The Human Side of Service Engineering, Vol. 62, 2022, 201–209
https://doi.org/10.54941/ahfe1002559

Predicting Economic Indicators Using Political Texts

Alessandra Amendola[1], Alessandro Grimaldi[1], and Walter Distaso[2]

[1]University of Salerno, Salerno, Italy
[2]Imperial College Business School, London, UK

ABSTRACT

The paper proposes a new text-based indicator aimed at assessing the impact over time of political debate on economy. Textual data from the plenary verbatim reports of the Italian Parliament are pre-processed and relevant themes, whose temporal evolution allows predicting fluctuations in fundamental macro-economic variables, are estimated via a Correlated Topic Model. Specifically, a Political Debate Index is derived based on a time-varying weighting function of the estimated topic proportions. The capability of the proposed approach in improving the predictability of selected economic indicators is evaluated considering different predictors. The reached results seem to support the evidence that qualitative information conveyed by the daily political debate does have an impact on the economic dynamic over time and can be usefully used to improve the economic predictions performance.

Keywords: NLP, Topic modeling, Text as data, Parliamentary debate, Time series, Economic indices

INTRODUCTION

Recently, applications of topic modeling in economics have dramatically increased with researchers mainly focused on: the analysis of the evolution of the economic literature over time; the prediction of stock prices, returns, and volatility; the analysis of the effects of central banks communication; the development of text-based indices and economic. See, for instance, Lüdering and Winker (2016), Adämmer and Schüssler (2020), Cerchiello and Nicola (2018), Baerg and Lowe (2020), Hansen, McMahon, and Prat (2017), Angelico et al. (2021), Larsen and Thorsrud (2019).

So far, the preferred sources of texts have been social media, newspapers, and, in some cases, transcripts of Central Bank Governors or Presidential speeches. Other types of texts such as politicians' speeches, which also might be of interest from an economic perspective, have been somewhat disregarded in the field, although they are quite frequent in social and political science works. See, for instance, Gentzkow, Shapiro, and Taddy (2019), or Grimmer (2010), among others.

In this work, we explore parliamentary records through topic models whose output is utilized to construct economic indices integrating qualitative information conveyed by texts with purely quantitative information from traditional economic measures.

DATA DESCRIPTION

The construction of the proposed indices requires two types of data: the transcripts of parliamentary debates and the time series of economic indicators. The procedure described is entirely replicable for any language and Country. We choose to focus on Italy motivated by the following considerations. First, this Country is a rather peculiar case when it comes to the complexity of its political system – for instance, in the considered 24 years, 14 Governments succeeded one another. Second, with few exceptions (e.g. Angelico et al. (2021) or Larsen and Thorsrud (2019)), languages other than English are still quite rare in such literature.

Moving to more technical information, we measure time in quarters. Sample period (S) instants are indicated as $t = 1, \ldots, T = 98$, corresponding to calendar quarters 1996:Q2 – 2020:Q3.

The Italian Senate Parliamentary Reports

We analyze more than 4,300 parliamentary verbatim reports of the *Italian Senate of the Republic* from 09 May 1996 to 08 September 2020, i.e. 6 Legislatures – from the 13th to the (still ongoing) 18th.

After extraction from *.pdf* format, thanks to their consistent structure we split the reports into the single speeches given by orators, and construct the data frame used for standard cleaning (Banks et al., 2018; Denny and Spirling, 2018) and stemming – i.e. the reduction of inflected words to their base form – to reduce data dimensionality. To reduce noise and increase efficiency, very short speeches of less than 10 words are filtered out. The remaining speeches are aggregated on a daily basis before rearranging them in a *document-term matrix* (DTM). Precisely, to check topic model sensitivity to the dictionary magnitude, 5 different DTM are constructed by selecting the top 20, 40, 60, 80, and 100 percent of most relevant words in terms of their *Term Frequency-Inverse Document Frequency* (TF-IDF) transformation (Sammut and Webb, 2011). Details are given in Table 1.

The Economic Variables

Similar to Larsen and Thorsrud (2019), to link parliamentary debate and economic dynamics, we use 8 national accounts statistics. Specifically, the time series of the *output* (Y), *imports* (M), *consumption* (C), *government expenditure* (G), *investments* (I), *exports* (X), *wages* (W), and *taxation* (T) are retrieved from the Italian National Institute of Statistics (ISTAT) website. In detail, Y is the gross domestic product at market prices; M and X are the imports and exports of goods and services; C is the final consumption expenditure of households and non-profit institutions serving households; G is the consumption of general government; I is the gross fixed capital formation; W is the domestic compensation of employees; T is the taxes minus the subsidies on production and imports. Aggregates are measured in current prices millions of euros. All the time series are quarterly based and seasonally adjusted by the source via the Tramo-Seats procedure which accounts also for calendar effects wherever they are present (Istat, n.d.).

Table 1. Aggregated DTM details.

DTM Statistics			Terms Counts Statistics											
Dimensions		Sparsity	Across Corpus						Across Vocabulary					
Sparsity size	Vocabulary size	(%)	Min	25%	Median	Mean	75%	Max	Min	25%	Median	Mean	75%	Max
2,858	73,638	96.98	38	7,390	11,821	12,312	16,812	68,538	1	1	3	478	17	459,095
2,858	58,828	96.23	38	7,388	11,816	12,306	16,803	68,531	1	2	5	598	31	459,095
2,858	44,006	94.98	38	7,387	11,802	12,298	16,787	68,524	1	3	11	799	66	459,095
2,858	29,208	92.49	38	7,376	11,779	12,279	16,748	68,496	1	10	32	1,202	178	459,095
2,858	14,557	85.35	38	7,331	11,686	12,198	16,654	68,304	1	65	179	2,395	749	459,047

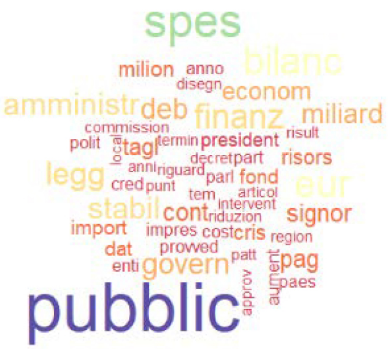

Figure 1: 100% TF-IDF corpus $K = 100$ CTM: Topic 61.

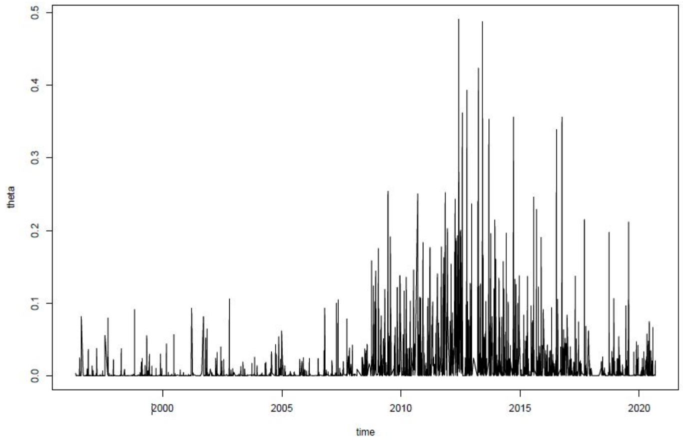

Figure 2: 100% TF-IDF corpus $K = 100$ CTM: Topic 61 daily proportions.

METHODOLOGY

The Correlated Topic Model

In the field, the Latent Dirichlet Allocation (LDA) "*of Blei, Ng, and Jordan (2003) is the most prominent and widely applied topic model*" (Adämmer and Schüssler, 2020). However, it does not allow for correlations among topics. Hence, as all documents belong to the same collection (Vayansky and Kumar, 2020), we apply the Correlated Topic Model (CTM) of Blei and Lafferty (2007) to each of the 5 CTM. As a result, we estimate the *per-topic term probabilities* β_k, where $k \in \{1, \ldots K\}$ and K is the total number of topics, and the *per-document topic proportions* θ_d, where $d \in \{1, \ldots, D\}$ and D is the total number of documents. Figure 1 shows an estimated topic (β), Figure 2 its corresponding document proportions (θ).

The Optimal Number of Topics

A key aspect in topic modeling is determining K. To address the issue, a data-driven strategy is adopted, similar to Adämmer and Schüssler (2020): several CTM are estimated and compared through diagnostic statistics. Specifically, for each of the 5 DTM previously built, 15 models are estimated with a K

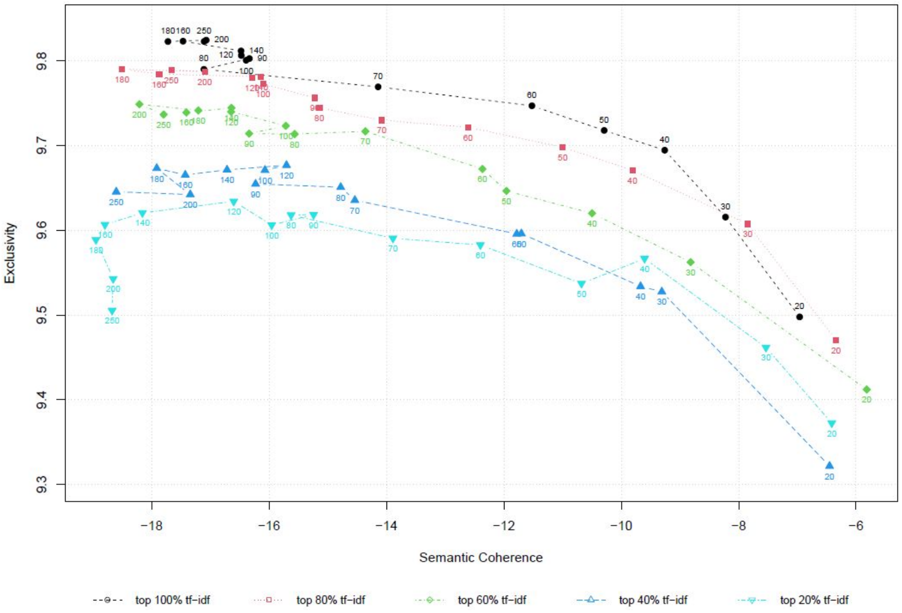

Figure 3: CTM Diagnostics by K and DTM.

Table 2. TPDI.

Index Type	ARX-AR Based	TVARX-TVAR Based
I_0	$\sum_{i=1}^{K} b_i\widetilde{\theta}_{i,t-1}$	$\sum_{i=1}^{K} b_{i,t-1}\widetilde{\theta}_{i,t-1}$
I_1	$\sum_{i=1}^{K} w_i b_i\widetilde{\theta}_{i,t-1}$	$\sum_{i=1}^{K} w_i b_{i,t}\widetilde{\theta}_{i,t-1}$
I_2	$\sum_{i=1}^{K} v_t b_i\widetilde{\theta}_{i,t-1}$	$\sum_{i=1}^{K} v_t b_{i,t}\widetilde{\theta}_{i,t-1}$
I_3	$\sum_{i=1}^{K} w_i v_t b_i\widetilde{\theta}_{i,t-1}$	$\sum_{i=1}^{K} w_i v_t b_{i,t}\widetilde{\theta}_{i,t-1}$

Table 3. Best model type per K and economic aggregate.

Number of Topics				Economic Variables				
K	Y	M	C	G	I	X	W	T
20	arx I2	arx I2	tvarx I2	arx I2	arx I2	arx I3	tvarx I2	tvarx I2
40	arx I2	arx I3	arx I3	arx I2	tvarx I3	arx I3	tvarx I2	arx I2
60	arx I2	tvarx I3	arx I3	arx I2	tvarx I3	tvarx I3	tvarx I3	arx I3
80	tvarx I3	arx I3	arx I2	tvarx I2	arx I3	tvarx I3	tvarx I3	arx I2
100	arx I3	arx I3	tvarx I3	arx I2	arx I2	tvarx I3	tvarx I3	arx I3
120	arx I2	arx I2	tvarx I3	arx I3	tvarx I2	tvarx I2	tvarx I3	arx I3
140	arx I3	tvarx I3	arx I3	arx I3	tvarx I2	arx I3	arx I3	arx I2

varying from 20 to 250. *Exclusivity* (Roberts et al., 2014) and semantic *coherence* (Mimno et al., 2011; Roberts et al., 2014) are the main metrics adopted for evaluation. High exclusivity means that words receiving high probabilities in a specific topic also receive low probability in other topics. High semantic coherence means that the estimated topics are easily interpreted by human

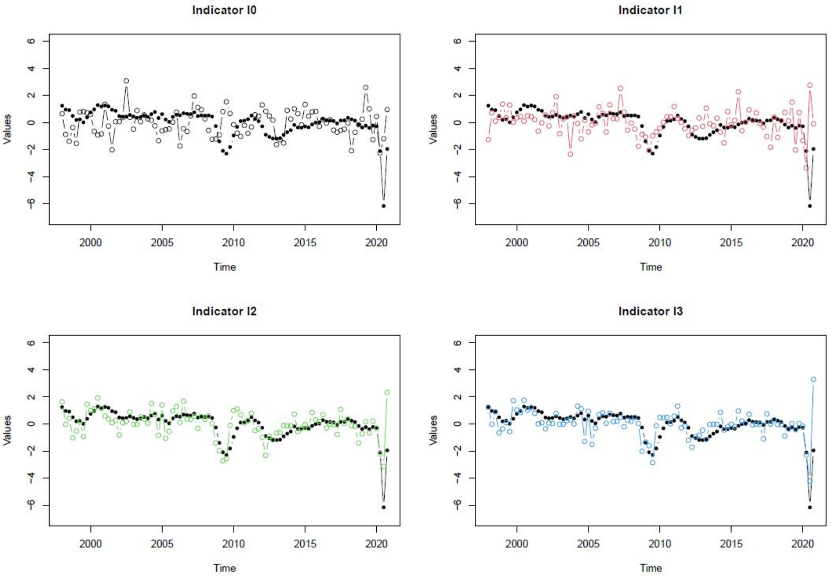

Figure 4: 100% TF-IDF DTM, $K = 100$ GDP ARX/AR TPDI vs GDP growth rate.

readers. As shown in Figure 2, there is a trade-off between the two. From graphical inspection, we choose $K = 100$ as the baseline for each of the five DTM as a good compromise between the two measures.

Textual Political Debate Indices (TPDI)

To construct the proposed Textual Political Debate Indices (TPDI), we first aggregate the estimated daily topic proportions θ on a quarterly basis to align them with the economic time series. The obtained $\widetilde{\theta}$ are the core of each TPDI. Then, to ensure stationarity, we consider the standardized year-on-year logarithmic differences of all economic series. Afterward, we model them with auto-regressive models, AR, (Box and Jenkins, 1970) and auto-regressive with exogenous inputs models, ARX, (Hannan, 1976) as well as time-varying parameters auto-regressive models (Cai, 2007), either with exogenous terms (TVARX) or without (TVAR). Based on Larsen and Thorsrud (2019), all auto-regressions are of order 1 and each (TV)ARX model includes the lagged values of $\widetilde{\theta}$ of one single topic at a time. Hence, for each of the 8 economic aggregates considered, there are $2(K + 1)$ regressions whose estimation allows us to measure the contribution, b_i (in case of ARX) or $b_{i,t}$ (in case of TVARX), of topic $i = 1, \ldots, K$, in explaining the economic variable at hand. Table 2 shows the formulae for the 4 types of TPDI under the time-varying and non-time-varying regression frameworks.

In Table 2, $0 \leq w_i, v_t \leq 1$ with $i = 1, \ldots, K$ and $t = 1, \ldots, T$ are the inter-topics and infra-time normalized weights, respectively. In formulae, $w_i = \frac{\widetilde{w}_i - \min(\widetilde{w}_i)}{\max(\widetilde{w}_i) - \min(\widetilde{w}_i)}$, where $\widetilde{w}_i = \frac{R^2_{i(TV)ARX}}{R^2_{i(TV)AR}}$ are the raw inter-topics weights and $R^2_{i(TV)ARX}$ are the R-squared of the (time-varying) regressions;

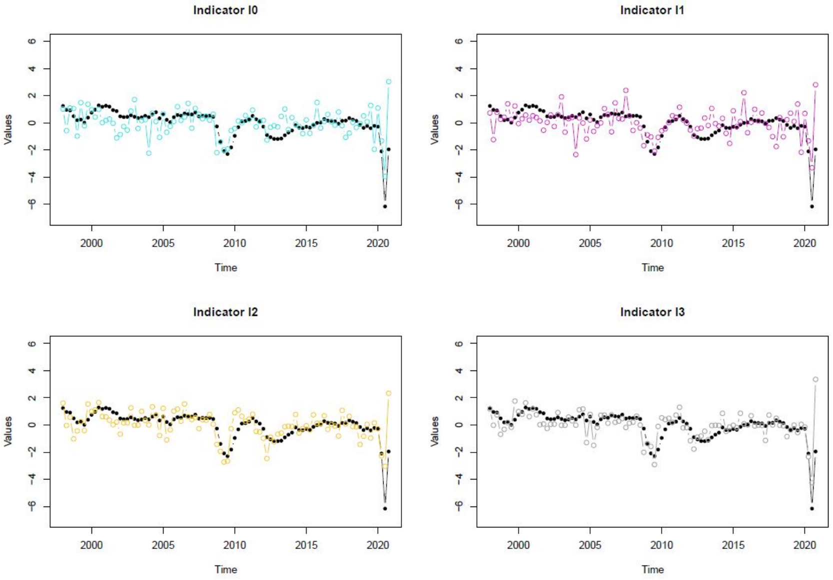

Figure 5: 100% TF-IDF DTM, $K = 100$ GDP TVARX/TVAR TPDI vs GDP growth rate.

$v_t = \frac{\tilde{v}_t - \min(\tilde{v}_t)}{\max(\tilde{v}_t) - \min(\tilde{v}_t)}$, where $\tilde{v}_t = \frac{\sum_{i=1}^{K} \tilde{u}^2_{i,t}}{\tilde{u}^2_{i,t}}$ are the raw infra-time weights with $\tilde{u}^2_{i,t} = \frac{\hat{u}^2_{i,t(TV)ARX}}{\hat{u}^2_{i,t(TV)AR}}$ being the ratios between the squared residuals $\hat{u}^2_{i,t(TV)ARX}$ from the (time-varying) regressions. For each of the 8 economic aggregates, after construction, the 4 types of indices are evaluated by performing the Model Confidence Set (MCS) procedure introduced by Hansen, Lunde, and Nason (2011). To check the sensitivity of the indices to the dictionary, the overall process of indices construction and evaluation is repeated 5 times: one per each of the DTM previously built. Results are shown in Table 3.

As shown in Figures 4 and 5, which refer to the GDP only, the proposed indices are able to closely mimic the dynamic of traditional economic measures.

CONCLUSION

In this work, we derive TPDI: a class of text-based indices to measure economic activity and capture qualitative information from political debate. Specifically, the proposed indices closely mimic the dynamic of 8 traditional national account statistics. Results are achieved through a topic models analysis of Italian Senate parliamentary reports over about 24 years through a data-driven procedure topics estimation does not rely on a predefined set of words – which is replicable for other languages and Countries. Moreover, differently from traditional measures, the proposed indices may be computed at higher frequencies, hence providing timelier information on current economic dynamics, potentially improving the forecasting accuracy of the state-of-the-art models.

REFERENCES

Adämmer, Philipp and Rainer A. Schüssler (May 2020). “Forecasting the Equity Premium: Mind the News!” In: Review of Finance 24.6, pp. 1313–1355. ISSN: 1572–3097. https://doi.org/10.1093/rof/rfaa007

Angelico, Cristina et al. (Feb. 2021). “Can we measure inflation expectations using Twitter?”. Bank of Italy, Economic working papers 1318. https://dx.doi.org/10.2139/ssrn.3827489

Baerg, Nicole and Will Lowe (2020). “A textual Taylor rule: estimating central bank preferences combining topic and scaling methods”. In: Political Science Research and Methods 8.1, pp. 106–122.ISSN: 2049-8470. https://doi.org/10.1017/psrm.2018.31

Banks, George C. et al. (2018). “A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App)”. In: Journal of Business and Psychology 33 (4), pp. 445–459. ISSN: 1573-353X. https://doi.org/10.1007/s10869-017-9528-3

Blei, David M. and John D. Lafferty (June 2007). “A correlated topic model of Science”. In: Ann. Appl. Stat.1.1, pp. 17–35. https://doi.org/10.1214/07-AOAS114

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). “Latent Dirichlet Allocation”. In: J. Mach. Learn. Res.3. Jan, 993–1022. ISSN: 1532-4435

Box, George EP and Gwilym M Jenkins (1970). Time series analysis: forecasting and control. 1st ed. Oakland: Holden-Day

Cai, Zongwu (2007). “Trending time-varying coefficient time series models with serially correlated errors”. In: Journal of Econometrics 136.1, pp. 163–188. ISSN: 0304–4076. https://doi.org/10.1016/j.jeconom.2005.08.004

Cerchiello, Paola, Paolo Giudici, and Giancarlo Nicola (2017). “Twitter data models for bank risk contagion”. In: Neuro computing 264. Machine learning in finance, pp. 50 –56. ISSN: 0925–2312. https://doi.org/10.1016/j.neucom.2016.10.101

Denny, Matthew J. and Arthur Spirling (2018). “Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It”. In: Political Analysis 26.2, 168–189. https://doi.org/10.1017/pan.2017.44

Gentzkow, M., Shapiro, J.M., Taddy, M.: Measuring group differences in high-dimensional choices: Method and application to congressional speech. Econometrica87(4), 1307–1340 (2019). https://doi.org/10.3982/ECTA16566

Grimmer, J.: A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. Political Analysis18(1), 1–35 (2010). https://doi.org/10.1093/pan/mpp034

Hannan, E. J. (1976). “The Identification and Parameterization of Armax and State Space Forms”. In: Econometrica 44.4, pp. 713–723. ISSN: 00129682, 14680262. http://www.jstor.org/stable/1913438

Hansen, Peter R., Asger Lunde, and James M. Nason (2011). “The Model Confidence Set”. In: Econometrica 79.2, pp. 453–497. https://doi.org/10.3982/ECTA5771

Hansen, Stephen, Michael McMahon, and Andrea Prat (2017). “Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach”. In: The Quarterly Journal of Economics133.2, pp. 801–870. ISSN: 0033–5533. https://doi.org/10.1093/qje/qjx045

Istituto Nazionale di Statistica (n.d.). I.Stat corporate data warehouse. http://dati.istat.it/

Larsen, Vegard H. and Leif A. Thorsrud (2019). “The value of news for economic developments”. In: Journal of Econometrics 210.1, pp. 203–218. ISSN: 0304–4076. https://doi.org/10.1016/j.jeconom.2018.11.013

Lüdering, Jochen and Peter Winker (2016). “Forward or Backward Looking? The Economic Discourse and the Observed Reality”. In: Jahrbücher für Nationalökonomie und Statistik 236.4, pp. 483–515. https://doi.org/10.1515/jbnst-2015-1026

Mimno, David et al. (2011). “Optimizing Semantic Coherence in Topic Models”. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, pp. 262–272. https://www.aclweb.org/anthology/D11-1024

Roberts, Margaret E. et al. (2014). “Structural Topic Models for Open-Ended Survey Responses”. In: American Journal of Political Science 58.4, pp. 1064–1082. https://doi.org/10.1111/ajps.12103

“TF–IDF” (2011). In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_832

Vayansky, Ike and Sathish A.P. Kumar (2020). “A review of topic modeling methods”. In: Information Systems 94, p. 101582. ISSN: 0306-4379. https://doi.org/10.1016/j.is.2020.101582