Exploring Sentiment in Social Media and Official Statistics: a General Framework

Emlio Sulis[1], Mirko Lai[1], Manuela Vinai[2] and Manuela Sanguinetti[1]

[1]Università degli Studi di Torino, Dipartimento di Informatica
c.so Svizzera 185, I-10149 Torino (Italy)
{sulis,milai,msanguin}@di.unito.it
[2] Q.R.S. soc. coop.
V.le C. Battisti 15, 13900 Biella
vinai@qrsonline.it

Abstract. The integration between official statistics and social media data is a challenging topic. This contribution aims to present a recently-designed framework to compare sentiment analysis on social media content with social and economic data. Such framework - which has already been applied, in a preliminary fashion, to the Felicittà project - is meant to integrate official statistics and correlate it with online social media data. Its ultimate goal, in fact, namely consists in giving a contribution to the definition of a measure of subjective well-being that could fully benefit from both traditional, well-established social indicators and dynamic data obtained from the web.

Keywords: Subjective Well-Being, Sentiment Analysis, Official Statistics, Social Media

1 Introduction

The significant growth of user-generated content on the web, and in particular the increased availability of data from online social media, has fostered the development of automatic techniques for the extraction and processing of such content for different purposes. This development is reflected, among other things, by the spread of scientific contests whose main track is the Sentiment Analysis (henceforth SA) of texts in different languages (see eg. SemEval[1] for English, and SENTIPOLC@EVALITA2014[2] for Italian).

In turn, these achievements are encompassed into a broader and interdisciplinary debate related to the study and definition of measures that could be considered as reliable indicators of the well-being of a community. In fact, a growing debate has recently involved the measurement of social and individual well-being. New statistical measures have been proposed besides the bare Gross Domestic Product (GDP), traditionally seen as the best way to measure national

[1] http://alt.qcri.org/semeval2014/task9/

[2] http://www.di.unito.it/~tutreeb/sentipolc-evalita14

economic results. Among such measures are a large amount of indicators that, in several ways and from different points of view, attempt to assess the degree of "happiness" and life satisfaction, also designated with the expression *Subjective Well-Being*, or simply SWB (see Section 2). Such measures are usually provided by governmental institutions or entitled research organizations, and they generally include social indicators measuring life quality and concerning all major areas of citizens' lives. However, such data are static and their recovery may require much efforts in terms of time and resources. Moreover, the increasing success of Sentiment Analysis (SA) techniques on social media has made it possible to develop alternative tools and measures, with respect to the latter, to assess the degree of happiness and well-being. Social media and their content can thus be used to complement and corroborate the information gathered from traditional data sources as regards SWB detection.

The work presented here is just part of this research context. In particular, the purpose of this paper is to describe a framework whose entire definition and completion is still in progress, for the analysis and assessment of the degree of "happiness" of a given community in Italy, taking into account and combining together the information gathered from two main data sources: *a)* social media content, and Twitter in particular; *b)* the socio-demographic information made available by the main suppliers of official statistical data, such as the Italian National Institute of Statistics (ISTAT)[3]. The first point in particular has actually been explored and developed under a recent project called Felicittà[4] [1], i.e an online platform for estimating happiness in Italian cities that daily analyzes Twitter posts and exploits temporal and geo-spatial information related to tweets, in order to enable the summarization of SA outcomes.

The present work is both an extension and a comprehensive reference framework of that project. As a matter of fact, its aim is manifold and includes: 1) the use and further development of techniques for the visualization of SA outcomes in Italian texts; 2) the study of the correlations between official statistics and user-generated media content; 3) providing a contribution to the debate on what can be considered effective and reliable indicators of social well-being.

The remainder of the paper is structured as follows: Section 2 provides a brief introduction to the notion of subjective well-being, summing up the more recent work carried out on this matter while Section 3 describes the whole architecture of the system, as currently conceived. Final remarks in Section 4 close the paper.

2 Background and Related Work

The present contribution covers the debate on the Subjective Well-Being as a social indicator and sheds some light on happiness studies based on the sentiment analysis of social media.

[3] http://www.istat.it/it/

[4] http://www.felicitta.net/

Subjective Well-Being. As stated in Diener [5], SWB includes reflective cognitive evaluations about the quality of life, such as life and work satisfaction, interest and engagement, and affective reactions to life events, such as joy and sadness. The common measurements of SWB are self-report methods and surveys with questionnaires. Social indicators and life quality research is a specific field of study grown over the years as witnessed, for instance, by the birth of the review "Social Indicators Research" and underlined by the initiatives of the Organization for Economic Co-operation and Development (OECD) since the Nineties[5]. Namely the OECD recently proposed a survey-based measurement of SWB at national level [16], as alternative to purely economic measures.

The well-being of the population: from Easterlin to GNH. The Gross Domestic Product (GDP) is today the main measure of the nation's economic activity. However, since late 70's, a huge debate has grown over this measure [8]. Easterlin [7] first identified the paradox for which the increase of economic well-being in wealthier countries has no further increases in subjective well-being [13]. As alternative to GDP, new concepts have arisen as sustainable socio-economic development, governance, environmental conservation and so on. Besides the OECD, several organizations and countries take into account new measurements, basically focused on the concept of *happiness*: see, for instance, the World Happiness Report [10] in a recent United Nations initiative, or the Gross National Happiness (G.N.H.) index developed by Bhuthan. In Italy, an inter-institutional initiative proposes a set of indicators on "Equitable and Sustainable Well-Being(BES)"[6].

Social Media and Well-Being. The analysis of textual expressions in social media contents on a Big-Data scale would offers an opportunity to economists and sociologists in the measurement of social well-being. There's an open debate on the topic and several works already investigated this subject with contrasting results. Wang et al. [20] examine Facebook's Gross National Happiness (FGNH) indexes and Diener's Satisfaction with Life Scale (SWLS), and finally criticize the idea that a well-being index can be based on the contents of a specific online social networks. Quercia et al.[17] explore the relationship between sentiment expressed in Twitter messages and community socio-economic well-being and, on the contrary, they found interesting correlations between sentiment and general well-being. Kramer [11] proposed a metric to represent the overall emotional health of the nation as a model of "Gross National Happiness". Our work aims to improve these studies by the analysis of a set of more extensive official statistics, better detailed in 3.2. Social media analysis also suggests the prediction of stock market [2], and of collective mood state [15]. Emotions have been considered with respect to social media and their dynamics [14] [12], also with geographical concerns [6]. We attempt to enrich and extend these studies by focusing on a

[5] See e.g. the Better Life Index `http://www.oecdbetterlifeindex.org/`

[6] The national statistical institute ISTAT and the National Council for Economics and Labour (CNEL) propose the BES Index which analyzes the changes in quality of life in Italy focusing on 12 different areas `http://www.misuredelbenessere.it/`

finer-grained administrative territorial division; as a matter of fact, our data describe the situation not only at a national level, but also with respect to regions, provinces and municipalities.

The challenge of visualization. The lecture of patterns and trends from spreadsheets or lists of numbers is a difficult task when we have to deal with large amounts of data. An improvement is often obtained by the use of graphs. Shapes and lines immediately create meanings and significance from data. In this way, data visualization allows us to present trends, to discover what is often hidden [4] and simplify the identification of patterns not easily detectable [21]. Several different tasks can be spotted in the design of a visualization system [18]. Some interesting works already dealed with social media data, highlighting aspects of public sentiment in the web [19] or public interest information[7]. Such works inspired, in their main principles, the design of our visualization module within the framework.

3 Framework Description

The present framework aims to include several approaches and techniques in order to detect the well-being of a community under a broader perspective. The steps entailed in the design phase was: a) the definition of the whole pipeline; b) the selection of data from official statistics to be correlated with the analysis performed by the SA module; c) the presentation of the most promising patterns emerging from the comparison between social media data and official statistics. In this section, we describe the general framework architecture with an overview of its modules.

3.1 Architecture

The whole framework architecture, as shown in Figure 1, consists of 5 main parts: Providers, Data Gathering, Data Analysis, Data Exposure and Data Visualization.

Providers. Providers are the data sources: i.e Twitter, from which we retrieve the geolocated[8] Italian tweets using the Stream API, and the various socio-demographic data sources (detailed in Table 1), that return demographic and socio-economic variables of different Italian administrative divisions.

Data gathering. This module is further divided in submodules, each one tackling one particular task:

- the `Collect` submodule collects data from different providers;

[7] http://twitter.github.io/interactive/sotu2015/

[8] For details on the geolocalization methods used, see [1]

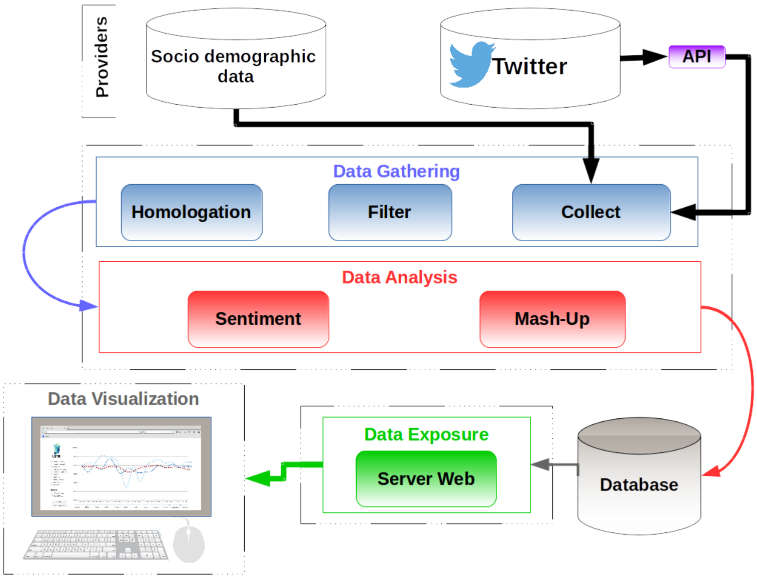


Fig. 1. Framework architecture.

- the `Filter` submodule filters the collected data in order to remove all the possible noisy data, such as duplicate records, empty voices, characters instead of numbers and other formatting errors; as possible correlations have been observed between sentiment and time of the day or day of the week (weekdays or holidays), or between sentiment and geographical areas in a given time frame due to the occurrence of some special event, during this step, we also intend to add a further filter that leaves out all the tweets that bear such temporal or geographical bias[9], as already made in [3], in the creation of the validation corpus.
- the `Homologate` submodule is devoted to the proper organization of collected and filtered data into a unified format. For example, 1420070400, 01/01/2015 and Thu, 01 Jan 2015 00:00:00 GMT indicate the same date, and 058091, [41.53,12.28] and Roma indicate the same city. The Homologate submodule converts dates in YYYY/MM/DD format, and administrative divisions in the ISTAT code[10].

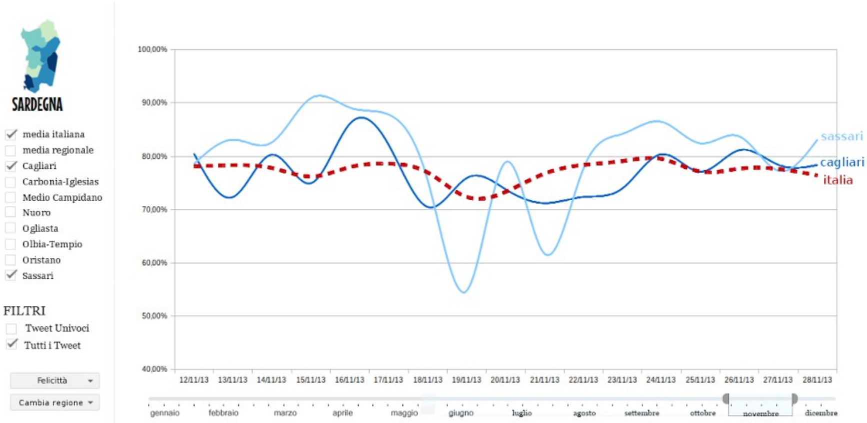
Data Analysis. First, the `Sentiment Analysis` submodule returns for each tweet a mood value (positive, negative, neutral); the SA engine is the one developed in Felicittà, as described in [1]. Then, the `Mash-Up` submodule aggregates Italian geolocated tweets by regions, provinces and municipalities. In this way, data about moods and social indicators can be grouped on the basis of the same

[9] Indeed, conventional expressions such as "*Happy New Year*", "*Merry Christmas*", and others, should not be considered as equally representative of, for example, joy.

[10] `http://www.istat.it/it/archivio/6789`

period and the same administrative level. The aggregate data are finally stored in a database. A correlation analysis across moods and, in turn, each statistic is performed, in order to quantify the strength of the relationship between the variables. As further detailed in 3.2, this is the most recent part of the project, that extends the one implemented in Felicittà.

Data Exposure. A web server exposes elaborated data by REST API. When a client runs a query, the server queries the database and returns the response.



SARDEGNA

☑ media italiana
□ media regionale
☑ Cagliari
□ Carbonia-Iglesias
□ Medio Campidano
□ Nuoro
□ Ogliasta
□ Olbia-Tempio
□ Oristano
☑ Sassari

FILTRI
□ Tweet Univoci
☑ Tutti i Tweet

Felicittà
Cambia regione

Fig. 2. A query result in Felicittà that shows the degree of happiness in relation to an event in particular, that is the flood that hit the northern area of Sardinia in 2013. The graph shows that, based on the analysis of tweets from that area, i.e. the province of Sassari, at that time-frame (November, 19th), a far lower degree of happiness is registered both with respect to other areas in Sardinia (such as the southern province of Cagliari) and the whole country.

Data Visualization. Finally, a web client presents the data obtained as response to the queries. For the time being, the visualization module allows to browse either the sentiment data (as in the example in Figure 2), or the sentiment data combined with demographic data, as shown in Figure 3. The part that shows socio-demographic statistics and correlations is yet to be completed.

3.2 Statistics

As a measure of the mood related to an area in a given period, we consider the percentage of positive tweets. In order to relate moods and numeric social indicators in different administration degrees, in Table 1 we summarized some social indicators that could provide an overview of the social well-being of a given community.

As data collection is not always an easy task and the Open Data is not yet widespread in Italian public administration, we realistically decided to focus

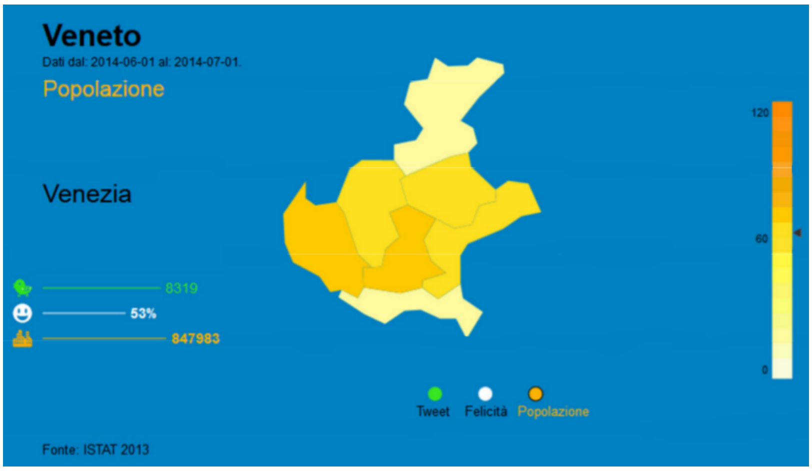


Fig. 3. Demographics and social media data in Felicittà. The provinces of each region (such as Veneto, in the picture) are coloured according to the number of inhabitants. Demographic data are combined with social media data (the number of tweets posted in Veneto in the interval of time) and mood (the percentage of positive tweets).

Description	**Source**	**Period**	**T.U.**
BES Measures	Istat	Y	R
Population by nationality, gender, marital status, and age	Istat	Y	M
Employed and unemployed by gender	Istat	M1	P
Workforce (Number of employees, artisans and so on)	INPS	M6	P
Retirements	INPS	M6	P
Companies registered and ceased by category	C.C.	Y	P
Exports / Imports	Istat	M3	R
Layoff	Istat	M1	P
Real estate market	A.E.	M3	P
Loans and bank deposits	B.I.	M6	P
Public debt of local governments	B.I.	Y	P

Table 1. Selection of Italian official statistics. Sources of data are Istat, Chamber of Commerce (C.C.), Italian Agency of Incomes (Agenzia delle Entrate, A.E.), Italian Welfare Institute (INPS) and Bank of Italy (Banca d'Italia, B.I.). Selected periods are Year (Y), month (M1), quarter (M3) or semester (M6) while Territorial Units (T.U.) are municipality (M), province (P), region (R).

our attention on data that could be easily accessed and retrieved from public administration web-sites. In order to detail different aspects of the society, we resort to different sources. In this way, we consider data from different fields and viewpoints, mainly demographic and economic.

As regards the demographic field, the main aspects considered are nationality, gender, age and marital status, since they are closely related to the perception of social well-being. We are interested, for example, in understanding whether and to what extent nationality may influence the sentiment expressed through social media, or whether married men are happier than singles.

Concerning the economic field, we consider both jobs data (e.g. the unemployement rate) and data about companies (e.g. enterprises demography). Our hypothesis is that people express negative sentiments more likely if they live in an area with significant unemployment rate or with a greater amount of cessations of business.

We also collected data about the real estate market, that we consider a typical indicator of the wealth of a territory. A correlation, in fact, is expected between this aspect and the overall mood detected in social media: the higher the prices (then the wealthier the area considered) and the greater the happiness may be. Similarly, we consider the amount of deposit and loan from the Bank of Italy as a measure of both individuals and public wealth. We selected this set of data as they are representative of different relevant social needs, with different time and granularity. A first integration between social media data and demographic data is shown in Figure 3.

Our current work then namely consists in exploring all the possible correlations between the indicators mentioned above and the output of the SA engine, and in improving the visualization module so as to better highlight such correlations and emerging patterns.

4 Conclusions and future work

In this paper we introduced an ongoing project on a framework for the analysis and assessment of the degree of “happiness” of a given Italian community, taking into account data from official statistics and SA data obtained from social media. We noticed at least two main problems: the representativeness of data and the role of ironic sentences. First, the diffusion of internet and the use of online social networks is not widespread in the same way over all kinds of population. Therefore, for instance, the sentiment of poorest people and elderly can be not represented or largely underrepresented. This is a classical problem of quite every sociological inquiries, mainly solved by representative sampling and qualitative research. A second issue is the presence of irony where the unintended meaning of words can often reverse the polarity of the message. We well know this problem and we state how exists a growing interest in this research subject, as we already investigate the role and the detection of irony and sarcasm[9]. As mentioned above, the work is still in progress and some issues limit the results, but there are also several expected positive impacts of the proposed approach.

First, we focus on the selection of data from official statistics that better correlate with social media data. An hypothesis is that a variation in the data on the labor market and, most of all, the youth employment situation in a given region entail a variation in the mood of the public opinion as expressed in online social media. Detecting the strength of the statistical relation between different variables could help in using social media as a tool for detection of social and economic trends. Another relevant concrete application of the present framework is the inclusion in the platform of Felicittà of the selected statistical data with the output emerging from the correlation analysis.

Acknowledgments. Part of the present contribution has been awarded in the 2014 Istat-Google Contest on "Producing official statistics with Big Data": `http://www.istat.it/it/archivio/144042`

References

1. L. Allisio, V. Mussa, C. Bosco, V. Patti, and G. Ruffo. Felicittà: Visualizing and estimating happiness in italian cities from geotagged tweets. In *Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013)*, pages 95–106, 2013.
2. J. Bollen, H. Mao, and X.J. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011.
3. C. Bosco, L. Allisio, V. Mussa, V. Patti, G. Ruffo, M. Sanguinetti, and E. Sulis. Detecting happiness in Italian Tweets: Towards an evaluation dataset for sentiment analysis in Felicittà. In *Proceedings of the 5th International Workshop on EMOTION, SOCIAL SIGNALS, SENTIMENT & LINKED OPEN DATA, (ESLOD 2014)*, pages 56–63, 2014.
4. A. Cairo. *The Functional Art: An introduction to information graphics and visualization.* New Riders, 2013.
5. E. Diener. Subjective well-being: The science of happiness and a proposal for a national index. *American psychologist*, 55(1):34, 2000.
6. P.S. Dodds, K.D. Harris, C.A. Kloumann, I.M.and Bliss, and C.M. Danforth. Temporal patterns of happiness and information in a global-scale social network: Hedonometrics and twitter. *PLoS ONE*, 2011.
7. R. Easterlin. Does money buy happiness? *The Public Interest*, 1973.
8. M. Forgeard, E. Jayawickreme, M. Kern, and M. Seligman. Doing the right thing: Measuring wellbeing for public policy. *International Journal of Wellbeing*, 2011.
9. D. Hernandez, E. Sulis, V. Patti, G. Ruffo, and C. Bosco. Valento: Sentiment analysis of figurative language tweets with irony and sarcasm. *Proc. Int. Workshop on Semantic Evaluation (SemEval-2015), Co-located with NAACL and *SEM. To appear.*, 2015.
10. J. Sachs J. Helliwell, R. Layard and Emirates Competitiveness Council. *World happiness report 2013.* Sustainable Development Solutions Network, 2013.
11. A. Kramer. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 287–290. ACM, 2010.

12. A. Kramer. The spread of emotion via facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 767–770. ACM, 2012.
13. E.W.Dunn L.B.Aknin, M.Norton. From wealth to well-being? money matters, but less than people think. *The Journal of positive psychology*, 4(6):523–527, 2009.
14. L. Mitchell, M. Frank, K. Harris, P. Dodds, and M. Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 2013.
15. T. Nguyen, B. Dao, D. Phung, S. Venkatesh, and M. Berk. Online social capital: Mood, topical and psycholinguistic analysis. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013.
16. OECD. *OECD Guidelines on Measuring Subjective Well-being.* OECD Publishing, Paris, 2013.
17. D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. Tracking gross community happiness from tweets. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*, 2012.
18. Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.
19. C.o Wang, Z. Xiao, Y. Liu, Y. Xu, A. Zhou, and K. Zhang. Sentiview: Sentiment analysis and visualization for internet popular topics. *Human-Machine Systems, IEEE Transactions on*, 43(6):620–630, 2013.
20. Kosinski M. Stillwell D.J. Wang, N. and J. Rust. Can well-being be measured using facebook status updates? validation of facebook's gross national happiness index. *Social indicators research*, pages 483–491, 2014.
21. N. Yau. *Visualize this: the FlowingData guide to design, visualization, and statistics*. Wiley Publishing, 2011.