The Italian Social Mood on Economy Index and the Covid-19 Crisis Implications

Righi A., Iannaccone R., Zardetto D., Zurlo D. (ISTAT)

1. Introduction

In recent years, the Italian National Institute of statistics has investigated whether social media messages may be successfully exploited to assess the Italian mood on the country's economic situation. In October 2018, this effort led to the first release of the Social Mood on Economy Index (SME), an experimental high-frequency sentiment index based on Twitter data. Although available only since February 2016, this new index, derived from samples of public tweets in Italian captured in real time, has already gained a good spread among economic analysts for short term analysis or forecasting models.

This paper is aimed at presenting this new statistical tool for economic analysis studying the relationships of this new indicator with some daily and monthly macroeconomic indicators coming from traditional and nontraditional sources. This analysis, performed in the period before and during the Covid-19 pandemic, allows better understanding the actual contribution in terms of information coming from the SME index. We used several nontraditional sources to produce time series to relate to the SME, as the daily number of Covid-19 deaths and new positive cases reported by the Civil Protection Department or macroeconomic indicators coming from Target2 and BI-COMP series (on POS and ATM transactions), but also the Bank of Italy electronic card transaction and e-commerce transaction monthly series.

This work is organized as follows. Section 2 reviews some previous research experiences showing the usage of Twitter information to perform sentiment analysis and the research questions we pose. Section 3 describes Twitter data and the methods used for the construction of the SME index and other traditional and nontraditional data used in our analysis. Section 4 shows the main results of the work: on the one hand, a test of the robustness of the SMEI during the Covid-19 pandemic, on the other hand, the analysis of the relationships between the SMEI and other daily and monthly economic series observed during the Covid-19 pandemic. A discussion on the opportunities and limits of the approach concludes the paper.

2. Background and research questions

Sentiment analysis is an increasing area of research and application, due to the enormous amount of unstructured data currently available. Many studies are published presenting main algorithms to be used for text mining and sentiment analysis (Gandomi and Haider 2015; Feldman and Sanger 2007). Further methodological developments, focused on sentiment analysis on social media short texts performing the analysis in different ways. Firstly, there are methods based on unsupervised learning techniques (Pak and Paroubek 2010), lexicon-based methods (Taboada et al. 2011), and combinations of two previous approaches (Kolchyna et al. 2015). Secondly, other methods are proposed based on the Twitter specificities as is the case of polarization, controversy and topic tracking in time, developed through network measures and clustering techniques (Garimella et al. 2016) or the methods based on the use of the hashtags classification developed through probabilistic models (Coletto et al. 2016b).

Many studies making use of lexicon-based methods refer to the English language (Miller 1995, Strapparava and Valitutti 2005; Esuli and Sebastiani 2006), and there are only a few studies adopting the method for the Italian language due to a minor development of the lexicon in this language. However, Basile and Nissim (2013) developed the Sentiment Italian Lexicon (Sentix), a vocabulary whose lemmas are associated with precomputed sentiment scores, which is the result of the alignment of several sources. Istat followed the lexicon-base method and used the Sentix lexicon in developing the experimental high-frequency sentiment index on the overall economic situation from Twitter data (Fabbri et al., 2018 and Zardetto, 2018).

Research questions

The research questions we pose in this study refer, primarily, to the robustness of the SMEI during the pandemic. How did the SME index react during the pandemic, was it sufficiently inclusive of the messages on the pandemic and the emergency? Secondly, we check which of the daily or monthly economic indicators have had a similar trend to the SMEI during the pandemic using both traditional macroeconomic indicators and nontraditional sources.

3. Data and methods

3.1 The construction of the SME index

Twitter's Streaming API is used to collect samples of public tweets matching a filter made up of 60 keywords relevant for the study of the general and personal economic dimension. All the filtered tweets reported in a single day are processed as a single block to compute daily index values. Messages are cleaned and normalized and then undergo a sentiment analysis procedure, which calculates positive and negative sentiment scores for each tweet.

Using an Italian sentiment lexicon, we calculated the scores of matched words, which we averaged to yield tweet-level scores. Tweets are then clustered according to their sentiment scores into three mutually exclusive classes (Positive, Negative and Neutral). 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 and is linearly transformed[1].

Special care has been devoted to making the index robust against possible contaminations by off-topic tweets that might pass the filter. An automatic outlier detection procedure is set to discriminate truly anomalous data from proper data within the daily time series. Daily index values classified as truly anomalous are eventually imputed via a method of multivariate interpolation (nearest-neighbor interpolation).

However, our estimates can possibly be affected by bias due to different causes: 1) the Italian Twitter users cannot be considered a representative sample of the Italian population, due to different Twitter penetration rates among various sub-population (e.g., young people is overrepresented among Twitter users); 2) the free download from Twitter API we used did not give full access to all the tweets generated by the nine-million Italian active Twitter users, but rather to a subset (a non-probability sample) of those tweets.

3.2 The treatment of the SMEI series

The identification of the seasonal component for daily time series presents the difficulty of distinguishing between the trend component and the annual seasonality, but also the considerable advantage of including 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 seasonal adjustment approaches adopted

[1] The transformation makes the SME long-run mean, referred to the period 10 February 2016 – 30 September 2018, equal to zero.

for high-frequency series, a methodology that applies modified exponential smoothing models compared to the standard models has been selected (De Livera, Hyndman, Snyder 2011). This methodology considers each series as a combination of different components (a first one describing the level of the series, a second one for the growth rate/trend, a third one capturing the seasonal movements and an irregular one). The model for the identification of these components is then represented in a state-space form whose parameters are estimated by the *tbats* function of R package *forecast* (Hyndman and Khandakar 2007).

The seasonal adjustment procedure of the SME Index series is carried out in two phases: the first is the identification and estimation of deterministic effects through the introduction of appropriate dummy variables, and the second one for the identification of the components of the linearized series. The chosen reference period spans from 1 March 2016 to 30 September 2020. The dummy variables are introduced to take into account some effects of a holiday as Christmas and weekend and to correct for the presence of outlier.

3.3 Other traditional and nontraditional data used

We used some traditional macroeconomic indicators and even nontraditional series to study the evolution of their relationships with the SME index. Among the other traditional macroeconomic monthly indicators, we considered the Consumer confidence. Even if SME index and the Consumer climate are both sentiment indices on the economic situation, 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 is harmonized at European level and has a long and relevant tradition in short-term analysis and forecasting.

Furthermore, a scientific collaboration with Bank of Italy gave us access to daily and monthly time series on card transaction (from the issuer side) with granular sectoral information. The source is BI-COMP, a multilateral clearing system created to guarantee the settlement in central bank money of euro-denominated retail interbank payments. BI-COMP handles the transactions and the bilateral balances between banks entered by the clearing systems and operates a multilateral clearing in six daily compensation cycles. At the end of each cycle, BI-COMP calculates a multilateral credit or debit balance for each participant that is sent for settlement in central bank money on the accounts held by intermediaries in TARGET2 (Trans-European Automated Real-Time Gross Settlement Express Transfer System). Newly available series refer to the shares of the daily transaction by merchant category code groups (e.g., clothing, hotels and restaurants, home, work, retail, services, telecom, web, travels and transport) in terms of indices (2014=100), and e-commerce monthly transaction (from the acquirer side, namely, the expenditure of Italians on Italian firms) in terms of indices (2014=100).

3.3.1 New sources and indicators for Covid pandemic

We further exploited the information coming from the SME index to produce a new daily Twitter series, that is, the share of tweets containing the terms “Coronavirus” or “Covid” in the text (among those used in the measure of the SME) out of the total tweets used to measure the SME index. This series, available for the period from 1 January 2020 to 15 October 2020, has been normalized by mapping to 100 its peak value in the period (31%, registered on 24 February 2020).

Using Google Trends as a source, we calculated another daily series for the same period computing the weighted average of the query shares referring to the keywords “Coronavirus” and “Covid” for the Italian territory, using as weights the relative "popularity levels" of the two keywords in the period supplied by Google Trends (18 for "Coronavirus" and 5 for "Covid"). We normalized also this Google Trends daily series by mapping to 100 its peak value in the period (observed on 23 February 2020).

We used also the daily series on the number of Covid-19 newly infected people and deaths coming from the "health risk associated with Coronavirus infection database" of the Civil Protection Department of the Presidency of the Council of Ministers.

4. Main results

From the beginning of March 2020, SMEI showed a strong increase in the volume of tweets, which in two days more than doubled (from around 67 thousand tweets on 3 March 2020 to around 115 thousand ones on 5 March 2020) to reach 144 thousand on 28 March; then, from the end of May, the tweets returned to a volume comparable to the one before the Covid-19 crisis.

As for the SMEI levels, through the *tbats* procedure implemented in *R* software, we removed multiple seasonal components identified through a careful inspection of the spectral density of the linearized series and we extracted the level component of the linearized series. Trend component recorded a strong decline from 18 February to the second half of April and in the second quarter of 2020, it showed a fluctuating trend, marking an upward trend until the end of May, and then undergoing a new decline in the first 10 days of June. From mid-June, it showed an upward trend peaking in September 2020. Seasonal component showed positive peaks in December 2019 and from the end of March to early April 2020, whereas a negative peak started in the summer months and peaks in September (figure 1).

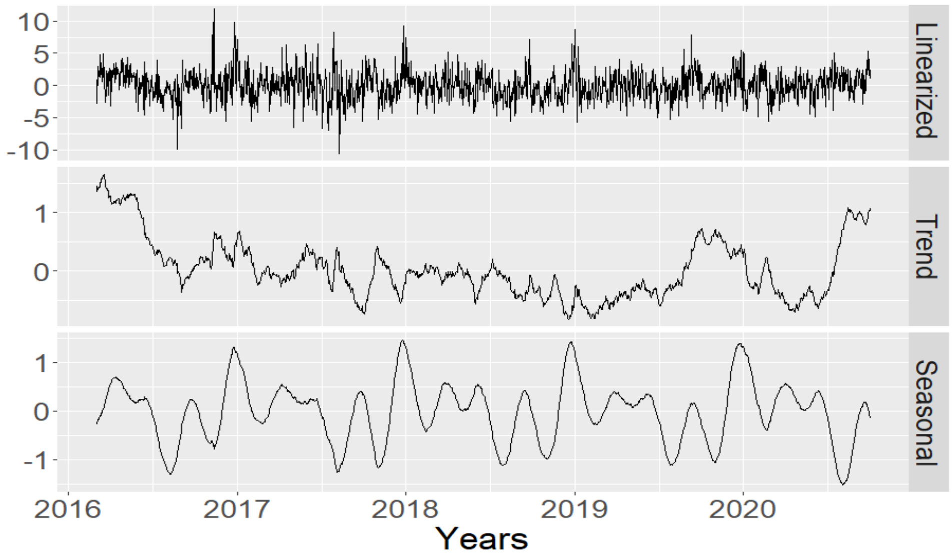

Figure 1 - Decomposition of the daily time series of the Social Mood on Economy Index

4.1 A test on the robustness of SMEI during the Covid-19 pandemic

To answer to our first research question, on the robustness of the SMEI during the pandemic, we tried to understand how the SME index reacted during the pandemic; in particular, if the considered tweets in the measurement were sufficiently inclusive of the talks on the pandemic and the problems related to the health emergency and economic crisis.

As the series are too short, we could not perform other econometric analysis. Thus, with graphical analysis, we compared the filtered Covid SME series to the Google Trends series referring to Covid/Coronavirus shares of queries and to the series of the death by Coronavirus observed in Italy.

Signals extracted are quite similar, in both cases series peak on 23-24 February 2020. They present a negative peak at the beginning of March and an increase before mid-March. After March, the interest for the keywords reduced by 50% and the decrease continued until June, when both the shares remain between 10% and 20%. The increase of Google Trends series started again in August (with a peak registered on 22 August 2020), afterwards opening a declining phase with a following rapid increase observed in October. In this latest phase, the increase is less evident for the filtered Covid SME series, even if we observe peaks at the beginning and at the end of September.

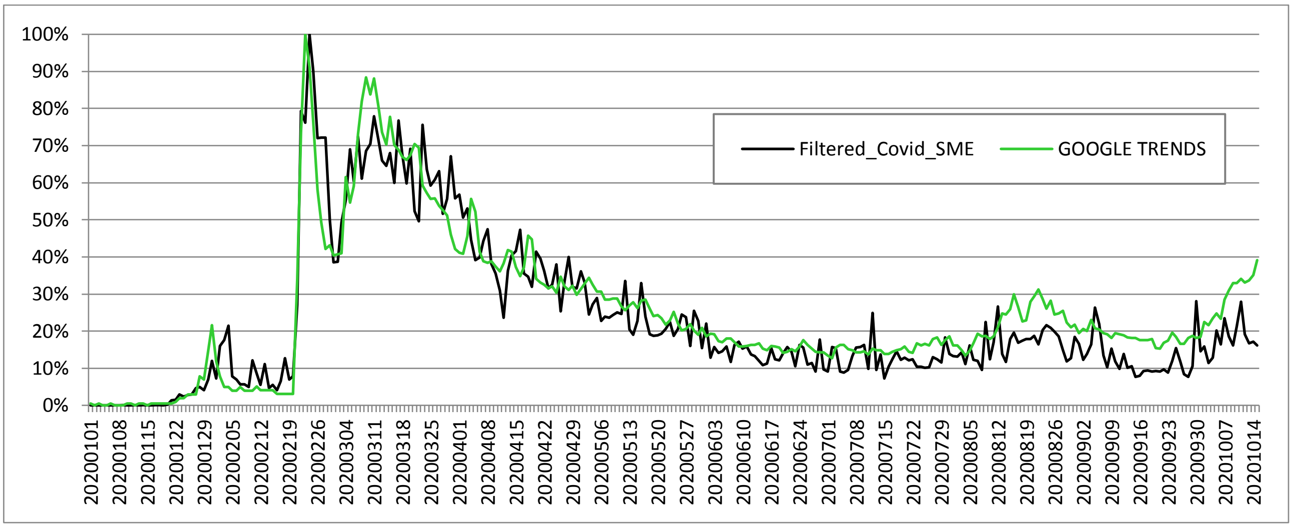

Figure 2 - Comparison of the Istat filtered Covid SME and the Google Trends Covid/Coronavirus query shares series – 01 Jan 2020 – 15 Oct 2020

We made comparisons also with the daily series on the number of newly infected persons and deaths coming from the Civil Protection Department health risk associated with Coronavirus database. The series of filtered Covid SME and the Covid deaths behave more similarly after 24 March 2020; the series evolved similarly decreasing very rapidly and remaining to almost the same level (between 10% and 20%, in the case of Twitter SMEI) from the beginning of June 2020 onward. Whereas the series of newly positive cases register an increase after the summer period not observed in the filtered Covid SME series.

The interest in the Covid-19 issue among the Italians declined with the reduction of the severity of the sanitary emergency and the increase in the new cases in September and October 2020 caused a sligh increase of the number of the filtered Covid SME conversations.

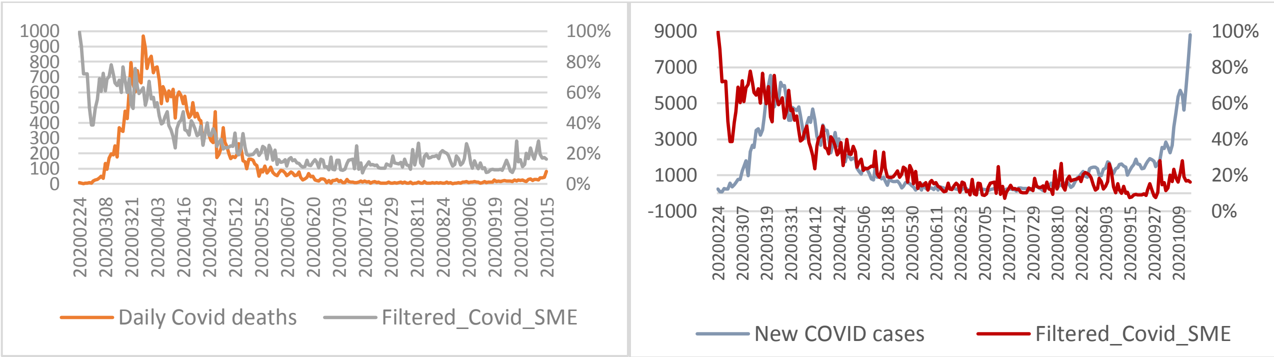

Figure 3 - Comparison between the Istat filtered Covid SME series and the series of the number of daily new COVID cases and deaths in Italy - 01 Jan 2020 – 15 Oct 2020

4.2 Relationships between SMEI with other daily and monthly nontraditional economic series

The SME index has been compared with other economic indicators both on monthly and daily frequency.

Monthly indicators considered are the consumer confidence indicators (total and main components), the BI-COMP card transaction by Merchant category code groups and the e-commerce monthly transaction both in terms of indices (2014=100). As the comparison with the seasonally adjusted monthly indicator and the monthly average of the daily series of the level of the SME index showed a low contemporaneous correlation and a weak predictive power of the SME index for the traditional monthly indicators. We think that averaging the SME index we are losing some important information given by the high-frequency characteristic of the index.

We look for more daily time series on the payment system of financial transaction, in particular, BI-COMP series referring to POS and ATM transactions and Target2 series. The daily time series have been pre-treated for the outlier detection procedure and the removal of other deterministic effects with a similar approach for the SME index. To consider the problem that transaction data are not available for holidays and weekends, a procedure to spread data on Mondays for the weekends has been implemented according to a weighting procedure.

Between the daily series the one that shows better performance in terms of correlation is the time series on POS. In Figure 4, the two series are plotted together for the most recent period in which the correlation is higher.

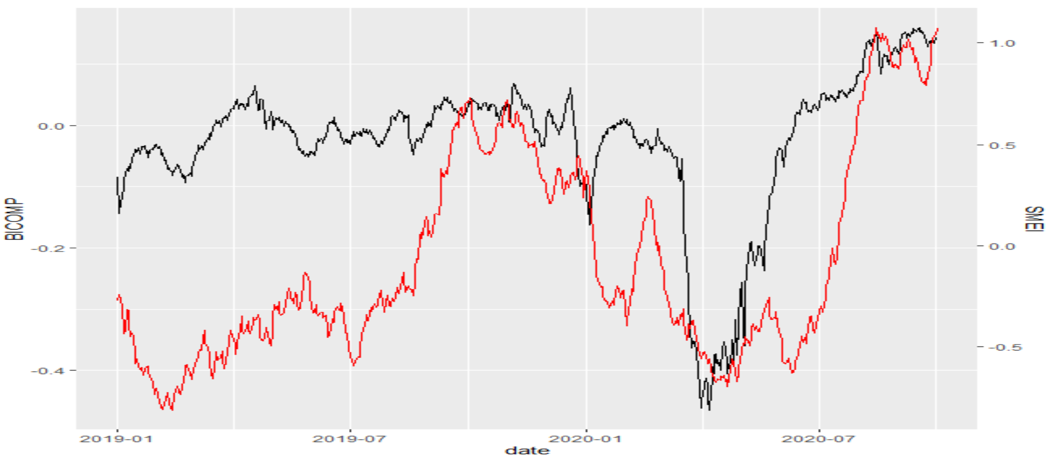

Figure 4 – Comparison between SMEI (in level, in red line right axes) and BI-COMP POS (level, in black line, left axes) series

A more detailed analysis has been carried out on both series to answer to our second research question on the relationships between SMEI and other macroeconomic indicators. The SME index and the POS series are both integrated. A cointegration test showed that the series are cointegrated. For this reason, we estimated a Vector Correction Error Model for each series in the following terms:

$$\Delta POS_t = -0.011(POS_{t-1} - 0.15SMEI_{t-1}) + 0.14\Delta POS_{t-1} - 0.001\Delta SMEI_{t-1} + 0.07\Delta POS_{t-2} - 0.02\Delta\, SMEI_{t-2}$$

$$\Delta SMEI_t = 0.037(POS_{t-1} - 0.15SMEI_{t-1}) - 0.15\Delta POS_{t-1} + 0.3\Delta SMEI_{t-1} - 0.04\Delta POS_{t-2} - 0.03\Delta\, SMEI_{t-2}$$

The results of the VCE models, reported in tables 1 and 2, show a positive cointegration relationship between the two variables indicating a co-movement between SMEI and BI-COMP POS series.

Table 1 - Coefficients of POS equation of VECM with relative standard errors

Delta POS(t)	Estimate	Std.error	t-value	p-value
POS(t-1) -0.15 SMEI(t-1)	-0.01123	0.003913	-2.869	0,00418
constant	0.000241	0.000293	0.824	0.41028
Delta POS(t-1)	0.136987	0.027634	4.957	8.10E-07
Delta SMEI(t-1)	-0.00984	0.007498	-1.312	0.18962
Delta POS(t-2)	0.072992	0.027767	2.629	0.00867
Delta SMEI(t-2)	-0.02402	0.007824	-3.07	0.00219

Table 2 - Coefficients of SMEI equation of VECM with relative standard errors

Delta SMEI(t)	Estimate	Std.error	t-value	p-value
POS(t-1) -0.15 SMEI(t-1)	0.037532	0.014432	2.601	0.00941
Constant	-0.00032	0.00108	-0.297	0.76667
Delta POS(t-1)	-0.15749	0.101925	-1.545	0.12255
Delta SMEI(t-1)	0.297268	0.027656	10.749	2.00E-16
Delta POS(t-2)	-0.04453	0.102415	-0.435	0.66378
Delta SMEI(t-2)	0.031468	0.028856	1.091	0.2757

5. Discussion and conclusions

We studied the evolution of the Istat SME index during the Covid-19 pandemic performed. We wanted better understand if the index had correctly caught the worries of twitterers for the health emergency and the related economic crisis. We observed that the signals on the tracking of terms Covid and Coronavirus in the SMEI tweets were comparable to the evolution of the search for the same terms in Google Trends and to the evolutions of the new Covid positive case and deaths and we found several similarities. Thus, we argued that the index correctly reported the Crisis.

Furthermore, new nontraditional sources allowed us a better understanding of the explicative economic power of SMEI. We could verify that SMEI, a relevant Istat experimental indicator for the analysis and the short-term forecasting, and other traditional monthly macroeconomic indicators have very different trends during the Covid-19 crisis period.

Leveraging the timeliness of the high frequency of the daily series, we could observe a positive correlation between the SMEI and the BI-COMP POS daily transaction series. As a further development, we should try to understand if the pandemic has widened or reduced the correlation and cointegration among SMEI and payment series.

Considering these positive results, we will continue to investigate through the SMEI the impact of the economic crisis and try to include some information into Istat economic and forecasting models.

References

Basile, V., Missim, M. (2013). Sentiment analysis on Italian tweets. *4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA 2013*, June 14th 2013, (pp. 100-107). Atlanta, U.S.A.

Coletto, M., Lucchese, C., Orlando, S., Perego, R. (2016b). Polarized user and topic tracking in Twitter. *SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, July (pp. 945-948).

De Livera, A. M., Hyndman, R. J., Snyder, R. D. (2011), Forecasting Time Series with Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American Statistical Association*, 106, 1513-1527. https://doi.org/10.1198/jasa.2011.tm09771

Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC* (Vol. 6, pp. 417-422).

Fabbri, C., Iannaccone,R., Righi, A., Scannapieco, M., Testa, P., Valentino, L., Zardetto, D., Zurlo, D., (2018) *The Social Mood on Economy Index. Methodological note*. Rome: Istat.

Feldman, R., Sanger, J. et al. (2007). *The Text mining Handbook: advanced approaches in analysing unstructured data*. Cambridge: Cambridge U.P.

Gandomi, A., Haider, M. (2015). Beyond the hype. Big data concepts, methods and analytics. *International Journal of Information management*, 35(2):137-144.

Garimella, K., De Francisci Morales, G., Gionis, A. & Mathioudakis, M. (2016). Quantifying controversy in social media. *The 9th ACM International Conference on Web Search and Data Mining WSDM '16*, (pp. 33–42) San Francisco, California.

Ghyels, G., Synko, A., Valkanov, R. (2007), MIDAS regressions: further results and new directions. *Econometric Reviews*, 26, 53-90

Hyndman, R. J., & Khandakar, Y. (2007). *Automatic time series for forecasting: the forecast package for R* (No. 6/07). Clayton VIC, Australia: Monash University, Department of Econometrics and Business Statistics.

Kolchyna, O., Souza, T. T., Treleaven, P. & Aste, T. (2015). Twitter sentiment analysis: lexicon method, machine-learning method and their combination. arXiv preprint arXiv:1507.00955.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41.

Pak, A., Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Seventh Conference on International Language Resources and Evaluation. LREC 2010*, 17-23 May 2010, Valletta, Malta, (pp. 1320–1326).

Strapparava, C., & Valitutti, A. (2004, May). Wordnet affect: an affective extension of wordnet. In *LREC* (Vol. 4, No. 1083-1086, p. 40).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. (2011). Lexicon-based methods for sentiment analysis, *Computational linguistics*, 37 (2), 267–307.

Zardetto, D. (2020). Using Twitter Data for the Social Mood on Economy Index, Atti della XIII Conferenza nazionale di *statistica*, Rome, 4-6 July 2018, ISBN 978-88-458-2016-8, (pp. 385-390).