IMT
SCHOOL
FOR ADVANCED
STUDIES
LUCCA

#04
2016

ISSN 2279-6894
IMT LUCCA EIC WORKING
PAPER SERIES 04
July 2016

RA **Economics and institutional change**

Tweet-tales: Moods of Socio-Economic Crisis?

Grazia Biorci
Antonella Emina
Michelangelo Puliga
Lisa Sella
Gianna Vivaldo

Research Area
Economics and institutional change

Tweet-tales: Moods of Socio-Economic Crisis?

Grazia Biorci
CNR- IRCrES, Genova

Antonella Emina
CNR- IRCrES, Moncalieri

Michelangelo Puliga
IMT School for Advanced Studies Lucca

Lisa Sella
CNR- IRCrES, Moncalieri

Gianna Vivaldo
IMT School for Advanced Studies Lucca

ISSN 2279-6894
IMT LUCCA EIC WORKING PAPER SERIES #04/2016

Piazza San Ponziano 6, 55100 Lucca

Tweet-tales: Moods of Socio-Economic Crisis?

Biorci Grazia, CNR-Ircres (Genova), grazia.biorci@ircres.cnr.it +39 0102465459
Emina Antonella, CNR-Ircres (Moncalieri), antonella.emina@ircres.cnr.it +39 0116824929
Puliga Michelangelo, IMT (Lucca), michelangelo.puliga@imtlucca.it
Sella Lisa, CNR-Ircres (Moncalieri), lisa.sella@ircres.cnr.it +39 0116824926
Vivaldo Gianna, IMT (Lucca), gianna.vivaldo@imtlucca.it

Key Words
Big data, social media, Twitter, hierarchical clustering, unemployment.

Abstract

The widespread adoption of highly interactive social media like Twitter, Facebook and other platforms allow users to communicate moods and opinions to their social network. Those platforms represent an unprecedented source of information about human habits and socio-economic interactions. Several new studies have started to exploit the potential of these big data as fingerprints of economic and social interactions.

The present analysis aims at exploring the informative power of indicators derived from social media activity, with the aim to trace some preliminary guidelines to investigate the eventual correspondence between social media indices and available labour market indicators at a territorial level. The study is based on a large dataset of about 4 million Italian-language tweets collected from October 2014 to December 2015, filtered by a set of specific keywords related to the labour market. With techniques from machine learning and user's geolocalization, we were able to subset the tweets on specific topics in all Italian provinces. The corpus of tweets is then analyzed with linguistic tools and hierarchical clustering analysis. A comparison with traditional economic indicators suggests a strong need for further cleaning procedures, which are then developed in detail. As data from social networks are easy to obtain, this represents a very first attempt to evaluate their informative power in the Italian context, which is of potentially high importance in economic and social research.

1 Introduction and motivation

In last years, the enhancing development in Big-data science provided tools for collecting and analyzing an unprecedented amount of information about human habits, socio-economic interactions, and collective decision-making. Data from search engines, search query volumes, Internet users, and so on provide useful information for predicting collective behaviour (Goel et al. 2010, Choi and Varian 2012, Moat et al. 2014).

Bentley et al. (2014) specifically explore the use of massive data from highly interactive social network platforms, that allow users an instantaneous and pervasive communication of moods and opinions to their social networks (e.g., Twitter and Facebook), as a new information source to investigate collective

decision-making. In a different perspective, many other studies conceive social media data as genuine fingerprints of socio-economic interactions, such as local economic development and the relation between mobility fluxes and unemployment (Eagle et al. 2010, Llorente et al. 2015).

Focusing on labour market flows, different Big-data sources and approaches are currently under exploration. Internet job search query indices have been extensively exploited to enhance the nowcasting of contemporaneous economic activity in Israel (Suchoy 2009), Italy (D'Amuri 2009), Germany (Askitas and Zimmermann 2009), and the U.S. (Choi and Varian 2009). Antenucci et al. (2014) propose a social media index of job loss for the U.S., derived from counts of job-related expressions in Twitter data. Their real-time index fits an interesting tracking of initial claims for unemployment insurance data, which performs better predictions than both consensus forecast and lagged data, thus showing that Tweets incorporate pieces of information which are not reflected elsewhere.

Arising from their experience, this paper presents a preliminary approach to analyze Italian-language Twitter data, aiming at investigating critical issues in the Italian labour market. In particular, an integrated approach is described, that blends data science, textual/linguistic, and statistical techniques with the aim to trace some preliminary guidelines to investigate eventual correspondence between social media indices and available labour market indicators at a provincial level.

Analyses are performed on a large corpus of about 3 million Italian-language Tweets, dated from October 2014 to December 2015 and filtered by a set of specific keywords that are semantically related to the labour market. By means of machine learning and users' geo-localization techniques, the Tweets are subset in all Italian provinces. Given both the wide semantic richness of Italian language and the not truly satisfactory performance of automatic feature selection algorithms (Antenucci et al. 2013), the corpus is then analyzed by domain knowledge linguistic techniques, with the double aim of inspecting the corpus general contents and moods and of extracting textual signals that potentially correlate with socio-economic indicators of job lack. Since preliminary statistical cluster analysis on that signal fits unclear comparison with traditional economic indicators, several noise cleaning procedures are developed, including users' detection strategies.

The rest of the paper is organized as follows: after a brief corpus description, a linguistic and textual knowledge domain analysis is presented, that explores job- and unemployment-related issues. Then, count variables from unemployment-related tweets and unemployment rates at a provincial level are compared by a hierarchical clustering approach. Finally, the effect of noise is assessed, and advanced corpus cleaning procedures are proposed as a starting point for further analysis.

2 Corpus description

The corpus is a collection of tweets from Twitter recorderd from Oct 2014 to Dec 2015 using the Twitter Stream API (https://dev.twitter.com/streaming/public) filtered using a bag of words (BoW now on) that consists of Italian words like contracts, (un)employment, job, layoff, young, govern, wage, act, workers union, as well as the names of some important political leaders. The BoW considers also several topics of interest to the Italian labour market, such as *articolo* 18 and jobsact, that are respectively an entry on the Worker's Statute about layoff, and a job market reform proposed by Mr. Renzi, the present Italian Prime Minister. Each term of the BoW is extracted by case insensitive procedures, and hashtag versions are considered too (i.e., renzi, #Renzi).

The initial collection contains more than 12 millions tweets with texts in several languages, that are filtered in two ways: a) looking to the "language" field given by Twitter (a machine learning system able to distinguish among several languages in real time as all tweets are marked with the recognized language), b) using a tool for language detection based on a large multilingual machine learning training dataset (langid https://github.com/saffsd/langid.py). This double check allows to reduce the risk to misinterpret the language: assuming that Twitter has a precision of 90-95% and our tool langid has a comparable precision (90% as declared in the guide), this means that with the double check we are now misinterpreting 1 in 100 tweets.

Moreover, we were interested in selecting the geographical features of the tweets focusing on the location that is declared by the user at the time of his/her Twitter subscription. This feature must not be confused with the actual user's geo-location, that is his/her actual position when the smartphone GPS system is on. In fact, for our purposes the GPS position can be strongly misleading, since it captures mostly the travel situations rather than the user's provenance.

The location field is declared by the users in the 20%-30% cases, and it can be parsed to extract a large fraction of reliable locations. For instance, the location Lucca refers to a city in Italy, and it is clear and unequivocal, so we accept it; the location Italy or Tuscany is instead refused, being too generic. The geolocation process is made using a *gazetteer*, i.e. a long list of geographical Italian cities and villages, their coordinates, and their administrative subdivisions. Since our work focuses on the Italian provinces (i.e., subdivisions of regions), each tweet geolocation is associated with the correspondent province. The procedure is refined for common mistakes such as capital letters (i.e., ROME instead of Rome) and punctuation removal. A conservative approach is preferred, i.e. only the most safe users' locations are kept. This choice reduces the corpus to a final dimension of 3.209.715 tweets, including repeated tweets (retweets and reposts). The tweets were generated by a total 250.743 unique users.

3 Knowledge domain: linguistic and textual analysis

The corpus consists, thus, of a portion of lexicon which is heavily connoted to the domain of *work* in all its possible declinations, so *job/work* are our superordinated words. We started our queries testing expressions concerning *job loss* and *unemployment* domains.

For this purpose, we

a) set a new BoW[1], a selected series of linguistic patterns, words and syntagms, which are more likely expected to describe events concerning job loss, firing, unemployment;

b) verified their presence in the selected tweets;

c) analysed semantic value and salience of the co-occurrences in the tweets;

d) drafted a sort of twitter-thesaurus centred on *disoccupazione* (unemployment) and *lavoro* (job/work).

As first survey on social mood concerning unemployment and loss of job, we chose to start our test with the stemmed words *disocc** and *lavor**, following the hypothesis that the semantic values of the co-occurrences around such node words might suggest adequate hints. Within these sub-corpora, we read and listed all the co-occurrences attested in the co-texts. We proceeded in grouping words (nouns, verbs and adjectives) according to their semantic domain and observing their frequency and salience among the sub-corpora. We were able then to shape a sort of twitter-thesaurus centred on mood expressions concerning unemployment and job, so we could highlight the main topics of tweets clustered under conceptual superordinated words (Tognini–Bonelli 2001).

By inquiring the terms in the co-texts and by figuring, quite significantly, their semantic domain, it has been possible to define a sort of descriptive grid of this portion of the corpus. In particular we:

- produced a list of concordances of the node words identified as salient and semantically important for our aims. The length of the concordance string was decided to be 140 characters, which is exactly the length of a standard tweet. We had, so, a complete, and significant, visualization of the co-texts;
- obtained a list of collocations, in which are displayed words, which are more frequent or statistically more probably close to the node word "disocc*". In this way, syntagms and co-occurrences of "disocc*" emerge and connote the salience of each node word;
- tested both collocations and clusters of node words and compared the results with those obtained by a close reading of the concordances.

From our text analysis and linguistic point of view, the problems we had to face may be resumed in two main topics: the software used and the significant presence of noise. As regard as the chosen text-analysis software, *AntConc*, although innovative and usually well performing, it showed some constraints in creating

[1] *perso il lavoro, perdere il lavoro, perdere il posto di lavoro, perderlo, senza lavoro, mancanza d* lavoro, mancanza lavoro, manca lavoro, manca il lavoro, lavoratori a casa, lasc* a casa, rest* a casa, giovani a casa, licenzi*, disocc*.*

and exploiting sub-partitions of the corpus and in exporting data. The presence of noise, conversely, was due to different causes: the automatic/robot re-tweeting, the presence of the emoticons codes, appearing as a series of letters/characters producing non-words, the presence of many symbols and diacritical signs used for emphasis and rhetoric purposes.

4 Statistical clustering: preliminary comparisons with economic indicators

As a first step, the geolocalized tweets of the 110 Italian provinces were analyzed by hierarchical clustering[2] methodologies, in order to extract a common behaviour in terms of their Twitter activity. At this preliminary stage, the weekly provincial Twitter activity was computed by counting the raw occurrences of the stemmed word *disocc** inside each unique Twitter per province per week[3]. As a first result, we observed the lack of convergence of clustering algorithms, even at a quite high iteration rate, suggesting a serious instability in detecting the suitable number of possible communities present in the dataset. Just few clusters were detected. In particular, the most populated provinces Rome and Milan emerged as isolated and dominating components, pointing that weekly count data need to be normalized for the province Twitter population. Since these data are unavailable, the total amount of Twitter users for each province was estimated by the province population itself, under the preliminary hypothesis that the users are equally distributed among Italian provinces, i.e. it exists a direct proportionality between the population of a given province and its Twitter users. As the number of unique users changes in time for seasonal effects, we took the monthly average number of unique users per province, and we plotted it against the 2015 province population, revealing the presence of a strong linear relationship. Thus, two further weakly variables were computed: the number of unique users per estimated Twitter users (*users/Twitter_users*), and the amount of total weekly counts of *disocc** per estimated Twitter users (*counts/Twitter_users*).

A first visual Cattell (1966) *Scree test* on both *users/Twitter_users* and *counts/Twitter_users* variables suggested that the number of clusters was not trivial to assess, since no evident structural break clearly emerged between major and minor (trivial) factors. Hierarchical clustering was then performed.

According to preliminary results, *counts/Twitter_users* records were not stable to changes in the clustering method. Single and complete algorithms were unable

[2] Hierarchical clustering was performed by Ward (1963) *minimum variance*, *complete* and *single linkage* methods (Murtagh and Legendre 2014). The distance matrix was computed by *Euclidean* metric. Clusters uncertainty was assessed at the 95% c.l., following the approach of Shimodaira (2004) and Suzuki and Shimodaira (2006).

[3] Raw counts were standardized to zero mean and unit variance in order to ensure their reciprocal comparability.

to isolate significant components, while *Ward's* methodology detected four clusters at the 95% confidence level (Fig. 1, central panel).

The analysis was slightly more robust for *users/Twitter_users* records. A small cluster including Avellino and Parma provinces turned out to be more stable to algorithm changes, while just *Ward's minimum variance* method was able to detect other three clusters at the 95% c.l. (Fig. 1, right panel). Some correspondences clearly emerge between the two normalized variables.

The visual mapping of unemployment rates at provincial level in December 2014 (Fig. 1, left panel) showed the well-known partition in low, medium and high unemployment areas, that does not have any clear counterpart in the two proposed social media indices. In fact, the *counts/Twitter_users* clustering shows a sort of bipartition between low- (blue) and high-unemployment (red) areas, but some low-unemployment Northern regions fall in the red cluster. On the contrary, the expected polarization is not observed in the second indicator at all (*users/Twitter_users*).

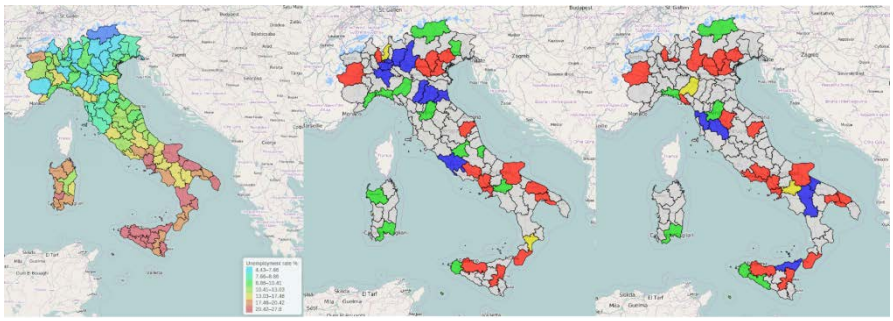

Fig. 1 Unemployment statistics vs. Twitter unemployment indicators. *Left panel*: official unemployment rates per province, Dec. 2014. Source: our elaboration on Istat data. *Central panel*: hierarchical clustering on counts/Twitter_users. *Right panel*: hierarchical clustering on users/Twitter_users. Source: our elaborations on Twitter data. Algorithm: Ward's (1963) minimum variance.

5 Discussion and preliminary conclusions: advanced filtering issues

This study focuses on statistical analyses of unemployment-related tweets in Italy, with the aim of exploring the informative power of indicators computed from social media activity in describing well known socio-economic phenomena. Due to a large noise, preliminary clustering results were stable.

In fact, the initial cleaning procedure generating the unemployment-connoted subcorpus was scarcely able to describe geographical characteristics according to the well known patterns of unemployment in Italy. Instead, data tended to be mixed, and hierarchical clustering results were not robust across algorithms. Hence, preliminary analyses suggested the need for further advanced filtering.

To improve the original filtering procedure, we investigated several tweet features, especially the interarrival time of two consecutive tweets for each user, that undercover the eventual presence of automatic retweeting systems (bots). The hypothesis is that no human user can tweet with a rate of one tweet per second or lower. As a matter of fact, the Twitter platform is particularly suited for automatic systems that can repost or retweet contents to increase the chances of spreading a message to a large audience. This characteristic can affect the descriptive/predictive power of signals extracted from Twitter data, when the analysis aims at measuring socio-economic phenomena from the echo they have in people conversations. In fact, in the corpus we identified job posting and press agencies tweets, as well as tweet reposts (retweets or reply) from automatic systems. Fig. 2 (right panel) shows the distribution of interarrival times for each user. The curves display different behaviour: while most recurrent blue peaks refer to users that tweet on average every 15 minutes or 1 hour, other users are tweeting every few seconds and in large numbers. These users are very likely to be bots, adding non negligible noise to the corpus.

Another class of users eventually generating noise are the professional users, such as press agencies, that usually issue posts at fixed times (e.g., 1 hour). This behaviour is represented by the largest red area in the inter-arrival map plot (Fig.2, left panel). We can make the dataset cleaner removing this kind of users and their tweets, too.

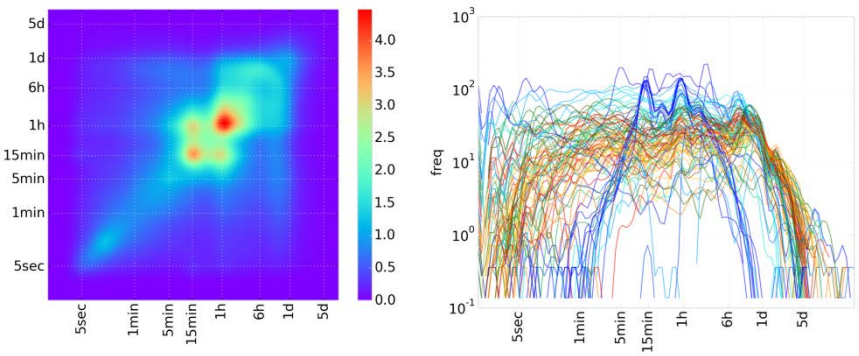

Fig. 2 Interarrival time plots. *Left panel*: tweets interarrival time map across users. Red-orange regions represent in ln scale higher tweet activity for the corresponding interarrival time couple. *Right panel*: tweet frequencies versus interarrival times for each user. Colours are proportional to individual average frequency. Source: our elaboration on Twitter data.

Finally, as further filtering improvement, we could isolate and look at the peaks in the volume of tweets in time. When a new law or act is announced in the media, users react with an higher rate of comments. This phenomenon is known as *agenda setting* and can be used to evaluate the diffusion of the same news across different areas in Italy. We can measure the response to each tweet peak in terms

of intensity (number of tweets) and persistence (decaying rate of each peak). A check of the geographical distribution of those parameters can be interpreted as a measure of interest of each topic and in each part of the country. The signal of the peaks is more clean than the baseline signal, that is likely to be more noisy and less focused on the labour market topics. However, the peak signal is strongly dependent on the specific issue discussed by the media and this can be misleading in turn. We can easily identify the topic of the peak just looking at the media tweets that are responsible for the initial peak surge.

6 References

Antenucci, D., Cafarella, M. J., Levenstein, M., Ré, C., Shapiro, M. D. (2013). Ringtail: Feature Selection For Easier Nowcasting. *WebDB*, 49-54. http://www-cs.stanford.edu/people/chrismre/papers/webdb_ringtail.pdf. Accessed 29 April 2016.

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., Shapiro, M. D. (2014). Using social media to measure labor market flows. *NBER Working Paper Series*, w20010, National Bureau of Economic Research.

Askitas, N., Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *German Council for Social and Economic Data (RatSWD) Research Notes*, 41.

Bentley, R. A., O'Brien, M. J., Brock, W. A. (2014). Mapping collective behavior in the big-data era. *Behavioral and Brain Sciences*, 37(1), 63-76.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276.

Choi, H., Varian, H. (2009). Predicting initial claims for unemployment benefits. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.549.7927&rep=rep1&type=pdf. Accessed 29 April 2016.

Choi, H., Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88, 2–9.

D'Amuri, F. (2009). Predicting unemployment in short samples with internet job search query data. *MPRA Paper*, 18403, University Library of Munich, Germany.

Eagle, N., Macy, M., Claxton, R. (2010). Network diversity and economic development. *Science*, 328(5981), 1029-1031.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National academy of sciences*, 107(41), 17486-17490.

Llorente, A., Garcia-Herranz, M., Cebrian, M., Moro, E. (2015). Social media fingerprints of unemployment. *PloS one*, *10*(5), e0128692.

Lloyd, S. (1982). Least squares quantization in pcm. Information Theory. *IEEE Transactions*, 28, 129–137.

Moat, H. S., Preis, T., Olivola, C. Y., Liu, C., Chater, N. (2014). Using big data to predict collective behavior in the real world. *Behavioral and Brain Sciences*, 37(1), 92-93.

Murtagh, F., and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, 31(3), 274–295.

Shimodaira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling, *Annals of Statistics*, 32, 2616–2641.

Suchoy, T. (2009). *Query indices and a 2008 downturn: Israeli data*. Bank of Israel. Research Department.

Suzuki, R. and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering, *Bioinformatics*, 22 (12): 1540–1542.

Tognini–Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins Publishing.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*, 58, 236–244

CAPTIONS

Fig. 1 Unemployment statistics vs. Twitter unemployment indicators. *Left panel*: official unemployment rates per province, Dec. 2014. Source: our elaboration on Istat data. *Central panel*: hierarchical clustering on counts/Twitter_users. *Right panel*: hierarchical clustering on users/Twitter_users. Source: our elaborations on Twitter data. Algorithm: Ward's (1963) minimum variance.

Fig. 2 Interarrival time plots. *Left panel*: tweets interarrival time map across users. Red-orange regions represent in ln scale higher tweet activity for the corresponding interarrival time couple. *Right panel*: tweet frequencies versus interarrival times for each user. Colours are proportional to individual average frequency. Source: our elaboration on Twitter data.

Piazza San ponziano 6, 5100 Lucca, Italy
www.imtlucca.it