mining and Machine Learning techniques.
- *Analytics Console* shows the final output to the user, by providing him with different widgets and data visualization paradigms.

In a typical pipeline, a user interacts with the framework by defining her own extraction heuristics as well as the social networks she wants to analyze. Next, once the extraction processes have started, all the content is processed by the Semantic Tagger and the Sentiment Analyzer. This step is performed in background and the output is locally stored. Finally, the information is aggregated and is presented to the user through an interactive interface which is updated in real-time. The way the information is aggregated and the kind of widgets which are presented typically depend on the analysis and the outcomes the user wants to obtain: in some cases it can be useful to plot on a pie chart the sentiment of the population about a certain fact or brand, or to check the evolution of the sentiment over a certain period of time, while in other scenarios the user could ask to put all the geotagged content on a map in order to analyze the spread of a certain topic over different areas and so on. The analysis which could be performed through such framework can be potentially infinite.

In the next sections a thorough description of each of the above-described components, along with the design choices, will be provided.

4.1. Social Extractor

The *Social Extractor* is the essential component of the pipeline implemented in CrowdPulse. Given some *extraction heuristics*, the component connects to the most popular social network platforms in order to extract some content which matches the heuristics and to feed the database of contributions. This framework implements the bridges towards Facebook[10] and Twitter[11] by exploiting their official API. This design choice is due to the fact that, as showed by recent statistics[12], most of the online discussions arise on these social networks, thus they can provide good and reliable snapshots of feelings and opinions of the online population. As regards Twitter, the content is extracted by querying the official Streaming APIs, while for Facebook, due to privacy reasons, only the content coming from specific pages or specific groups has been extracted.

As extraction heuristics, six different alternatives are available to the users:

- *Content* extracts all the Tweets which contain a specific term.
- *User* extracts all the Tweets posted by a specific user, given its user name.
- *Geo* extracts all the available (geolocalized) Tweets, given latitude, longitude and radius.
- *Content+Geo* extract all the available geolocalized Tweets which match the terms indicated.
- *Page* extract all the posts coming from a specific page (the main post as well as the replies).
- *Group* extract all the posts coming from a specific group (the main posts as well as the replies).

Clearly, the first four heuristics regard Twitter while the last two are used to extract data from Facebook. Even if Facebook API permits the extraction of more content (e.g. all the *likes* of a specific user and all the discussions in a specific timeline), we only took into account the content labeled as *public*, since the goal of the platform is to perform a large-scale massive extraction and mining of content, without the need of an explicit authorization of the users.

It is worth to note that each of the above-mentioned extraction heuristics can be made more precise by associating a specific language to each of them, in order to better drive the extraction process on the ground of the requirements of a specific scenario (e.g. *all the content extracted around London and written in Chinese*). Language detection is performed by adopting state-of-the-art open source libraries[13].

Given these heuristics, the framework allows us to perform a broad range of analysis. As an example, it is possible to analyze the opinion of the people about different facts in different areas of the city, to analyze what are they Tweeting or posting about or even to see how a certain topic can spread over time or over the town. As regards our specific applications, in the *Italian Hate Map* scenario the Social Extractor was launched by defining a set of *sensible terms* for each of the above-mentioned intolerance dimensions. Due to the complexity of the task,

[10] http://developer.facebook.com
[11] http://dev.twitter.com
[12] http://techcrunch.com/2013/12/30/pew-social-networking/
[13] https://code.google.com/p/language-detection/

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

the definition of the lexicons associated to each dimension was performed by psychologists with specific experience in this domain. The final list contained 76 terms which were used to set the CONTENT heuristics during the extraction process. In this specific scenario, only Twitter was used as source to extract intolerant content. Indeed, due to Facebook policies, no groups or pages with a clear homophobic or racist intent are available on the platform. On the other side, by considering Twitter, the simple usage of the terms (with or without hashtags) clearly identifies the intent of the post.

In the case of *Social Urban Network*, the SOCIAL EXTRACTOR has been launched by using several heuristics. As regards Facebook, specific pages and groups managed by citizens of L'Aquila (especially those focusing on the discussions about the consequences of the earthquake) have been analyzed. For Twitter, both the GEO heuristic (set on the latitude and longitude of L'Aquila) and the USER one have been exploited. In the first case, all the Tweets localized in a range of 50 km from the city of L'Aquila have been extracted, while in the latter all the Tweets posted by the main local newspapers as well as the reTweets and the mentions to such articles by other users have been considered.

Regardless of the specific scenario, all the extracted information is then anonymized and locally stored in a database of *contributions*. MongoDB was chosen as storage solution[14], since its document-based storage model perfectly fits the requirements of our framework. Furthermore, as proved by recent statistics[15], it is gaining more and more attention and interest from both research and development communities.

Hereafter, we will refer to all the contributions as *social content*, regardless of the source they come from.

4.2. Semantic Tagger

All the social content gathered by the *Social Extractor* needs to be further processed before being aggregated, filtered and presented in the *Analytics Console* with which the user will interact. This is due to the fact that official API made available by social networks return the available content by adopting a simple keyword-based matching. As a consequence, due to the well-known problems of ambiguity of natural languages [36], a lot of noisy content is extracted, especially when *polysemous* terms are used in the extraction heuristics. The next example will clarify this aspect.

Let us suppose the term "L'Aquila" is used in the *Social Urban Network* project as extraction heuristics to get all the Tweets where people talk about the city. Unfortunately, as shown in Fig. 4, L'Aquila is a polysemous term, since in Italian it is the translation of the term *eagle*, as well. As a consequence, even if the first Tweet actually discusses about the problems of the city after the earthquake, the latter is about the risk of extinction of the eagle[16].

The typical solution to this issue is to couple the extraction algorithms with a Natural Language Processing (NLP) pipeline able to produce a less noisy representation of the content gathered by the *Social Extractor*. Moreover, given that each analysis performed by CrowdPulse is supposed to extract and process a lot of content (just think about how many Tweets about L'Aquila are posted every day in Italy), it makes sense to further improve the representation and the organization of the information by implementing *topic modeling* algorithms [63], such as the well-known Latent Dirichlet Allocation (LDA) [8], in order to provide the user with a more general and abstract overview of the topics the extracted content is about.

In order to match with these requirements, in our *Semantic Tagger* we implemented a pipeline of *entity linking* algorithms able to produce a transparent, richer and fine-grained semantic content representation relying on Wikipedia-based features (hereafter, we will refer to these features as *concepts*). In our approach, each *social content* has been processed through a pipeline of state-of-the-art entity linking algorithms. Specifically, we chose DBpedia Spotlight[17], Wikipedia Miner[18] and Tag.me[19].

Entity Linking (EL) [53] techniques share a common insight, since they all aim to map an input text (composed of n terms, $w_1 \ldots w_n$) to k entities ($e_1 \ldots e_k, k \leq n$ since it may happen that some terms do not map to any entity) that are mentioned in it. The EL process is typically carried out in two steps: definition of the *knowledge base* and definition of the *linking methodology*.

First, it is necessary to define a *knowledge base* which contains all the possible entities that can be linked. In our case, we chose Wikipedia thanks to its reliability and its broad coverage. As regards the *linking methodologies*, each approach has clearly its own peculiarities. However, all linking processes usually perform *mention detection*, ranking *candidate entities* and (eventually) *entity disambiguation*. These steps are carried out by exploiting a combination of statistical calculations and Machine Learning techniques, all relying on the information stored in Wikipedia, which is used as input corpus to build the model.

As an example, the *keyphraseness* [42] (the ratio between the number of times a term is used in Wikipedia to mention a particular entity by the number of times the term appears in the corpus) and the *commonness* [38] (the ratio between the number of times a term is used in Wikipedia to mention a particular entity by the number of times the term is used to mention other different entities) are typical statistical measures adopted to select and rank the candidate entities associated to a word form.

Given a set of candidate entities, the disambiguation step is performed by combining pure *lexical approaches* (as the calculation of the overlap between the input text and the description of the Wikipedia page of the candidate entity) with more sophisticated Machine Learning techniques (as classifiers [42] based on the analysis of surrounding words,

[14] https://www.mongodb.org/

[15] http://db-engines.com/en/ranking

[16] For the sake of simplicity, both Tweets are reported in Italian. The translation of the first one is 'Cialente sends out an SOS, L'Aquila is going to die' while the translation of the latter is 'Wolf, eagle, otter and black stork are rare and precious animals that live in Irpinia and are threatened'

[17] http://dbpedia-spotlight.github.io/demo/

[18] http://wikipedia-miner.cms.waikato.ac.nz/

[19] http://tagme.di.unipi.it/

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

Fig. 4. Example of ambiguous Tweets.

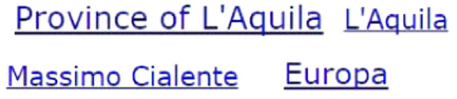

Fig. 5. Entity-based representation of a Tweet.

POS tagging information and the occurrence of entity-specific terms).

Typically, the output of the process is a set of entities each of which is provided with a confidence score. Thanks to this methodology it is possible to capture the mention of a name even if it is implicitly stated (this phenomenon is called name variations, as Barack Obama or Obama, that refer to the same entity). This is possible thanks to the fact that the algorithm is trained based on the input corpus in order to effectively capture the ways a particular entity is typically referred to.

Fig. 5 provides the output of the processing of the first Tweet. As shown in this figure, thanks to entity linking algorithm our framework is able to understand that a certain Tweet is about L'Aquila and Massimo Cialente (Mayor of the city). It is worth to note that this methodology is able to identify a mention to the Mayor even if its *first name* was not explicitly mentioned in the original Tweet.

It immediately emerges that such a representation, beyond being very transparent and more lightweight than classical topic modeling techniques such as LDA, automatically incorporates stopwords removal, bigrams recognition as well as entities identification and disambiguation. Furthermore, given that each entity is mapped to a Wikipedia page, we decided to browse the Wikipedia categories' tree to further enrich content representation by introducing the most relevant ancestor categories of that page. By considering the previous example, given the concept *Massimo Cialente*, the representation is enriched by adding as extra features concepts such as *Democrats Politics* and *L'Aquila Mayors*, thus extending the representation with other relevant features that may be of interest to understand what the content is about. This step is performed with no costs, by just browsing to the Wikipedia categories which are attached to the entity retrieved by the algorithm. It is important to further emphasize that entity linking algorithms, different from typical topic modeling techniques, can also enrich the representation by introducing features which not explicitly occur in the text, and this is tremendously important in order to obtain a more transparent and richer representation of social content.

To sum up, the goal of the Semantic Tagger is to process all the previously extracted content by exploiting entity linking algorithms, in order to obtain a richer and more fine-grained semantic representation which is very valuable and useful for final users, in virtue of its transparency and readability. It immediately emerges that this module plays a key role for both scenarios: as shown in the previous example, without a semantic processing step the SUN project would have taken into account a lot of noisy Tweets, thus providing a non-reliable snapshot of people discussions and opinions about the city. Similarly, the Italian Hate Map would have contained a lot of non-relevant Tweets since many seed terms used to extract intolerant content are polysemous (the Italian term *finocchio*, which is in the list of seed terms, can refer to both *queer* and *fennel*).

4.3. Sentiment Analyzer

Going back to the previous example regarding the city of L'Aquila, it is very important to understand the opinion of individuals about the topic they are posting about. To this aim, we implemented in CrowdPulse a *Sentiment Analyzer* module, whose goal is to associate a polarity to each social content.

State of the art approaches for sentiment analysis are typically classified in two categories: *supervised approaches* [29,47] learn a classification model on the ground of a set of labeled data, while *unsupervised* (or *lexicon-based*) ones [61,18] infer the sentiment conveyed by a piece of text by relying on (external) lexical resources which map each term to a categorical *(positive, negative, neutral)* or numerical sentiment score. As an example, terms such as *wonderful*, *beautiful* and *joy* have a positive sentiment score while terms such as *fear* and *sadness* have a negative one.

Even if recent work in the area showed that supervised approaches slightly overcome lexicon-based ones [44,54], in CrowdPulse we preferred the latter, since they have the advantage of being effective also in the absence of pre-labeled training data, and this aspect is very important for a framework which performs real-time computations such as CrowdPulse. Different from the Semantic Tagger, for the Sentiment Analyzer we did not rely on any external algorithm, but we defined one on our own.

Our lexicon-based algorithm is based on the insight that the polarity of a piece of text can be obtained on the ground of the polarity of the micro-phrases which compose it. Formally, the sentiment S conveyed by a social content C is defined as the sum of the polarity conveyed by the *micro-phrases* $m_1 \ldots m_n$. A new *micro-phrase* is built

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

whenever a *splitting cue* is found in the text. As *splitting cues* punctuations and conjunctions were used.

If we take into account the first Tweet reported in Fig. 4 *"Cialente sends out an SOS, L'Aquila is going to die"* it is split into two micro-phrases, delimited by comma.

Next, the polarity of each *micro-phrase* depends on the sentiment score of each term in the micro-phrase, labeled as *score*(t_j), which is obtained from an external lexical resource. As lexical resources we compared two state-of-the-art solutions:

- *SentiWordNet:* SentiWordNet [7] is a lexical resource devised to support Sentiment Analysis applications. It relies on WordNet [43], a lexical database for the English language. WordNet organizes all the English terms by grouping them in *synsets* (portmanteau of *synonym sets*). A *synset* is a group of word forms belonging to the same lexical category (*nouns*, *adjectives*, etc.) roughly sharing the same meaning. Moreover, WordNet automatically encodes binary relations between synsets such as hyponymy, meronymy, and hypernymy. WordNet's latest Online-version[20] contains 155,287 words organized in 117,659 synsets. SentiWordNet extends WordNet by associating to each synset three numerical sentiment scores (positivity, negativity, neutrality).
  Clearly, given that this lexical resource provides a synset-based sentiment representation, different senses of the same term may have different sentiment scores. As shown in Fig. 6, the term *terrible* is provided with two different sentiment associations.
- *SenticNet:* SenticNet [13] is a lexical resource for *concept-level* sentiment analysis. It relies on the Sentic Computing [12], a novel multi-disciplinary paradigm for Sentiment Analysis. Different from SentiWordNet, SenticNet is able to associate polarity and affective information also to concepts such as *accomplishing goal* and *celebrate special occasion*. At present, SenticNet provides sentiment scores (in a range between −1 and 1) for 14,000 common sense concepts.

In our approach, when *SentiWordNet* is used as lexicon, each term t_j is processed through a NLP pipeline to get its POS tag. Given that the structure of Tweets and Facebook posts is very different from typical textual content, we performed POS-tagging by exploiting a tool specifically designed for social content such as TweetNLP[21], described in [23].

Next, all the synsets mapped to that POS of the terms are extracted. Finally, *score*(t_j) is calculated as the weighted average of all the *sentiment scores* of the sysnets. A similar approach is performed for *SenticNet*, since the knowledge-base is queried and the polarity associated to that term is obtained. However, given that SenticNet also models common sense concepts, the algorithm tries to match more complex expressions (as *bigrams* and *trigrams*) before looking for simple unigrams. Moreover, we also evaluated a combined approach which associates to t_j the average of the score obtained from both SentiWordNet and SenticNet.

Next, given the score t_j gathered from a lexical resource, a list of modifiers (intensifier and downtoners) is exploited to update the sentiment score of the terms occurring in a fixed-size window near the modifier. Specifically, if a modifier is found, the score of the terms in the window is updated through the following formula:

$$score(t_j) = score(t_j) + (score(t_j) * w_{mod}) \quad (1)$$

where *score*(t_j) is the original score associated to the term t_j, while w_{mod} is the weight associated to the modifier. As an example, to the term *less* is associated a modifier score of −1.5, while *extremely* has a score of 0.35. The size of the window was set to 2 through a rough heuristics. As modifier list we adopted the one presented in [11], which consists of 175 concepts.

By referring to the previous example again, the Tweet in Fig. 4 contains only a sentimental term (*die*), which influences negatively the overall score (its SenticNet score[22] is −0.235).

Finally, we defined two different implementations of such approach: BASIC and EMPHASIZED. In the BASIC formulation, the sentiment of the social content *C* is obtained by first summing the polarity of each micro-phrase. Then, the score is normalized through the length of the whole text. In this case the micro-phrases are just exploited to invert the polarity when a negation is found in text:

$$S_{basic}(C) = \sum_{i=1}^{n} \frac{pol_{basic}(m_i)}{|T|} \quad (2)$$

$$pol_{basic}(m_i) = \sum_{j=1}^{k} score(t_j) \quad (3)$$

The EMPHASIZED version is an extension of the basic formulation which gives a bigger weight to the terms t_j belonging to specific part-of-speech (POS) categories:

$$S_{emph}(C) = \sum_{i=1}^{n} \frac{pol_{emph}(m_i)}{|T|} \quad (4)$$

$$pol_{emph}(m_i) = \sum_{j=1}^{k} score(t_j) * w_{pos(t_j)} \quad (5)$$

where $w_{pos(t_j)}$ is greater than 1 if $pos(t_j)$ = *adverbs, verbs, adjectives*, otherwise 1.

As regards the example about L'Aquila, in both cases the overall polarity of the Tweet is negative since it contains only a negative term. In the Experimental Evaluation session the effectiveness of the lexicons as well as of the variant of the approach has been evaluated against a subset of the Tweets gathered for the Italian Hate Map. Indeed, in such scenario it is tremendously important to put on the map only the Tweets expressing a negative opinion and containing an intolerant lexicon, thus we evaluated the effectiveness of our algorithm.

[20] http://wordnet.princeton.edu/wordnet/download/current-version/

[21] http://www.ark.cs.cmu.edu/TweetNLP/

[22] http://sentic.net/api/en/concept/die/

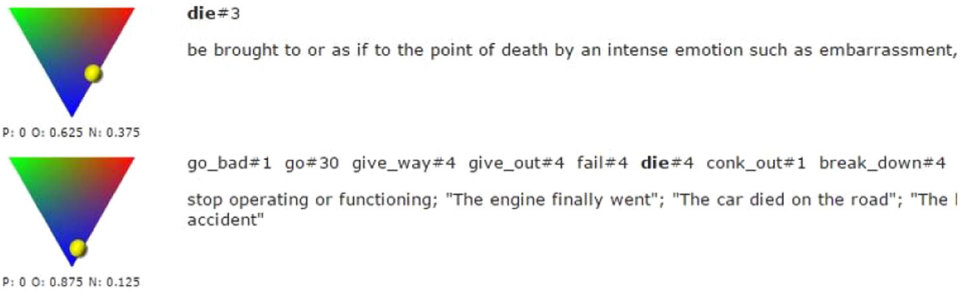

Fig. 6. An example of sentiment association in SentiWordNet.

4.4. Domain-specific processing

Each of the analysis which is carried out by CrowdPulse typically requires some further processing in order to produce the desired outcomes. This processing is typically domain and scenario-dependent, and it may range from the application of Machine Learning or Data Mining techniques (as text classification or social networks mining) to the definition of simple heuristics to filter out or to enrich the data previously extracted. To this aim, in our framework we exploited the Weka APIs[23], since it integrates a broad range of state-of-the-art techniques for Machine Learning and Data Mining. Hereafter, we will describe what kind of processing we implemented for the scenarios we already carried out.

In the *Italian Hate Map* project, as domain-specific processing we defined some heuristics to increase the number of geolocalized Tweets. As confirmed by recent studies [35], the amount of geolocalized Tweets is around 1% of the total number of Tweets everyday posted. Given that the goal of the project was to put on the maps as more Tweets as possible, we defined a methodology to increase the number of Tweets with geo-location information. To this aim, we exploited social network official APIs to extract the *location* attribute for all the users who posted intolerant content. When a specific location was indicated, all the content coming from that specific user inherited the information about the location. Similarly, we extracted all the content posted by each user in a 7-day windows. If other content (regardless it was intolerant or not) contained information about the location, the location itself was used to label all the intolerant Tweets from that user.

Furthermore, as regards the final output, some extra processing has been performed to provide a better snapshot of the most at-risk Italian areas. Indeed, the final maps do not show the rough number of Tweets localized in that specific area, since the value has been normalized by considering also a sample of (not intolerant) Tweets extracted from that area in the same period of time. In this way we showed on the maps the ratio of intolerant Tweets over the total, which is a more reliable value.

On the other side, as regards the *L'Aquila Social Urban Network*, we implemented a specific *Content Scoring and Classification* algorithm. As previously introduced, this processing is needed to map each content to one (or more) of the social indicators and to associate a score to each of them. This step has been carried out by comparing the effectiveness of two different approaches: in the first case, we exploited a set of labeled examples to learn a multi-class classification model to associate a specific Tweet or a specific post with the social indicator it refers to. As features, all the keywords as well as all the Wikipedia concepts (entities and categories) returned by the Semantic Tagger have been taken into account. In the latter, a list of sensible terms associated to each social indicator was defined by the team of psychologists who worked on the project, with the insight that the more the *sensible terms* appear in the content, the more the influence of that post on the social indicators.

Next, in order to provide each social indicator with a *score*, each content has been processed through the SENTIMENT ANALYSIS module as well. The overall score of the social indicator has been obtained by summing the sentiment score conveyed by all the citizens when they post something on the social networks about that indicator. Specifically, the merge of the sentiment of all the content coming from all the citizens about a specific indicator over a certain period of time provides the synthetic score which represents the snapshot of the feelings of L'Aquila's citizens.

The processing carried out by the domain-specific module developed for the SUN project is summarized in Fig. 7. Given a Tweet coming from a citizen of L'Aquila[24], the classification algorithms associate to that Tweet (it is about the idea of introducing new sustainable buildings in the town) two social indicators: *Sense of Belonging* and *Trust*. Next, the Sentiment Analysis algorithm associates to that content a positive sentiment score, which is inherited by both social indicators the content refers to. In this case, their score is slightly increased thanks to the sentiment conveyed by that user in her post. This process is performed in real-time, in order to continuously update the scores associated to each social indicator over time, as new content is published on the social networks.

In the next section the effectiveness of the domain-specific processing implemented for both scenarios has been evaluated.

[23] http://www.cs.waikato.ac.nz/ml/weka/index.html

[24] For the sake of simplicity, it is reported in its original version in Italian language

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

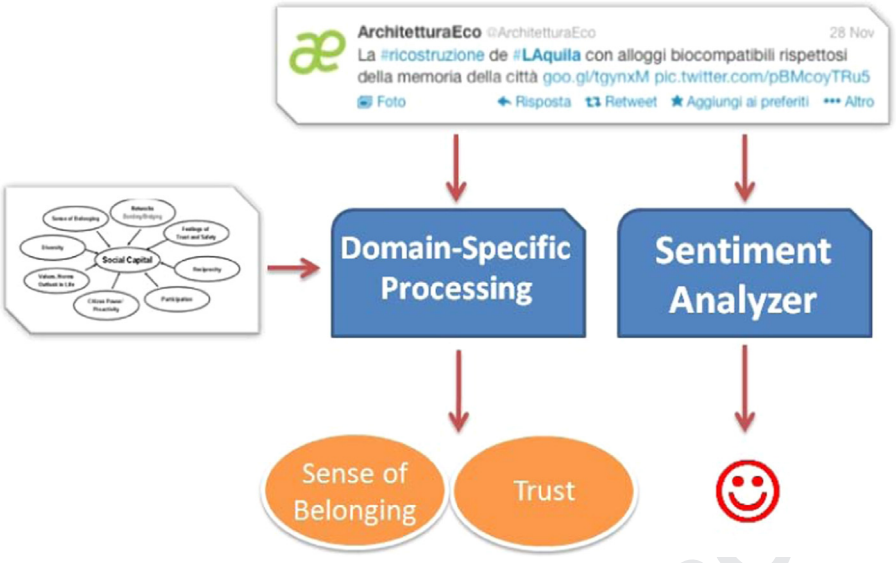

Fig. 7. Mapping Social Capital – Example.

4.5. Analytics Console

The goal of the Analytics Console is to provide the user with a set of tools which let him visualize and interact with the aggregated results of the analysis he launched. Specifically, we provided CrowdPulse with three different visualization widgets: MAPS, TAG CLOUDS and CHARTS, which are used to describe the output of the Extraction, Semantic Tagging and Sentiment Analysis modules, respectively.

MAPS (see Fig. 8) are used to immediately let the user visualize the geographical distribution of the data gathered by the extraction component. This can be very useful for several scenarios, as for example, to check the opinion of the citizens about recent administrative measures over different areas of the town, or to examine how popular is a topic in a particular area. Data visualization is performed by adopting the popular *heat map* formalism: the more intense the red, the more the content extracted from that particular location. Clearly, the maps produced as output by CrowdPulse are not static: the user can interact with them in the Analytics Console and can change the level of details by zooming in or zooming out, in order to obtain a broader or a more specific overview of data distribution.

TAG CLOUDS are used in CrowdPulse to aggregate and organize the output produced by the Semantic Tagger. As shown in Fig. 9, we designed three different kinds of tag clouds, one for each of the output produced by the Semantic Tagger. The *concepts* tag cloud reports the entities returned by the entity linking pipeline, while the *content* tag cloud shows the most popular terms and the hashtag used in the analysis requested from the user. Finally, the *categories* tag cloud is based on the Wikipedia categories attached to the entities identified in the text. Clearly, the size of each tag is related to the amount of content (Tweets or posts) in which the concept is used. All the elements in the tag cloud are not static, since the user can click on them. By clicking on a tag, the platform will update the widget by showing the most popular tags which are used in co-occurrence with the tag the user clicked on. This lets the user deepen the analysis by focusing on some specific subsets of the content (e.g. the posts about Massimo Cialente where people talk about *terremoto* – earthquake, in English). It is worth to note that, thanks to this visualization, the most relevant concepts as well as the most relevant topics which occur in the discussions immediately emerge and it is very simple to get for the user a quick overview of the distribution of the data gathered by the Extraction component.

Finally, CHARTS are used to report useful information about the trends of the data which have been extracted. As an example, in CrowdPulse the distribution of the sentiment over the posts stored for a certain analysis is plotted on a *pie chart*, while a *line chart* is used to show the amount of Tweets posted over time about a certain topic. As shown in Fig. 10, pie charts can summarize the amount of positive, negative and neutral Tweets, by providing a quick overview of the overall sentiment of a specific analysis. On the other side, through line charts also the temporal dimension can be taken into account as well, since the figure shows the trend in a specific time window (8–18 January) of positive (green line), negative (red line) and neutral Tweets (blue line).

The combination of maps, charts and tag cloud lets the user to analyze data and have some aggregate views of the latent information hidden in them, in order to obtain some valuable and reliable insights and analytics from the rough information gathered from social networks. In the Appendix, we will show some of the output produced by CrowdPulse for both scenarios.

5. Evaluation

In this section we report some details about the experiments performed for each of the use cases previously presented. As regards the Italian Hate Map project, we carried out an in vitro experiment aiming at measuring the effectiveness

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

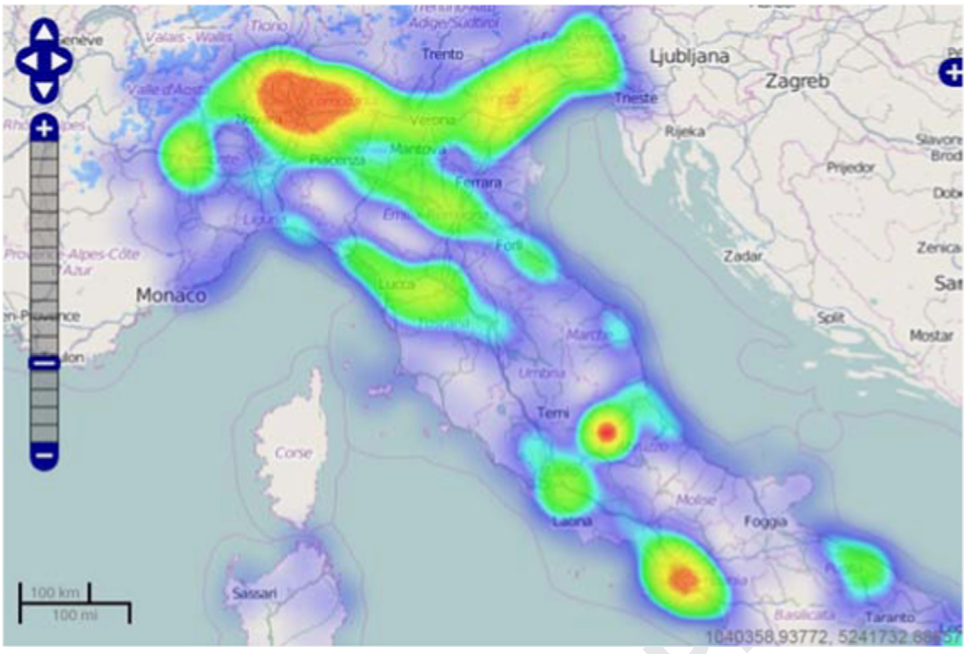

Fig. 8. An example of map-based visualization.

Fig. 9. An example of the semantic tag clouds implemented in CrowdPulse.

of the Sentiment Analysis algorithm in this specific scenario. Given that the goal of the project was to report on the map only the intolerant content, to have an algorithm able to filter out all the Tweets where the sensible terms are used without a specific intolerant intent is tremendously important. Similarly, in the scenario of SUN project, we compared the effectiveness of both strategies we proposed for *content classification*. Indeed, according to project's goals, it is very important to implement an algorithm able to classify the content with respect to the social indicator it actually refers to, in order to obtain a real and valuable snapshot of the trends of the social indicators over time. In the following, the experimental design as well as the results of the experiment is reported.

5.1. Sentiment analysis for the Italian Hate Map

The goal of this experiment was twofold:

1. To evaluate the effectiveness of our sentiment analysis technique in filtering out non-intolerant content from the map.
2. To identify the best lexicon as well as the best strategy to calculate sentiment scores.

As shown in Table 1, from January to October 2014 more than 1,800,000 Tweets were extracted (43,000 of them were geolocalized, around 2.3%) and were analyzed by the psychologists who worked on the project. Specifically, they analyzed the amount of the Tweets for each area and for each intolerance dimension, looking for some patterns in the distributions of the data (e.g. some areas with many intolerant Tweets for more than one dimension). Moreover, they studied terms usage in intolerant Tweets, by looking for interesting co-occurrences of terms or co-occurrences of concepts in a specific intolerance dimension. This analysis provided several interesting outcomes: as an example, it emerged that racist Tweets show a peak in conjunction with football matches and sports events, while Tweets against women are very frequent during TV-shows with showgirls. As previously explained, the Extraction component was fed with a set of 47 *sensible terms* defined by the team of psychologists over the five *intolerance dimensions*.

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

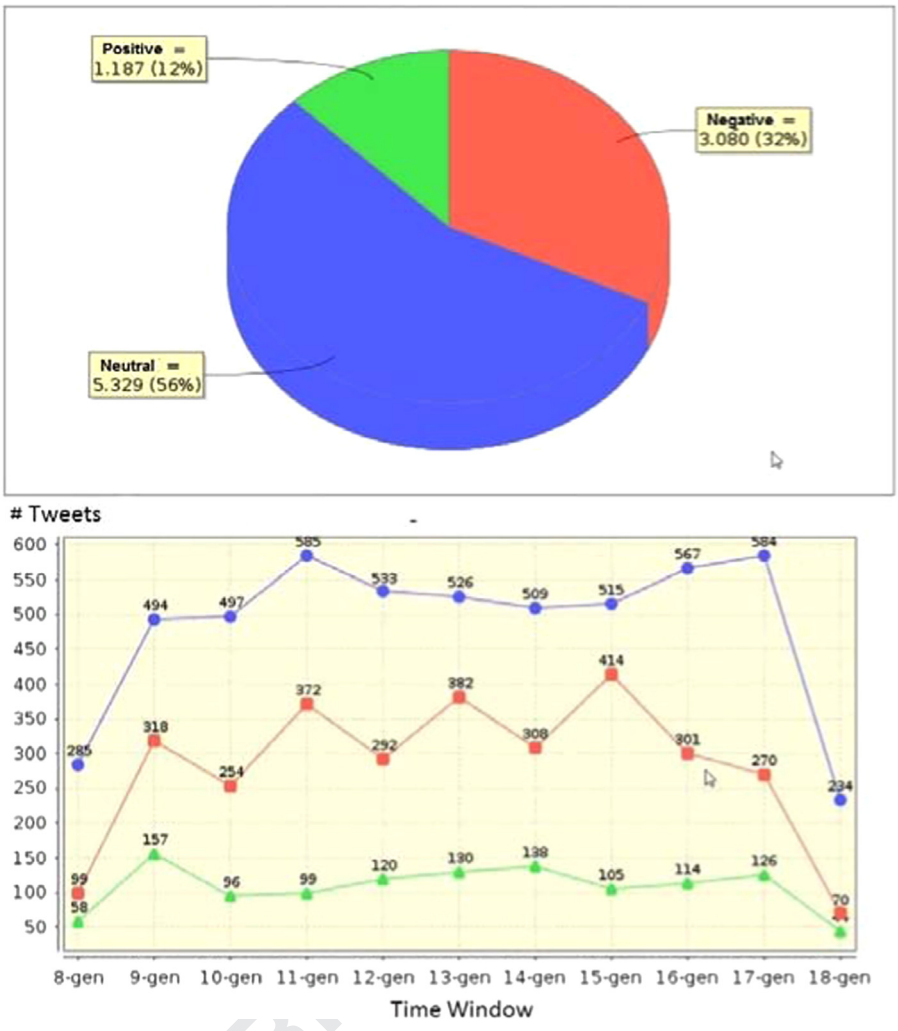

Fig. 10. An example of the charts plotted in CrowdPulse. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

Table 1
Italian Hate Map – statistics about extracted content.

Dimension	#Tweets	#GeoTweets	%GeoTweets	Timelapse
Homophobia	110,774	8501	7.66	January–October 2014
Racism	154,170	1940	1.24	January–October 2014
Violence	1,102,494	28,886	2.62	January–October 2014
Disability	479.654	3410	0.75	January–October 2014
Anti-Semitism	965	174	18.03	January–October 2014

To perform the evaluation, we extracted a little sample of 30,000 Tweets obtained by randomly sampling the whole set of Tweets. The Tweets were manually labeled by three different persons in order to obtain the *ground truth*. Next, 30% of the Tweets were used to learn the optimal classification threshold (the score above which a Tweet can be labeled as neutral or positive) by following a Greedy strategy. Next, the Sentiment Analysis algorithm was run on the remaining Tweets and the effectiveness of the techniques was evaluated by calculating the *F*1-measure [56] on the *Negative class*, since we were mainly interested in the precision of the algorithm on classifying negative Tweets. Specifically, nine different configurations of the algorithm have been compared, on varying of the lexicon adopted (SentiWordNet (SWN), Sentic.net and Mixed) and of the score calculation (Basic, Emphasized at 150% and Emphasized at 200%).

As shown in Table 2, the mixed strategy obtained the best results with very promising scores. Indeed, the best-performing configuration obtained an *F*1-measure of 76.3%. A quick analysis of the results showed that the *emphasis-based* algorithm did not improve the results. This

Table 2
Results of the experiment. The score reports the *F*1 measure obtained by the configuration.

	SWN (%)	**SenticNet** (%)	**Mixed** (%)
Basic	73.8	74.4	**76.3**
Emphasis 150	73.5	72.5	75.6
Emphasis 200	73.0	74.1	71.3

Table 3
Results of the experiment. The score reports the *F*1 measure obtained by the configuration.

	Classifier (%)	**Lexicon** (%)
Sense of Belonging	**46.52**	42.86
Diversity	33.33	**45.24**
Citizen Power	36.00	**50.00**
Participation	77.62	**80.00**
Trust	43.39	**80.95**
Overall	49.22	**69.00**

Table 4
Partial Lexicon for the Racism intolerance dimension.

Italian	**English**
Negro	Nigger
Rumeno di merda	Romanian Shit
Albanese di merda	Albanian Shit
Zingaro	Gypsy
Terrone	Southerner
Muso Giallo	Gook
Ebreo di Merda	Jew Shit
Crucco	Kraut
Kebabbaro	Kebabbaro
Rabbino	Rabbi
Giudeo	Jew

is probably due to the fact that intolerant behavior is typically expressed by using *nouns*, while emphasis-based configuration increases the score of different POS-categories (as verbs and adjectives) which do not have any influence in conveying intolerant content. As regards the effectiveness of single lexicons, the *F*1-measure obtained by SWN and SenticNet does not differ in a significant way. The improvement obtained by the Mixed configuration can be justified in virtue of the better coverage of intolerant words which comes from the merge of both lexicons. Indeed, the set of the terms which are modeled by SentiWordNet e SenticNet is not totally overlapping, thus a merge of the information coming from both information sources can lead to an improvement of the overall precision of the algorithm. To sum up, it is worth to note that these results are just preliminary, since they have been obtained without any tuning on the specific scenario. It is likely that a more in-depth analysis of the behavior of the algorithm could produce a further improvement of the overall results. However, even the algorithm as it is can produce reliable results which can be taken into account for the specific Italian Hate Map scenario.

5.2. Content classification in the SUN project

The goal of this experiment was to evaluate the effectiveness of the strategies which were implemented in the Content Classification module. As introduced in Section 4.4, we compared an approach based on a multi-class classification algorithm to a lexicon-based classification technique.

As regards the text classification algorithms, a randomly chosen sample of 5000 social content was manually annotated by three people. Next, the LibLinear library [20], an open source library for large-scale linear classification, was chosen to learn a classification model relying on labeled examples.

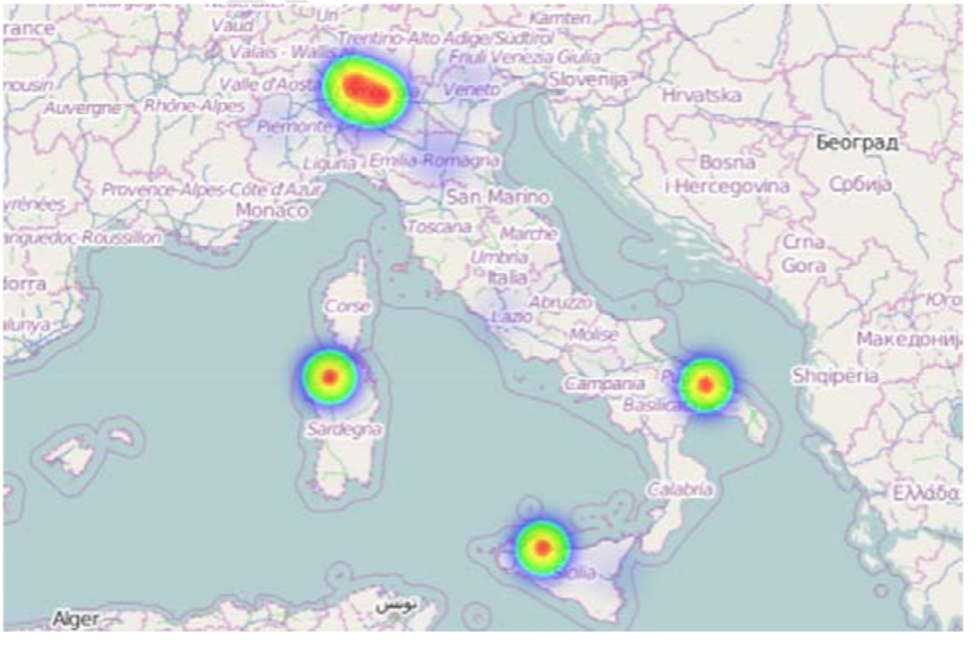

Fig. 11. Italian Hate Map – anti-semitism. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

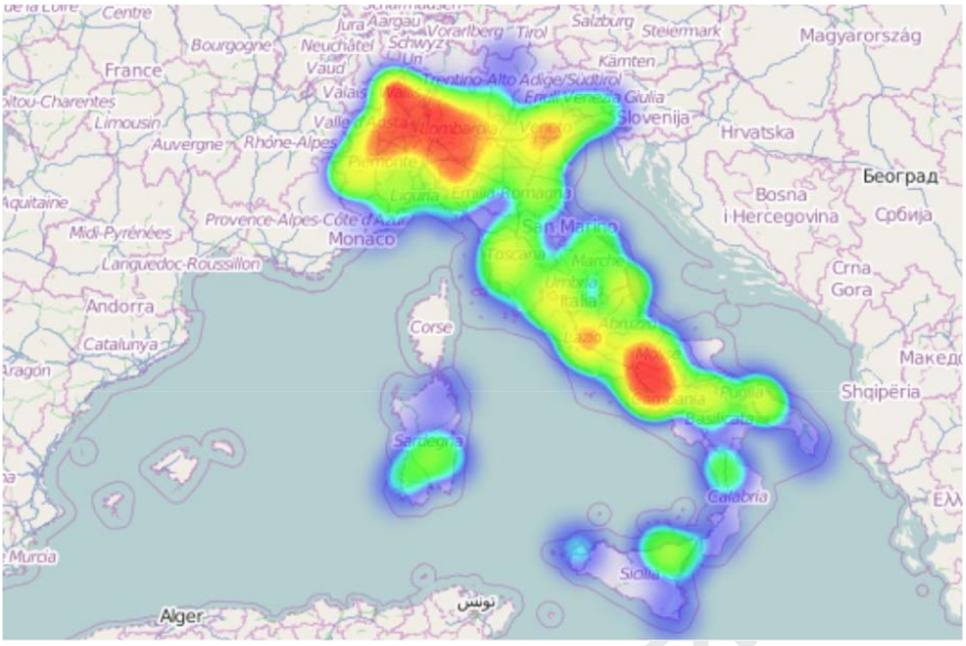

Fig. 12. Italian Hate Map – disability. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

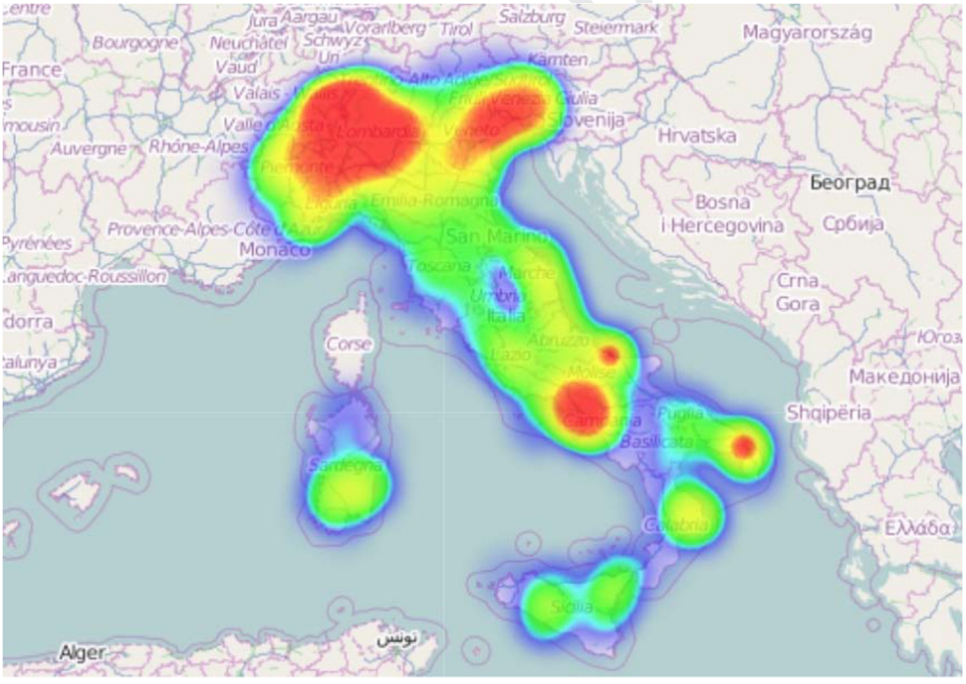

Fig. 13. Italian Hate Map – violence against women. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

LibLinear supports logistic regression and linear Support Vector Machines (SVM) [31]. In our specific scenario, SVM with linear kernel was used as learning technique, since several work suggested the adoption of that kernel for text classification problems [28].

Each social content was represented by merging the keywords, the Wikipedia concepts and the Wikipedia categories in which the entities are referred to. As Wikipedia concepts we exploited the entities returned by the entity linking pipeline, while as regards the keywords the content was processed through a NLP pipeline consisting of stopwords removal, stemming and POS-tagging processes. To better deal with the particular lexicon used in micro-blogs, the processing

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

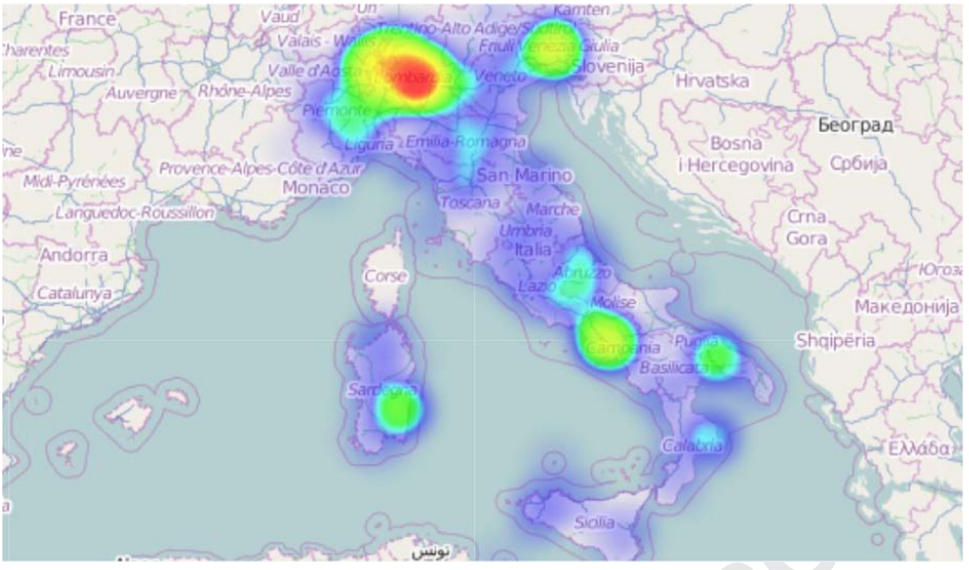

Fig. 14. Italian Hate Map – homophobia. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

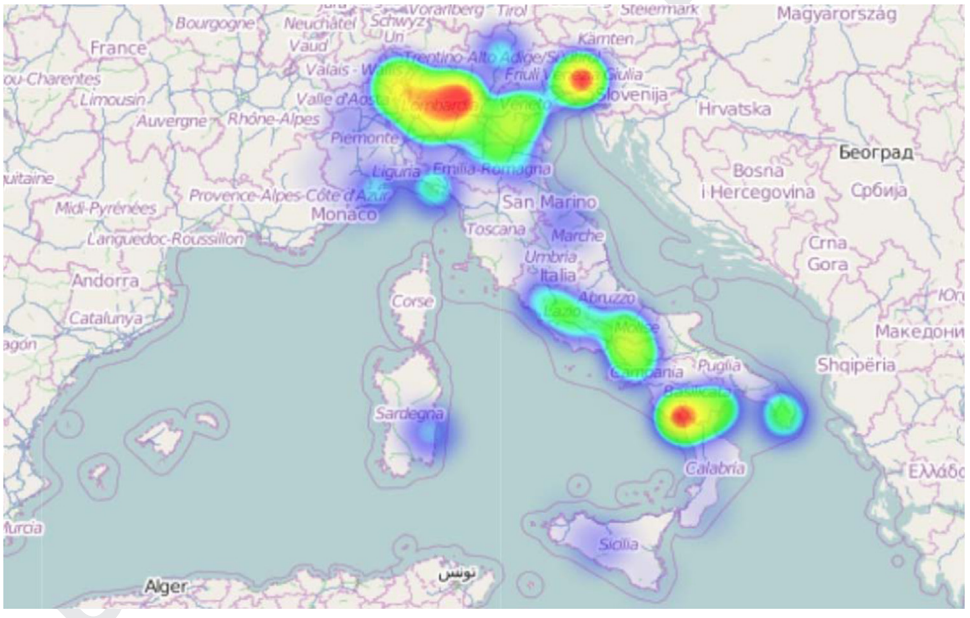

Fig. 15. Italian Hate Map – racism. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

pipeline was provided with a specific enriched list of stop-words (e.g. *btw* used as abbreviation of *by the way*). For the sake of simplicity, only 5 out of 8 *social indicators* were used in this experiment, due to small number of labeled examples obtained for the remaining categories. Specifically, the social indicators taken into account have been *Sense of Belonging*, *Participation*,*Trust*, *Diversity* and *Citizen Power*. Experiments were run by performing a split (70% training–30% test) of the sample, built by maintaining the ratio between the cardinality of each class, and the effectiveness was evaluated by calculating *F*1-measure on the test set.

On the other side, experiments for the lexicon-based approach were performed by defining a set of 40 sensible terms (8 for each social indicator, on average). The set has been defined by the team of the psychologists. Given that this approach did not need any training, experiments were performed on the same test set used for the text classifier. Each item in test set was classified by calculating the overlap between the features of each social content and the lexicon describing each social indicator. Clearly, each social content was classified in the social indicator which contained more overlapping terms. Results of the Experiments are reported in Table 3.

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007

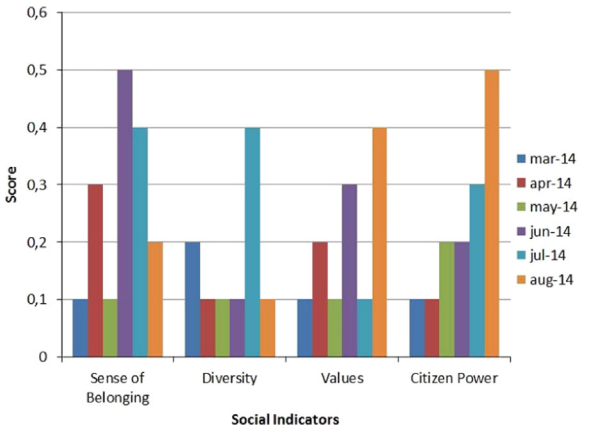

Fig. 16. L'Aquila SUN – social indicators (part 1).

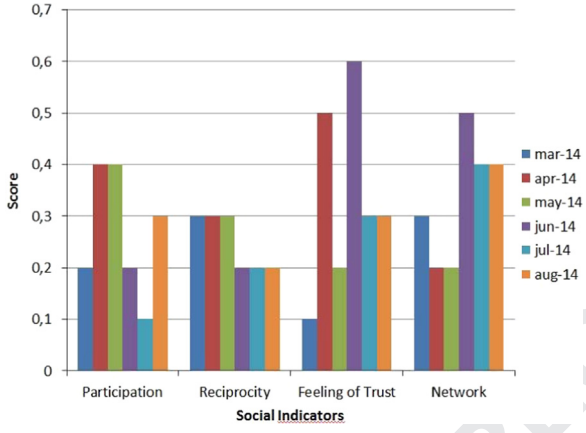

Fig. 17. L'Aquila SUN – social indicators (part 2).

The main outcome of the experiment is that, despite its simplicity, the lexicon-based approach outperforms the text classifier for 4 out of 5 social indicators. Results are particularly significant especially for the *Trust* and *Citizen Power* indicators. This behavior is due to the fact that the nature of the input makes very difficult to learn a reliable classification model, since labeled examples are very noisy and short (a Tweet can contain only 140 characters. The average length of the input examples was around 9 features).

On the other side, a simple but effective approach based on the calculation of the overlap between lexicons is able to provide a more precise classification. The effectiveness of the approach is also confirmed by the lower number of false positives (12.2%, against 17.7% obtained by the text classifier). Indeed, in such scenario, it is very important that the overall score of the social indicator depends on the sentiment of the social content which actually refers to it, thus a very low number of false positive is very important. These outcomes are not surprising since the effectiveness of lexicon-based approaches for real-time text analytics was already assessed in a similar work by Melville et al. [39], in which a platform operated by UNICEF Uganda for the automatic classification of SMS messages is described.

To sum up, the overall results in terms of *F*1-measure were encouraging and confirmed the effectiveness of our content classification algorithm. The best version of the algorithm has been integrated in the framework and has been exploited to calculate the score associated to each *social indicator*.

6. Conclusions and future work

In this work we presented CrowdPulse, a framework for real-time analysis of human-generated textual streams. We showed the architecture of the framework as well as the design choices behind all the modules which compose it. We designed a framework where the combined use of techniques for semantic representation and sentiment analysis can rapidly provide a valuable and reliable snapshot of people feelings, opinions and sentiments in several domains. As semantic representation we proposed a new methodology based on the combination of several entity linking algorithms, while for sentiment analysis an approach based on the combination of two state-of-the-art lexicons has been presented.

We also showed two real use cases of the framework: the Italian Hate Map and the SUN project for the city of L'Aquila. In both cases the framework showed its effectiveness, since it was able to easily reach the project goals with just some simple adaptation to domain-specific requirements. As regards the Italian Hate Map project, we gave a real evidence of the power and the effectiveness of tools and techniques for monitoring and mining data for social goods. The outcomes coming from psychological analysis of the maps provide several valuable insights to learn more about intolerant behaviors and to prevent it through specific initiatives. In this specific scenario, thanks to the analysis of big data, it has been possible to aggregate rough information about user intents and behaviors in order to build a valuable snapshot describing the current situation of the Italian country. On the other side, for SUN project, we showed that an integrated and multi-disciplinary approach combining psycho-social research with computer science can be exploited for mining social data to obtain a valuable and interesting snapshot of people feelings, sentiment and opinion about the current state of the town. Moreover, those information can be exploited to plan some specific intervention aimed at empowering or recovering the situation of the indicator whose score gets worse over time. We can also state that the goodness of the design choices was confirmed by the experiments performed in both scenarios.

However, the work is still ongoing so there is a lot of space for future development. The very modular structure of the framework makes very easy to extend it by introducing more modules to process or to analyze the content gathered from social networks. As an example, we plan to extend processing modules by introducing techniques for social network analysis and to link content-based information with the information coming from the Linked Open Data (LOD) cloud.

Acknowledgments

The Italian Hate Map project was made possible thanks to cooperation between the Department of Computer

Science of the University of Bari, Department of Dynamic and Clinical Psychology[25] of the University of Rome, University of Milan and Vox Diritti[26], an agency involved in human and civil rights-related projects.

Appendix A

A.1. Lexicons

In this subsection we provide a partial lexicon for the *Racism* intolerance dimension[27]. The lexicon is provided in Table 4. As shown in the table, some of the terms are very culture-dependant (as Southerner, which is used in Italy to insult people from Southern Italy) or language-dependant (as *kebabbaro*, the person who cooks *kebab*, which is typically used to insult people from Turkey and the Middle East).

A.2. Analytics console – output

In this subsection we show the output produced by CrowdPulse for both scenarios. For the Italian Hate Map project, all the Hate Maps are reported, while for L'Aquila Social Network a chart plotting the evolution of the indicators in the timelapse taken into account within the project is presented.

Even if the psychological analysis and the description of the concrete outcomes of both projects are out of the scope of this paper, a quick analysis of the images can provide some interesting findings: as regards the Italian Hate Map project, each of the images (from Figs. 11–15) adopts the above-described *heat map* formalism to identify the areas with a higher ratio of intolerant Tweets. Specifically, all the maps share a common interesting pattern, since the area around Milan and Lombardy Region is colored red regardless of the specific intolerance dimension. It is also worth to note that the map showing the geolocalization of Tweets against women is the one with the highest amount of Tweets, specifically in the areas around Naples and in Northern Italy. As regards L'Aquila Social Urban Network project, Figs. 16 and 17 summarize the score obtained by each social indicator according to the social content extracted and classified by our framework. Generally speaking, there is no clear trend in the data. This means that the satisfaction of the citizens as well as the social capital of the city is not getting recovered over time (with the exception of the Citizen Power indicator). All the indicators have a peak, but it is likely that it depends on particular events happening in L'Aquila that influence people's Tweeting activity in a specific (short) period of time, thus influencing the overall score of an indicator in a certain month.

However, regardless of the patterns emerging from the output, these figures give a real evidence of the effectiveness of the platform in carrying out the task of each scenario. Indeed, thanks to this output, it is possible to get a general overview of the data extracted from social networks. Finally, thanks to the exploitation of Machine Learning techniques, complex phenomena as those we took into account can be examined and understood from different (novel) point of views.

[25] http://www.psicologia1.uniroma1.it/

[26] http://www.voxdiritti.it/

[27] We do not provide the lexicons for the other intolerance dimensions due to the explicit terms they contain

References

[1] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, Ke Tao, Twitcident: fighting fire with information from social web streams, in: Proceedings of the 21st International Conference Companion on World Wide Web, ACM, 2012, pp. 305–308.

[2] Charu C. Aggarwal, Tarek Abdelzaher, Social sensing, in: Managing and Mining Sensor Data, 2013, Springer, pp. 237–297.

[3] M. Albakour, Craig Macdonald, Iadh Ounis, et al., Identifying local events by using microblogs as social sensors, in: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, 2013, pp. 173–180.

[4] Raian Ali, Carlos Solis, Mazeiar Salehie, Inah Omoronyia, Bashar Nuseibeh, Walid Maalej, Social sensing: when users become monitors, in: Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering, ACM, 2011, pp. 476–479.

[5] Leonardo Allisio, Valeria Mussa, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Felicittà: visualizing and estimating happiness in Italian cities from geotagged tweets, in: ESSEM@ AI* IA, Citeseer, 2013, pp 95–106.

[6] M.G. Armentano, D. Godoy, A.A. Amandi, Followee recommendation based on text analysis of micro-blogging activity, Inf. Syst. 38 (8) (2013) 1116–1127.

[7] Stefano Baccianella, Andreai Esuli, Fabrizio Sebastiani, SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of LREC, vol. 10, 2010, pp. 2200–2204.

[8] David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[9] Johan Bollen, Huina Mao, Alberto Pepe, Modeling Public Mood and Emotion: Twitter Sentiment and Socio-economic Phenomena, 2011, pp. 450–453.

[10] Johan Bollen, Huina Mao, Xiaojun Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.

[11] Julian Brooke, A semantic approach to automated text sentiment analysis (Ph.D. thesis), Simon Fraser University, 2009.

[12] Erik Cambria, Amir Hussain, Sentic Computing, Springer, 2012.

[13] Erik Cambria, Daniel Olsher, Dheeraj Rajagopal, Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis, in: AAAI, Quebec City, 2014, pp. 1515–1521.

[14] Andrew T. Campbell, Shane B. Eisenman, Nicholas D. Lane, Emiliano Miluzzo, Ronald A. Peterson, People-centric urban sensing, in: Proceedings of the Second Annual International Workshop on Wireless Internet, ACM, 2006, p. 18.

[15] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber, Bigtable: a distributed storage system for structured data, ACM Trans. Comput. Syst. 26 (2) (2008) 4.

[16] Kristina Chodoro, MongoDB: The Definitive Guide, O'Reilly Media, Inc., 2013.

[17] Hafedh Chourabi, Taewoo Nam, Shawn Walker, José Ramón Gil-Garcia, Sehl Mellouli, Karine Nahon, Theresa A. Pardo, Hans Jochen Scholl, Understanding smart cities: An integrative framework, in: 2012 45th IEEE Hawaii International Conference on System Science (HICSS), 2012, pp. 2289–2297.

[18] Xiaowen Ding, Bing Liu, Philip S. Yu, A holistic lexicon-based approach to opinion mining. in: Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, 2008, pp. 231–240.

[19] Andrea Esuli, Fabrizio Sebastiani, SentiWord-Net: a publicly available lexical resource for opinion mining, in: Proceedings of LREC, vol. 6, 2006, pp. 417–422.

[20] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.

[21] Paolo Ferragina, 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, ACM, 2010, pp. 1625–1628.

[22] Raghu K. Ganti, Nam Pham, Hossein Ahmadi, Saurabh Nangia, Tarek F. Abdelzaher, Greengps: a participatory sensing fuel-efficient maps

application, in: Proceedings of the Eighth International Conference on Mobile Systems, Applications, and Services, ACM, 2010, pp. 151–164.

[23] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, Noah A Smith, Part-of-speech tagging for twitter: annotation, features, and experiments, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, vol. 2, Association for Computational Linguistics, 2011, pp. 42–47.

[24] Daniel Gruhl, Meena Nagarajan, Jan Pieper, Christine Robson, Amit Sheth, Multimodal social intelligence in a real-time dashboard system, VLDB J.: The International Journal on Very Large Data Bases 19 (6) (2010) 825–848.

[25] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, Samee Ullah Khan, The rise of big data on cloud computing: review and open research issues, Inf. Syst. 47 (2015) 98–115.

[26] Abdelsalam Helal, Diane J. Cook, Mark Schmalz, Smart home-based health platform for behavioral monitoring and alteration of diabetes patients, J. Diabetes Sci. Technol. 3 (1) (2009) 141–148.

[27] José M. Hernández-Muñoz, Jesús Bernat Vercher, Luis Muñoz, José A. Galache, Mirko Presser, Luis A. Hernández Gómez, Jan Pettersson, Smart Cities at the Forefront of the Future Internet, Springer, 2011.

[28] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al., A Practical Guide to Support Vector Classification, 2003.

[29] Xia Hu, Lei Tang, Jiliang Tang, Huan Liu, Exploiting social relations for sentiment analysis in microblogging, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM, 2013, pp. 537–546.

[30] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, Patrick Meier, Practical extraction of disaster-relevant information from social media, in: Proceedings of the 22nd International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2013, pp. 1021–1024.

[31] Thorsten Joachims, Text categorization with support vector machines: learning with many relevant features, in: ECML, 1998, pp. 137–142.

[32] August Joki, Jeffrey A. Burke, D. Estrin, Campaignr: A Framework for Participatory Data Collection on Mobile Phones, Center for Embedded Network Sensing, 2007.

[33] Salil S. Kanhere, Participatory sensing: crowdsourcing data from mobile smartphones in urban spaces, in: 2011 12th IEEE International Conference on Mobile Data Management (MDM), vol. 2, IEEE, 2011 pp. 3–6.

[34] Rick Lawrence, Prem Melville, Claudia Perlich, Vikas Sindhwani, Steve Meliksetian, P. Hsueh, Yan Liu, Social media analytics, OR/MS Today (2010) 26–30.

[35] Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, Eric. Shook, Mapping the global twitter heartbeat: the geography of twitter, First Monday 18 (5) (2013).

[36] Christopher Manning, Hinrich Schütze, Text categorization, in: Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, USA, 1999, pp. 575–608 (Chapter 16).

[37] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela H. Byers, Big Data: The Next Frontier for Innovation, Competition, and Productivity, Mckinsey Institute Report, 2011.

[38] Olena Medelyan, Catherine Legg, Integrating cyc and wikipedia: folksonomy meets rigorously defined common-sense, in: Proceedings of the WIKI-AI: Wikipedia and AI Workshop at the AAAI'08 Conference, Chicago, USA, 2008.

[39] Prem Melville, Vijil Chenthamarakshan, Richard D. Lawrence, James Powell, Moses Mugisha, Sharad Sapra, Rajesh Anandan, Solomon Assefa, Amplifying the voice of youth in africa via text analytics, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 1204–1212.

[40] Prem Melville, Vikas Sindhwani, R. Lawrence, Social media analytics: channeling the power of the blogosphere for marketing insight, in: Proceedings of the WIN, 2009.

[41] Pablo N. Mendes, Max Jakob, Andrés García-Silva, Christian Bizer, DBpedia spotlight: shedding light on the web of documents, in: Proceedings of the Seventh International Conference on Semantic Systems, ACM, 2011, pp. 1–8.

[42] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge, in: Proceedings of the 16th ACM Conference on Information and Knowledge Management, ACM, 2007, pp. 233–242.

[43] George A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.

[44] Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, Theresa Wilson, Semeval-2013 Task 2: Sentiment Analysis in Twitter, 2013.

[45] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, Noah A. Smith, From tweets to polls: linking text sentiment to public opinion time series, in: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM), vol. 11, 2010, pp. 122–129.

[46] Franco Orsucci, Giulia Paoloni, Mario Fulcheri, Mauro Annunziato, Claudia Meloni, Smart Communities: Social Capital and Psycho-Social Factors in Smart Cities, 2012.

[47] Alexander Pak, Patrick Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA), May 2010.

[48] Bo Pang, Lillian Lee, Opinion mining and sentiment analysis, in: Foundations and Trends in Information Retrieval, vol. 2, nos. 1–2, 2008, pp. 1–135.

[49] Cecile Paris, Stephen Wan, Listening to the community: social media monitoring tasks for improving government services, in: CHI'11 Extended Abstracts on Human Factors in Computing Systems, ACM, 2011, pp. 2095–2100.

[50] Michael J. Paul, Mark Dredze, You are What You Tweet: Analyzing Twitter for Public Health, 2011, pp. 265–272.

[51] Philips Kokoh Prasetyo, Ming Gao, Ee-Peng Lim, Christie Napa Scollon, Social sensing for urban crisis management: the case of singapore haze, in: Social Informatics, Springer, 2013, pp. 478–491.

[52] Kiran K. Rachuri, Christos Efstratiou, Ilias Leontiadis, Cecilia Mascolo, Peter J. Rentfrow, Smartphone sensing offloading for efficiently supporting social sensing applications, in: Selected Papers from the 11th Annual {IEEE} International Conference on Pervasive Computing and Communications (PerCom 2013), Pervasive Mob. Comput. 10 (Part A) (2014) 3–21.

[53] Delip Rao, Paul McNamee, Mark Dredze, Entity linking: finding extracted entities in a knowledge base, in: Multi-source, Multilingual Information Extraction and Summarization, Springer, 2013, pp. 93–115.

[54] Sara Rosenthal, Preslav Nakov, Alan Ritter, Veselin Stoyanov, Semeval-2014 task 9: sentiment analysis in twitter, in: Proceedings of SemEval, 2014.

[55] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in : Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 851–860.

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

[57] Dongyoun Shin, Daniel Aliaga, Bige Tuner, Stefan Müller Arisona, Sungah Kim, Dani Zünd, Gerhard Schmitt, Urban sensing: using smartphones for transportation mode classification. Computers, Environment and Urban Systems (2014).

[58] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, The Hadoop distributed file system, in: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), IEEE, 2010, pp. 1–10.

[59] Viktor Slavkovikj, Steven Verstockt, Sofie Van Hoecke, Rik Van Walle, Review of wildfire detection using social media, Fire Saf. J. 68 (2014) 109–118.

[60] Stefan Stieglitz, Linh Dang-Xuan, Social media and political communication: a social media analytics framework, Soc. Netw. Anal. Min. 3 (4) (2013) 1277–1291.

[61] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2) (2011) 267–307.

[62] Peter D. Turney, Patrick. Pantel, From frequency to meaning: vector space models of semantics, J. Artif. Intell. Res. 37 (2010) 141–188.

[63] Hanna M. Wallach, Topic modeling: beyond bag-of-words, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 977–984.

[64] Stanley Wasserman, Social Network Analysis: Methods and Applications, vol. 8, Cambridge University Press, 1994.

[65] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, Robert Power, Using social media to enhance emergency situation awareness, IEEE Intell. Syst. 27 (6) (2012) 52–59.

[66] Daniel Zeng, Hsinchun Chen, Robert Lusch, Shu-Hsing Li, Social media analytics and intelligence, IEEE Intell. Syst. 25 (6) (2010) 13–16.

[67] Cai-Nicolas Ziegler, Michal Skubacz, Towards automated reputation and brand monitoring on the web, in: IEEE/WIC/ACM International Conference on Web Intelligence, 2006. WI 2006, IEEE, 2006, pp. 1066–1072.

Please cite this article as: C. Musto, et al., CrowdPulse: A framework for real-time semantic analysis of social streams, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.06.007