New Experimental Statistics at Istat: the Social Mood on Economy Index

**Keywords:** Experimental Statistics, Big Data, Social Media, Sentiment Analysis, High‑Frequency Indices

# Introduction

Nowadays millions of people all over the world use social media platforms to keep up with the news, to express their feelings and ideas, as well as to share or debate opinions on virtually every conceivable topic. This justifies the interest of National Statistical Institutes (NSI) towards social media as a means for “measuring” the public mood [1].

In recent years, the Italian National Institute of Statistics (Istat) has been investigating whether social media messages may be successfully exploited to develop domain‑specific sentiment indices, namely statistical instruments meant to assess the Italian mood about specific topics or aspects of life, like the economic situation, the European Union, the migrants’ phenomenon, the terrorist threat, and so on.

To this end, Istat researchers have developed procedures to collect and process only social media messages containing at least one keyword belonging to a specific filter, namely a definite set of relevant Italian words. Domain-specific filters have been designed by subject-matter experts with the aim of filtering out since the beginning messages that would very likely turn out to be off-topic for the intended statistical production goal.

Istat has recently released a new experimental statistic [2], based on Twitter data: the *Social Mood on Economy Index*. The index provides daily measures of the Italian sentiment on the state of the economy. These measures are derived from samples of public tweets in Italian, which are captured in real time.

Similar initiatives have been put in place by other NSIs in recent times, notably Statistics Netherlands’ attempts to “mimic” the time evolution of the Dutch Consumer Confidence Index by means of a sentiment index based on social media [3], and to derive daily measures of social tension from Twitter [4].

The rest of this paper provides an overview of the production pipeline of Istat’s *Social Mood on Economy Index*.

# Methods

Twitter’s Streaming API is used to collect samples of public tweets matching a filter made up of 60 relevant keywords (some are actual words, some are phrases). A subset of these keywords has been borrowed from the questionnaire items of the Italian Consumer Confidence Survey, a monthly sample survey that collects data in the first two weeks of each month and disseminates results by the end of the month. However, the phenomenon tracked by the Social Mood on Economy Index is much broader in scope, fuzzier, and only partially overlaps consumer confidence. Still, the index can detect and promptly point out events that influence consumer confidence but occur after the interview period of the consumer confidence survey: the Central Italy earthquake of 24th August 2016 is a striking example (see Section 3).

The sampling algorithm is almost a black box, as it is entirely controlled by Twitter’s Streaming API. Very little is known about it. The API is allowed to return *at most* a 1% sample of all the tweets produced on Twitter at a given time. However, if a filter is specified, the API returns *all* the tweets matching the request up to the “1% of all tweets” limit. From the above rules it follows that our data collection procedure actually captures *at least* 1% of all the relevant tweets (and very likely much more than that, given the specificity of the keywords employed).

To compute daily index values, all the tweets collected in a single day (about 47,000, on average) are processed as a single block, see Figure 1. Messages are first cleaned and normalized, then undergo a sentiment analysis procedure. For this purpose, an unsupervised, lexicon-based approach has been adopted (we had to discard supervised, Machine Learning methods because we have been unable to find large, high-quality collections of human-labeled tweets in Italian that could be used as training sets).



Figure 1: A schematic representation of the processing pipeline of the Social Mood on Economy Index

For each tweet, positive and negative sentiment scores are calculated. To this end, message words are matched against entries of an Italian sentiment lexicon, namely a vocabulary whose lemmas are associated to pre-computed sentiment scores. Atomic scores of matched words are then averaged to yield tweet-level scores. Subsequently, tweets are clustered according to their sentiment scores into three mutually exclusive classes: Positive, Negative and Neutral tweets. Lastly, the daily index value is derived as an appropriate central tendency measure of the score distribution of the tweets belonging to the Positive and Negative classes. As a post processing step, the daily index is linearly transformed in such a way that its long-run mean, referred to the baseline period 10th February 2016 – 30th September 2018, is zero.

Special care has been devoted to make the index robust against possible contaminations by off-topic tweets that might pass the filter. To this end, a surveillance system has been put in place, which continuously searches for anomalous values in the daily index time series by means of two independent and complementary outlier detection routines. Daily values detected as potential outliers cause the system to automatically generate a set of dedicated diagnostic reports. These are then sent to human reviewers in charge of deciding whether the detected values are actually proper data points or truly anomalous. The latter case typically arises when an off-topic tweet that happened to pass the filter becomes “viral” on Twitter. Being re-twitted and quoted thousands of times in a day, viral tweets may have an unduly impact on the daily index and introduce bias. As a consequence, all the daily index values classified as truly anomalous are eventually imputed via nearest-neighbor interpolation.

# Results

Data collection for the Social Mood on Economy Index started in February 2016 and has been active since then almost without interruptions. The first publication of the index as Experimental Statistic involved daily data covering the period 10th February 2016 – 30th September 2018. The plot in Figure 2 shows the time behavior of the index (green line), along with the corresponding 15-days (blue line) and 30-days (red line) moving averages. The higher the value of the index, the better is the sentiment of the day.



Figure 2: Time series of the Social Mood on Economy Index from 10th February 2016 to 30th September 2018

As evident from the plot, the daily time series of the index is very volatile. This comes as no surprise, since active users of Twitter typically react very quickly to the news, whose life-cycle is of course pretty short. However, the main peaks and main valleys of the index are clearly associated with significant events, which we have been able to identify by analyzing the most frequent terms in the texts of the daily corpus. These points are highlighted in Figure 2 with a small square.

Topics characterizing notable valleys in the daily index can be classified into two broad groups: ‘disasters and terrorism’ and ‘welfare and economy’. Examples of the first group are the Central Italy earthquake (24th August 2016), the Livorno flood (10th September 2017), the Andria-Corato train collision (12th July 2016), the truck attack in Nice (14th July 2016) and the Christmas market attack in Berlin (19th December 2016). The second group involves debates and worries about poverty (14th July 2016), women’s retirement regulation (9th August and 28th September 2017), youth unemployment (31st August 2016 and 3rd July 2017) and general unemployment (2nd May 2017). As for the peaks of the index, they are either linked to holydays or, again, to welfare and economy. Remarkable examples include school teachers’ expectations about the removal of the triennial mobility constraint introduced by law 107/2015 (8th and 10th November 2016) and the debate triggered by Pope Francis’ claim that the goal of universal employment should be considered superior to the one of universal income (27th May 2017).

Despite the daily time series in Figure 2 may seem second-order stationary at first glance, with the green line frequently crossing its long-run mean and evolving with nearly constant variance, a formal KPSS test for level-stationarity is rejected at significance level α = 0.05. However, the alternative hypothesis of the KPSS test, that is the presence of a unit-root, is also rejected at α = 0.05 by dedicated tests, like the Phillips-Perron test and the Augmented Dickey-Fuller test. With regard to autocorrelations, both the Box-Pierce and the Ljung-Box tests exhibit extremely small p-values, thus indicating strong evidence against the white noise hypothesis. Inspection of the pacf plot for the daily series shows that partial autocorrelations become negligible after the second lag.

# Conclusions

The *Social Mood on Economy* index is still under development. We are currently exploring more sophisticated natural language processing techniques and alternative approaches to sentiment analysis. Further work will be devoted to the study of the time evolution of the index. For instance, it will be investigated whether the index correlates with, or even can help to predict, other economic indices that are officially released by Istat at low-frequency (i.e. on a monthly, quarterly or annual basis). The hope is that the new index could either improve the performance of Istat’s forecasting models, or enrich existing statistical products (e.g. the Equitable and sustainable well-being - [BES](http://www.misuredelbenessere.it/index.php?id=51)), or even be disseminated as a new official statistic in its own right.

# References

1. ESSnet Big Data, WP7. URL:

<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page>

1. The publication of the Social Mood on Economy Index is scheduled for 23rd October 2018. URL:

<http://www.istat.it/en/experimental-statistics/experiments-on-big-data>

1. van den Brakel, J., Söhler, E., Daas, P. and Buelens, B. (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 43, No. 2.
2. <https://www.cbs.nl/en-gb/our-services/innovation/project/social-tension-indicator-based-on-social-media>