Felicittà: Visualizing and Estimating Happiness in Italian Cities from Geotagged Tweets

Leonardo Allisio, Valeria Mussa, Cristina Bosco,
Viviana Patti and Giancarlo Ruffo

Università degli Studi di Torino
Dipartimento di Informatica
c.so Svizzera 185, I-10149 Torino (Italy)
{bosco,patti,ruffo}@di.unito.it,
{leonardo.allisio,valeria.mussa}@studenti.unito.it

Abstract. Felicittà[1] is an online platform for estimating happiness in the Italian cities, which uses Twitter as data source and combines sentiment analysis and visualization techniques in order to provide users with an interactive interface for data exploration. In particular, Felicittà daily analyzes Twitter posts and exploits temporal and geo-spatial information related to Tweets in order to easy the summarization of sentiment analysis outcomes and the exploration of the Twitter data. By interactive maps it provides users with the possibility to have a comprehensive overview of the sentiment analysis results about the main Italian cities, and with the opportunity to zoom-in to a specific region to visualize a fine-grained map of the city or district as well as the location of the individual sentiment-labeled Tweets. The platform allow users to tune their view on such huge amount of information and to interactively reduce the inherent complexity, possibly providing an hint for finding meaningful patterns, and correlations between moods and events.

Keywords: Data Analysis and Visualization, Twitter, Sentiment Analysis

1 Introduction

The huge amount of information streaming from online social networking and micro-blogging platforms such as Twitter, are increasingly attracting the attention of many kinds of researchers and practitioners, such as sociologists, psychologists, communication and political scientists, as well as data journalists and computational linguistics scholars. From different perspectives, each observer looks for relationships from massively user generated data, in order to get insights that can lead to new conjectures, correlations, and causalities.

Even if the debate on how the social networking users generated content can be representative of the human behavior in everyday life is still open, a computational framework that analyzes and visualizes information at customizing scales

[1] The name Felicittà is a fusion of two Italian words, "felice" (*happy*) and "città" (which corresponds to both *city* and *town*).

of summarization is a valuable tool for experts with different backgrounds that do not want to deal directly with raw data or abstract models; in fact, statistical tools, machine learning techniques, large-scale network analysis and natural language modeling are hard to be applied by many skilled sociologists and communication scientists. On the other hand, a lot of quantitative research has been conducted on data, but the connection with theories that would qualitatively explain the observed and modeled phenomena is often missing.

In this paper, we introduce *Felicittà*, an online platform that allows the user to explore the result of sentiment analysis performed over geotagged Tweets. The contribution is two-fold. First, we propose a fully implemented visualization system to estimate the level of happiness in a given geographical area based on geotagged Tweets, that has been engineered in a modular way, and can overlay different analysis engines. Second, we describe a particular instantiation of the system, where a sentiment analysis engine for detecting Italian Tweets' sentiment polarity has been developed and employed in order to estimate happiness in Italian cities. Visualization techniques are adopted to support researchers and practitioners to explore the data about happiness in Italian cities. Maps, plots, tag clouds and other charts can be interactively requested on demand for giving more contextual and quantitative details. The paper is organized as follows: Section 2 contains a brief overview of related work. Section 3 describes the modules of the computational framework devoted to Tweets' retrieval and sentiment analysis. The front-end module of our online service, which provides summarization and visualization of user sentiments in Italian cities, is outlined in Section 4.Brief conclusions end the paper.

2 Related Work

Relevant contributions for the issues addressed in Felicittà can be found in a wide range of disciplines, ranging from computational linguistics and sociology to advanced interfaces for big data exploration.

The investigation of correlations between Twitter geotagged data (expressing in real-time sentiment of individuals) and emotional, geographic, demographic, and health characteristics has been matter on an interesting recent work [8], where a team of researchers analyzed ten million Tweets to map happiness in the United States and to create a sort of *geography of happiness*. The happiness score, computed by applying sentiment analysis techniques, is measuring something different but perhaps complementary to traditional survey-based techniques. Based on the Tweet analysis happiness maps of the United States are produced, but they are not web accessible neither inspectable in an interactive way. On the same line, an ongoing study is focussing on the geography of hate, by analyzing geotagged hateful Tweets in the United States. The resulting hate heat map shows where racist, homophobic tweets come from, and is available on-line: `http://users.humboldt.edu/mstephens/hate/hate_map.html`.

In Italy, the issue of measuring the happiness in the cities by analyzing Italian Tweets has been addressed recently in the context of the Voices from the Blog project[2] and lead to the development of an iPhone/iPad app, called iHappy. In this application, results of the sentiment analysis on Tweets are summarized and visualized to users mainly by *static* maps of happiness, by showing happiness indicators related to Italian cities, provinces or region in a seven-day time window, and by providing daily (or weekly) region or city happiness rankings.

The main contribution of Felicittà w.r.t. to the above mentioned systems is two-fold. On the one hand, the visualization component of Felicittà (Section 4) provides introduces more *zoom & filtering* tools and *details-on-demand* capabilities, that support users in the activity of interactively inspecting Tweet-generated happiness maps. On the other hand, our platform has been engineered to be modular w.r.t. Tweet-based analysis engines, and, in principle, can overlay sentiment analyzers specialized for different languages and domains.The issue of identifying key features of an information visualization tool has been matter of a pioneering work of the information visualization researcher Ben Schneiderman [13]. An information visualization tool should allow users "to gain an *overview* of the data under study, provide *zoom and filtering* capabilities, item-level *details-on-demand*, allow users to see *relationships* among items in a collection, and *extract* target data about specific subsets within the collection" [13]. On this line, recent works deal with the visualization of Twitter data [6, 7, 11], but without a specific interest on the estimation of happiness in given areas, which is the focus of the present work.

For what concerns the technical sentiment analysis task, comprehensive surveys are available in literature [3, 5, 16]. Only very recently some works focussed on analyzing the sentiment in *Italian* Tweets [1, 2, 9], let us mention, among others, the work in [2] which focusses on irony detection, or [9], which addresses the task of modeling political disaffection in Italy. The approach used in Felicittà is lexicon-based. The evaluation of the polarity of a Tweet is based on the word's polarity, and supported by state-of-the-art lexical and affective resources, i.e. MultiWordNet and WordNet-Affect [14], as we will explain below.

3 The Heart of Felicittà: the Sentiment Analyzer

Felicittà aims at automatic mining and estimating happiness of people living in a given location from the Tweets posted in that geographic area. The architecture is sketched in Fig. 1. Geotagged information is retrieved from social media APIs. Such data can be enriched and contextualized with other information accessible from a wide range of data sources, such as official statistics produced to offer services to citizens and policy makers. For example, ISTAT (Italian National Institute of Statistics) is the main producer of socio demographic information, and their data can be directly accessed. At the core of our architecture there is the analyzer that receives as input the data gathered from different sources and produces an estimation of the general sentiment at different geographical and temporal scales. The analysis layer communicates with other services through

[2] `http://www.blogsvoices.unimi.it/`

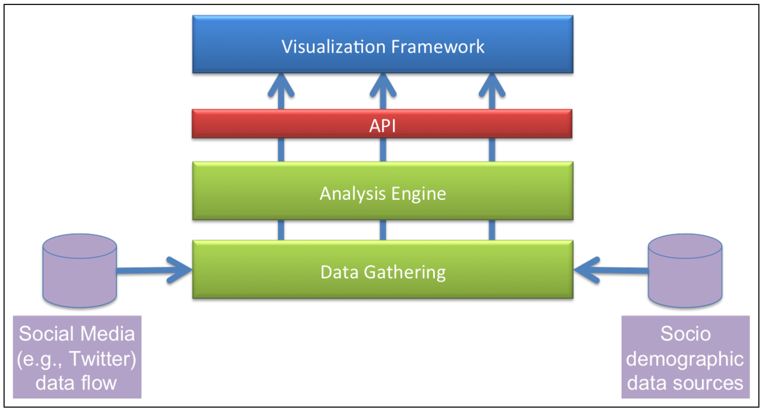


Fig. 1: Architecture of the Felicittà's framework

an API, keeping modules logically separated and easy to be modified or substituted without affecting other layers. On top of the architecture we have our visualization framework that is presented in the next section.

We focused on the main Italian towns, i.e. the 110 administrative Italian centers. By exploiting Twitter's APIs, the system collects daily all the Tweets freely downloadable (450,000) geolocated in these towns, and performs three steps of analysis for each Tweet[3] in order to classify it as positive or negative. At the end, it aggregates the polarity of all the Tweets according to their geotagging and thus evaluates the happiness of each town and region.

The pipeline of Felicittà's analyzer includes four steps (Fig. 2). The first one is the collection of the Tweets to be analyzed: assuming that each town T can be identified by latitude and longitude values, all the messages posted in a range of 10 km from the center of T the previous day are daily collected, by applying the Twitter's search function, and geolocated. In the second step the collected Tweets are cleaned by deleting emoticons, links, mentions of other users and redundant punctuation. Emoticons which are clearly expressing some emotion are substituted by the more similar emotional words, in order to maintain their affective value. Mentions, usually preceded by '@', are instead substituted by the word 'user'. Table 1 shows some example.

The third step consists in parsing the cleaned Tweets by Freeling[4], an open source tool for morpho-syntactic analysis of Italian and other languages, developed at the University of Catalunya (Spain). In particular, the grammatical category of each word is recognized allowing for the association with a lemma to be searched in the affective lexicon in the next step, e.g. the word "odio" (*hate*) is recognized as a Verb and associated with the lemma "odiare" (*to hate*). The

[3] In this paper, we capitalize the T in Tweet and Twitter as requested in the *Twitter Trademark and Content Display Policy* at `https://twitter.com/logo`.

[4] `http://nlp.lsi.upc.edu/freeling/`

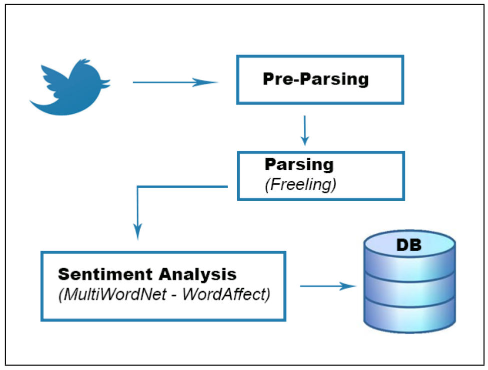


Fig. 2: Sentiment analyzer's pipeline

<table>
<tr><th>Regular Expression</th><th>Interpretation</th><th>Substitution</th></tr>
<tr><td><code>:[-]?[D]+</code></td><td><code>:D</code></td><td>gioia (joy)</td></tr>
<tr><td><code>[:;][-]?[Pp]+</code></td><td><code>:P ;P</code></td><td>ironia (irony)</td></tr>
<tr><td><code>[;T][_]+[;T]</code></td><td><code>;_; T_T</code></td><td>tristezza (sadness)</td></tr>
<tr><td><code>>[_]+<</code></td><td><code>>_<</code></td><td>rabbia (anger)</td></tr>
<tr><td><code>(mw|MW)*[hH]*([aAeEiI]+[hH]+)+[aAeEiI]*</code></td><td><code>ahah mwahah eheh</code></td><td>risata (laugh)</td></tr>
</table>

Table 1: Expressions and substitutions in pre-parsing

recognition of lemmas is especially needed for morphologically rich languages like Italian where the phenomenon of inflection, even if more notable in the case of Verbs, affects a large variety of grammatical categories where the agreement is required in the most of cases[5].

In the fourth step the sentiment analysis is applied on Tweets. First, all the *content words* of each Tweet, i.e. words carrying semantic content, and therefore useful for the detection of the affective meaning, are extracted. They are Nouns, Verbs, Adverbs and Adjectives associated to lemmas in the previous step. Second, for each extracted lemma a query is done on a lexical database in order to find its meaning(s) that can be related to some affective concept. The resources developed for Italian we exploited for this task are MultiWordNet[6] and WordNet-Affect [10, 14]. The former is a lexicon built at the FBK (Fondazione Bruno Kessler, Trento, Italy) for Italian and other languages, aligned with the WordNet database developed for English at Princeton University [4]. It is organized according to the grammatical category of words and for each lemma it includes the meaning(s) that the lemma can assume. WordNet–Affect is instead the affective extension of WordNet domains, aligned with MultiWord-

[5] Think for instance to Adjectives and Determiners that agree with Nouns or Verbs in participle (past and present) which agree with subject Nouns or ProNouns.

[6] `http://multiwordnet.fbk.eu/english/home.php`

Net, that associates to some lemmas affective concepts. If a lemma L occurs in MultiWordNet, the query results in a list of one or more *meanings*:

$$L =< m_1 ... m_n >$$

Each m_i is then searched in WordNet–Affect and when the association with an affective concept is found, an affective evaluation is expressed using one *polarity* label[7] among 'negative' (-1), 'neutral' (0), or 'positive' ($+1$):

$$p(m_i) = -1/0/+1$$

The *polarity of the lemma* is then described by the following formula:

$$p(L) = \begin{cases} -1 & \text{if } \sum_{m \in L} p(m) < 0 \\ 0 & \text{if } \sum_{m \in L} p(m) = 0 \\ 1 & \text{otherwise} \end{cases}$$

On the one hand, $p(L)$ is an indication of the prevailing polarity of the m_i associated by WordNet–Affect to L. Moreover, it should be observed that in each Tweet for each L, also associated with several meanings, only one single m_i is realized.

While the above described part of sentiment analysis is referred to single words only, the last part of the process concerns each full Tweet T considered as a bag of lemmas. The *polarity of a Tweet* T is described by the following formula:

$$p(T) = \begin{cases} -1 & \text{if } \sum_{L \in T} p(L)/m \leq -\epsilon \\ 0 & \text{if } -\epsilon < \sum_{L \in T} p(L)/m < \epsilon \\ 1 & \text{otherwise} \end{cases}$$

where m is the number of lemmas of T, and ϵ is an empirical estimable small value constant that may allow to extend the range of Tweets to be labeled as neutral (in our platform, we set ϵ to a temporary value of 0, because we need a more accurate evaluation in order to make such estimation). The polarity of a city/region is defined as the rate of positive Tweets geolocated in the city/region w.r.t. the total amount of Tweets geolocated in that area.

It should moreover be observed that because of the size of the WordNet–Affect lexicon, which includes around 4,700 words, the affective polarity can be detected only in a limited part of the Tweets collected every day by Felicittà. The results are based on around 35,000 out of the 450,000 daily collected Tweets.

[7] For the purpose of the present work, exploiting the hierarchical organization of the database, we have limited the granularity of the affective concepts included in the WordNet–Affect classifying with these three labels only a broad notion of polarity.

3.1 An Example

We conclude by showing the result of each step above described on a Tweet of our corpus. In Fig. 3, first, you can see the original Tweet (*I HATE that horrible place, I will be satisfied when it will be razed.* `>_<`). Then, it is shown the substitution of the emoticon `>_<` in the pre–parsing step, and how the Freeling parser analyzes and split the Tweet in columns in order to associate to each word the lemma and the morpho-syntactic features. We can see the result of the query on MultiWordNet and WordNet-Affect: two lemmas of the Tweet are detected as affective, that is "odiare" (*to hate*) and "rabbia" (*rage*), the former with a single meaning and the latter with two meanings, all with negative polarity. Finally, the polarity evaluation for the Tweet is reported.

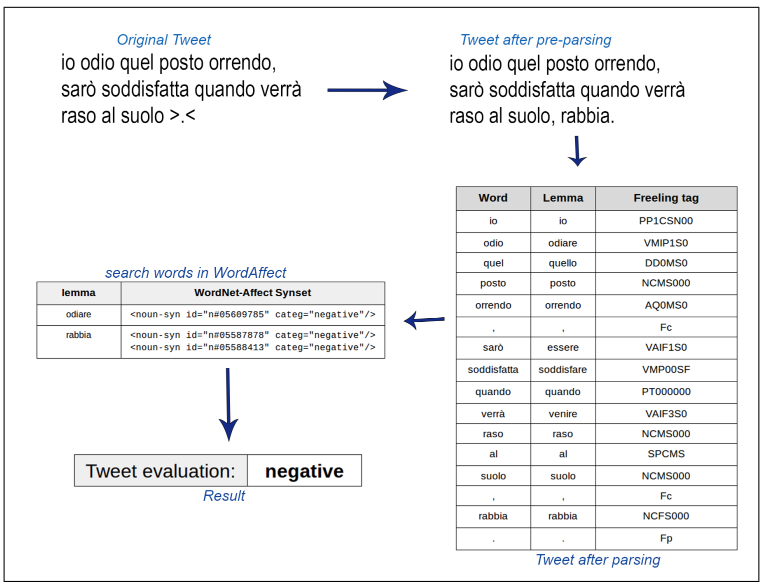


Fig. 3: A Tweet parsed by Freeling.

4 Visualizing and Interacting with Estimated Happiness

In this section we introduce the interactive user interface designed for geotemporal visualizations of happiness in Twitter. The results of Felicittà are displayed to the user according to different perspectives, as you can see at `http://www.felicitta.net`. The main views available to the user on the web site are three: Città (*Towns*), Regioni (*Regions*) and Top10.

The view Città (see Fig 4) is a map of Italy where the user can find a round marker for each town, which assumes different colors ranging from blue to red

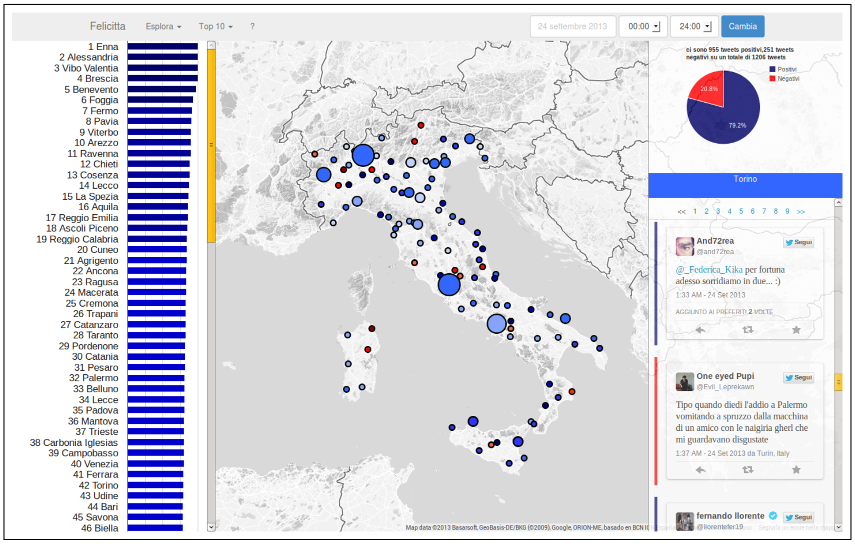


Fig. 4: Felicittà view Città

to represent the affective status, and different size to display the larger/smaller amount of messages posted by the users of the town. Also an ordered score of all the Italian towns is shown for each given day. By selecting a town on the map the user can see more details about the evaluation expressed by the system and also the Tweets posted there to test the reliability of the system evaluation.

The view Regioni (see Fig 5) shows instead the affective status of full regions calculated on the basis of the results obtained for all the towns of the region itself. This can be displayed in two different ways, a map of Italy where the regions are colored according to their affective status, as for towns, or in a score of the regions from the happiest to the less happy. The last view, i.e. Top10, displays the Tweets as geolocated on Italy. Also this view includes a map and a score. The map shows all the Tweets positioned within the area where they have been posted. By moving on the map, the user can zoom in to find the exact position of Tweets on the map and also to read them. Each Tweet is here represented by a marker colored according to its detected affective polarity (see Fig. 6). Nevertheless, only posts which are associated with a precise geolocation can be seen in the Top10 view, while several others, for which the geolocation is only referred to a town, are exploited in the computation of the affective polarity of a geographic area, but cannot be seen in this view by the user.

4.1 Looking for Relationships and Details on Demand

According to the visualization features identified by Schneiderman in [13], our interface includes several ways to expand other details, that may help the observer to look for hidden relationships and correlations. If we focus on a given

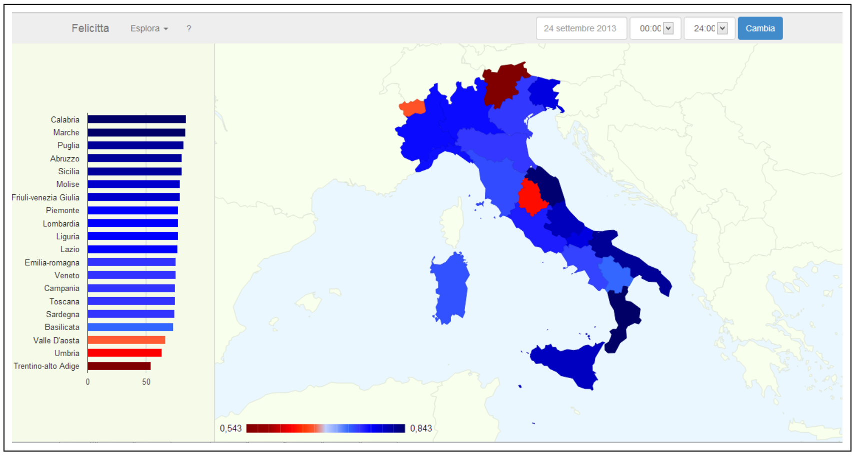


Fig. 5: Felicittà view Regions

city (for example: Turin), we can browse the Tweets that have been geotagged there using the rightmost column displayed in the city map view (See Fig. 4). However, we can further expand our search for information on a temporal dimension. In fact, we can view the number of Tweets that have been analyzed in a period of time (e.g., August 2013), as in Fig. 7.

If we want to observe how the level of happiness varies in time, we can open quarterly and daily based views. In the given example, and according to the sample of Twitter users we analyzed, August has been the happiest summer month in Turin (Fig. 8a). However, the distribution plotted on a daily basis (Fig. 8b), shows that happiness does not exhibit a regular behavior, and we have an happiest day (08/09) and a saddest one (24/09).

We decided to display in our framework also tag-clouds, a visual representation useful for quickly perceiving the most prominent terms involved in analyzed Tweets. For example, in Fig. 9a and 9b we show some of the words used in Turin during the happiest and saddest days of August 2013. Words are displayed with different sizes according to the number of their occurrences in the Tweets posted during those days. Quite interestingly, on 08/24, the soccer player Carlos Tévez made his first appearance with Juventus (one of the teams quartered in Turin), scoring the winning goal and beating U.C. Sampdoria 1-0 in their opening match of the 2013/14 season. It worths to be noted that Tévez wears the shirt number 10, i.e., "maglia numero 10" (see 9b). Maybe, the general bad mood reflected in Tweets posted in Turin in that day can be explained by the disappointment of the many citizens that do not support Juventus or by the anxiety ("ansia") resulted by the long-awaited season opening. However, the given interpretation is much beyond the scope of our tool, and the empirical co-occurrence of particular words in charts and a corresponding polarity evaluation can be used only to raise issues for further investigation.

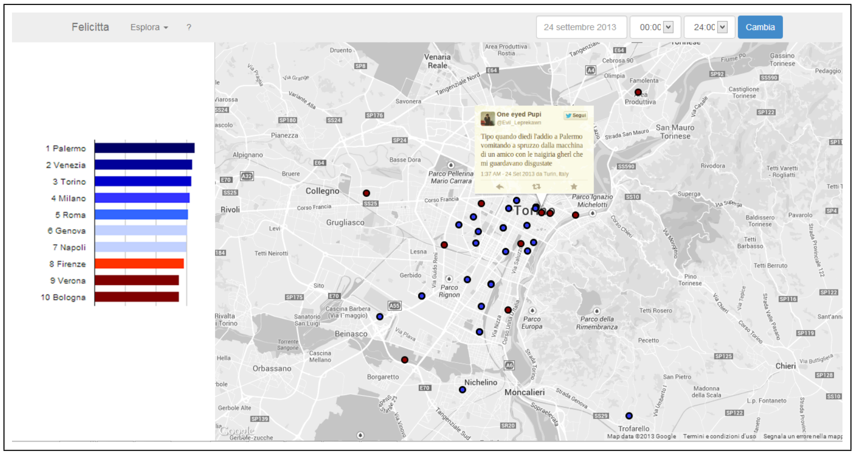


Fig. 6: Felicittà view Top10

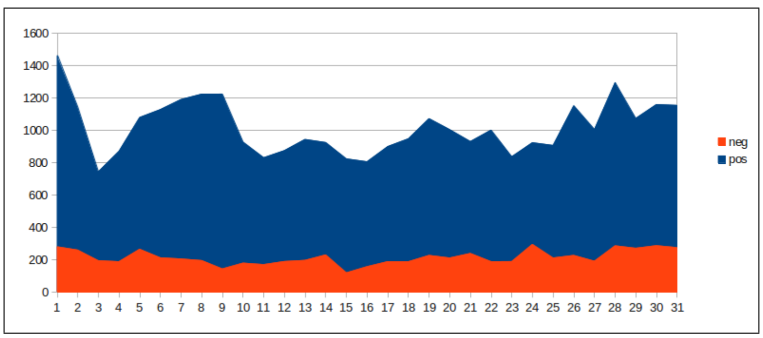
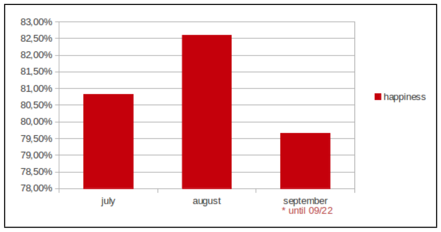


Fig. 7: August 2013: number of positive/negative Tweets in Turin.

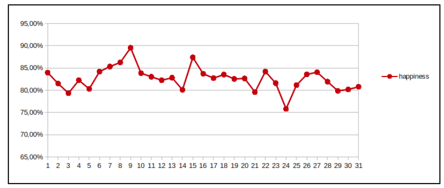
5 Conclusion and Future Work

The paper describes an online platform for estimating happiness in the Italian cities, which uses Twitter as data source and combines sentiment analysis and advanced visualization techniques in order to provide users with an interactive interface for the exploration of the resulted data.

For what concerns the visualization module of Felicittà, we aim at improving the browsing experience of the user as regards time-oriented information, by allowing users to visualize data within the specified window or time period, rather than within a single day as in the actual implementation. Moreover, we are studying the embedding of tools that enable users to personalize (and export) the views, the graphs and the statistics offered by Felicittà, according to their needs and goals, with the main aim to offer a richer support to sociologists, researchers,



(a) Quarterly happiness. Period: June–September 2013.



(b) Daily happiness. Month: August 2013. Happiest day: 08/09. Saddest day: 08/24.

Fig. 8: Views at different time scales in Turin

(a) August 09, 2013

(b) August 24, 2013

Fig. 9: Tag cloud of the happiest (a) and the saddest (b) days in August 2013

journalists, common users in detecting meaningful patterns, possible correlations and so on. For instance, it could be interesting to have the possibility to compare the results of Felicittà's estimation of happiness in a given district of a city and in a given time period, with other information, e.g. quality of public services in the area, events occurred in the area in the same time window, in order to give users food for thought about possible correlations between happiness expressed by Tweets and living in different districts of the same city.

For what concerns the sentiment analysis task, we currently rely on a simple lexicon-based approach, where the positive and negative polarity of a Tweet is calculated based on a dictionary of Italian words annotated with the word's polarity. In this work our first focus was on the sentiment visualization and summarization issue, but we are currently working to improve the analyzer and provide a reliable evaluation of it. For this purpose, we are developing a gold standard corpus of manually annotated Tweets that can be used as a testbed for evaluation and comparison with other systems.

Another interesting challenge to address is to apply emotion detection techniques in order to classify Tweets according to different emotions (e.g. the Ekman's basic emotions used in [12, 2], or the emotional categories from the Plutchik's model used in [15]) and to provide a sort of geography of emotions.

References

1. V. Basile and M. Nissim. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, 2013.
2. C. Bosco, V. Patti, and A. Bolioli. Developing corpora for sentiment analysis: The case of irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63, 2013.
3. E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.
4. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
5. P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha. Comparing and combining sentiment analysis methods. In *Proc. of the 1st ACM Conf. on Online Social Networks*, pages 27–38. ACM, 2013.
6. S. Kumar, F. Morstatter, and H. Liu. *Twitter Data Analytics*. Springer, New York, NY, USA, 2013.
7. K. McKelvey, A. Rudnick, M. Conover, and F. Menczer. Visualizing communication on social media: Making big data accessible. In *Proc. of CSCW Workshop on Collective Intelligence as Community Discourse and Action*, 2012.
8. L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5), 05 2013.
9. C. Monti, A. Rozza, G. Zappella, M. Zignani, A. Arvidsson, and E. Colleoni. Modelling political disaffection from twitter data. In *Proc. of Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM'13, pages 3:1–3:9, 2013.
10. E. Pianta, L. Bentivogli, and C. Girardi. MultiWordNet: developing an aligned multilingual database. In *Proc. of Int. Conference on Global WordNet*, 2002.
11. J. Ratkiewicz, M. Conover, B. Meiss, M.and Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011 (Companion Volume)*, pages 249–252, New York, NY, USA, 2011. ACM.
12. K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu. EmpaTweet: Annotating and detecting emotions on Twitter. In *Proc. of the 8th Language Resources and evaluation Conference, LREC'12*, pages 3806–3813, 2012.
13. B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *In IEEE Symposium on Visual Languages*, pages 336–343, 1996.
14. C. Strapparava and A. Valitutti. WordNet-Affect: an affective extension of WordNet. In *Proc. of the 4th Language Resources and evaluation Conference, LREC'04*, volume 4, pages 1083–1086. ELRA, 2004.
15. J. Suttles and N. Ide. Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing, CICLing 2013*, volume 7817 of *LNCS*, pages 121–136. Springer, 2013.
16. M. Taboada, J. Brooke, M. Tofiloski, K. D. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.