24th OCTOBER 2023

SOCIAL MOOD ON ECONOMY INDEX

A daily measure of the Italian sentiment on the economy based on Twitter data

Nowadays millions of Italians use social media platforms to be updated on the news, to express their feelings and ideas, as well as to share opinions or to virtually debate on every conceivable topic. This justifies Istat's interest towards social media sources and sentiment analysis techniques as means for "measuring" the Italian public mood [1].

In recent years, Istat has been investigating whether social media messages may be successfully exploited to develop *domain-specific* sentiment indices, i.e. statistical instruments meant to assess the Italian mood about specific topics or aspects of life, such as the economic situation.

To this aim, Istat researchers have developed procedures to collect and process social media messages containing keywords belonging to a specific *filter*, i.e. a pre-defined set of relevant Italian words. Domain-specific filters have been designed by subject-matter experts with the aim of filtering out off-topic messages for the intended statistical estimation goal.

Istat releases an experimental index, based on Twitter data: the Social Mood on Economy Index. This statistical instrument has been designed to enable high-frequency (i.e. daily) measures of the Italian sentiment on the state of the economy. These measures are obtained from real time samples of public Italian tweets. For the time being, the daily time series of the index is available since February 10, 2016.

During 2021 the time series exceeded 5 years, and due to the overwhelming enhancements of sentiment analysis methods, it was considered appropriate to carry out a methodological revision, below explained, aimed at increasing the quality and informative power of the index.

The time series has been recomputed within the new method

Data processing pipeline and adopted methods

Data processing procedure collects samples of public tweets matching a filter consisting of 60 relevant keywords (actual words or phrases). These keywords has been borrowed from the Italian consumer confidence survey questionnaire, 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 and fuzzier than consumer confidence, whose official measure relies on a standard methodology that

[1] This work has been realized as a part of the PSN IST-02589 project "Uso a fini statistici dei Big Data", defined within the National Statistical Program 2017-2019.

is harmonized at European level, and has a long and relevant tradition in short-term analysis and forecasting.

Starting from the fourth quarter of 2021 release, a second-level filter has been added, to exclude some first-level filtered tweets. More in detail, the second-level filter contains 115 keywords (words or lemmas) refining the first-level filter: for instance the word “well-being” has been declined into more specific composite expressions. No new keywords have been introduced and the second filter is a strict subset of the previous of 60 words, thus meaning that a sub-sample of the tweets selected by the previous filter of 60 words, has been drawn. This refinement allowed to identify tweets with specifically economic content with greater probability, even if the volume of selected and analysed tweets decreases significantly.

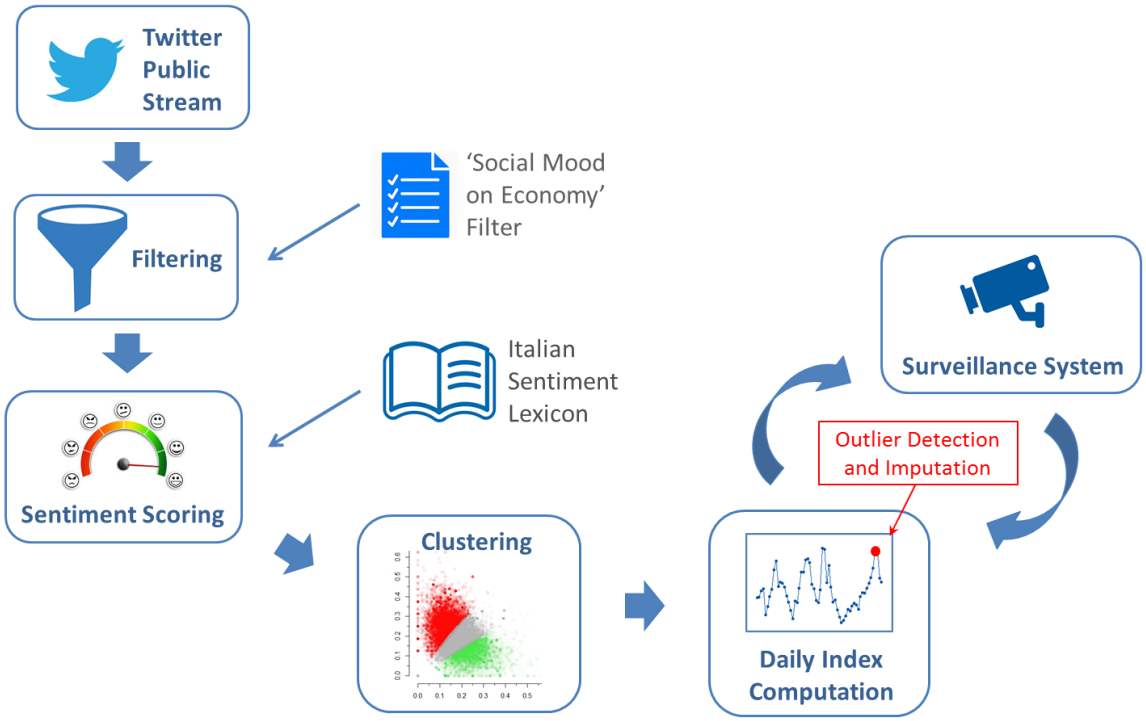

Figure 1: A schematic representation of the processing pipeline of the Social Mood on Economy Index

The procedure for producing the index processes all the tweets collected in a single day (about 26,000, on average currently) as a single block, see Figure 1. First, messages are cleaned and normalized, then undergo a sentiment analysis procedure. The adopted sentiment analysis method is lexicon based, i.e. it is based on an Italian vocabulary whose lemmas are associated to pre-computed sentiment scores. In occasion of The fourth quarter 2021 release introduces breaking changes in the index, i.e. the Italian lexicon has been replaced by a more recent one, which takes into account some neologism linked to the Covid-19 pandemic and which allows to more than triple the previous linguistic coverage.

The texts of all tweets are compared with the lexicon: on the basis of the scores of the matched words, the tweet positive and negative sentiment scores are obtained as weighted averages. 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 all the tweets.

In the previous procedure, since the vocabulary did not provide any neutral sentiment score, clustering was a crucial step for the classification of neutral tweets and the latter were excluded from the index calculation. The entries of the current vocabulary have positive, negative and neutral sentiment scores, therefore these operations are not necessary. Furthermore, having reduced the volumes of filtered tweets, these operations would have further reduced the field of observation. The index calculation formula has remained unchanged. The last step consists in removing the mean of the daily index, in such a way that its long-run mean, referred to the time window February 10, 2016 - September 30, 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 aim, a surveillance system has been developed, which periodically searches for anomalous values in the daily 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 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. Therefore, all the daily index values classified as truly anomalous are eventually imputed via nearest-neighbor interpolation.

Representativeness

Likewise, most Big Data sources, Twitter's data generating mechanism does not fall under the statistician's control and is unknown. This is evidently at odds with statistical surveys, whose data are designed by national statistical institutes, but also with administrative data sources, whose data generating mechanism is usually known to official statisticians. Therefore, there is no rigorous way to guarantee the general validity of statistical information derived from Twitter data. Italian Twitter users[2] cannot be considered a representative sample of the Italian population. For this reason, Istat cannot ensure the accuracy of the Social Mood on Economy Index when it is referred to the overall Italian population.

Privacy

The Social Mood on Economy Index is solely based on samples of public[3] tweets. No selection of Twitter's users is made during sampling and user activity is never tracked. The index uses only unlinked anonymized data. Tweets are never linked to Twitter's users during processing: Istat

[2] It has been estimated that, in 2021, there have been more than 11million Italian users active on Twitter on a monthly basis (source: Agcom).

[3] *Public* tweets are immediately visible to anyone, whether or not they have a Twitter account (this is the default setting of the Twitter platform). On the contrary, *protected* tweets are visible only to followers of the author of the tweet.

does not know who the authors of the messages are. Only the textual content of collected tweets is processed. Daily index values result from the aggregation of numeric scores attached to tens of thousands of messages. This is an irreversible process: no tweet can ever be reconstructed by analyzing the index.

Characteristics of the time series

In this section, in order to highlight properties useful for dynamic analysis and to identify non-observable components, the results of the treatments performed on the raw daily index are presented. The daily frequency of the Social Mood on Economy Index implied that for the identification of the seasonal component it was not possible to use the standard Istat methodology, which is the model-based procedure implemented in TRAMO-SEATS. The daily time series present, indeed, different methodological problems compared to the monthly or quarterly series due to the presence of multiple seasonality (midweek, weekly, monthly and/or annual). However, there are considerable advantages in analyzing daily time series with respect to their monthly aggregation, including timeliness, greater availability of information and better identification of the effects of working days and other daily events that influence the dynamic of the series.

Among the recent approaches adopted for the seasonal adjustment of high frequency series, the adopted methodology, as described in De Livera, A. M., Hyndman, R. J., Snyder, R. D. (2011), applies modified exponential smoothing models as compared to the standard models introduced by Holt and Winters. The methodology is based on the hypothesis that each historical series can be represented as a combination of different components: a component that describes the level of the series, one for the growth rate (trend), one that captures seasonal movements and, finally, an irregular one (noise). For the identification of these components, the model is then represented in a state space form whose parameters are estimated by the `tbats` function of the *forecast* package in R.

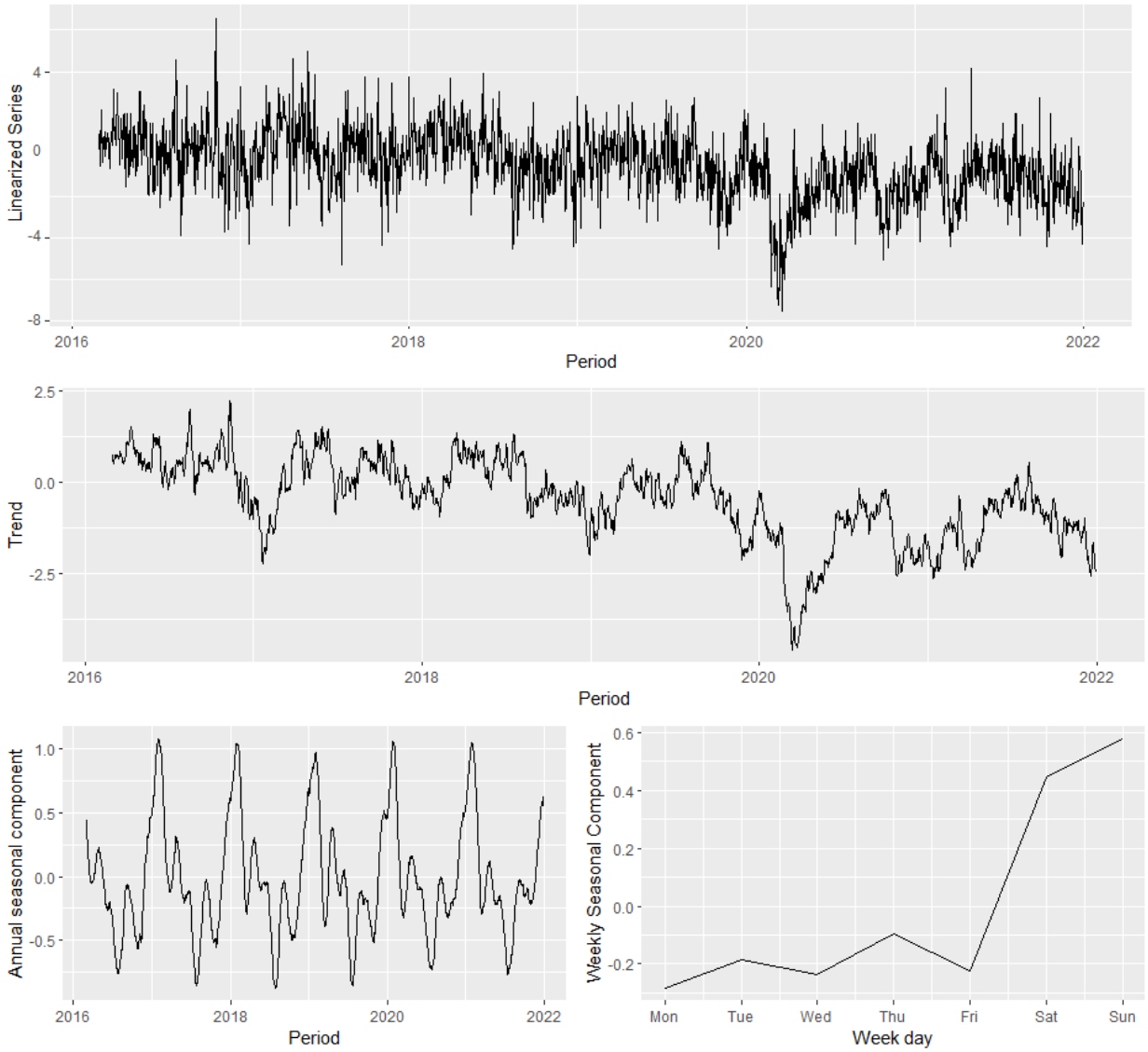

Figure 2: Decomposition of the daily time series of the Social Mood on Economy Index

In particular, the seasonal adjustment of the daily time series of the Social Mood on Economy Index was implemented in two phases: identification and estimation of deterministic effects through the introduction of appropriate dummy variables and identification of the components of the linearized series. The chosen reference period spans from March 1, 2016 (incomplete data collected during February 2016 have been discarded). The time window confirms the decision to use a seasonal adjustment procedure for daily data.

The results of the seasonal adjustment are illustrated in Figure 2, which shows the linearized series of the index along with its trend and the two seasonal components which have been identified by the seasonal adjustment procedure.

Main observed revisions

Though the production procedures haven't been altered, the introduced changes, such as the refinement of the filter and the change of lexicon, have a strong impact on the dynamic of the raw index.

On one side, the field of tweets' observation's change, produced a change in the annotation of the significant peaks and valleys of the daily index, enhancing their economic interpretability. On the other side, it must be stressed that some phenomena, such as those related to earthquakes or laws like option-woman, keep remaining observable, despite the introduced revisions in both filter and lexicon, therefore confirming the robustness of the procedure.

The lexicon change structurally modifies the raw index, by significantly reducing the volatility of the time-series. Finally, both the dynamics of the raw series and the trend show pronounced differences and the most manifest effects are recorded since the beginning of March 2020, where the current index shows a significant breakdown not previously present.

The Social Mood on Economy Index has been produced and analyzed by:

- M. Bruno, E. Catanese, R. Iannaccone, A. Righi, M. Scannapieco, P. Testa, L. Valentino, D. Zardetto, D. Zurlo.