Predicting economic indicators using textual data from Parliament verbatim reports

Alessandra Amendola & Alessandro Grimaldi
Università degli Studi di Salerno
Dipartimento di Scienze Economiche e Statistiche
alamendola@unisa.it; algrimaldi@unisa.it

1 Aim

We believe that what elected representatives say when discussing and making decisions in legislatives assemblies does have a *measurable* impact on the economy.

Therefore, we use topic models to investigate texts of a large sample of parliamentary records and try to quantify such relation.

Finally, we construct new economic indicators which are innovative in that they integrate qualitative information conveyed by textual data with purely quantitative information coming from standard economic measures such as, for instance, GDP.

2 Data

Italy is a rather peculiar case when it comes to the stability if its political system.

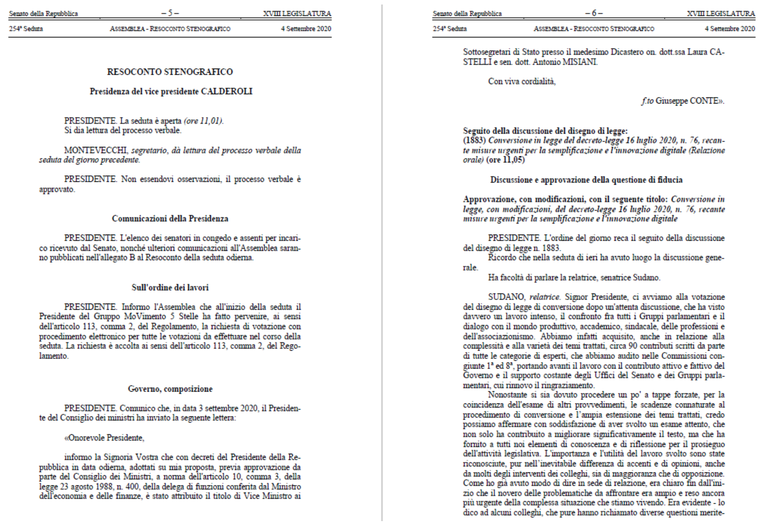
In this project we focused on the texts of the Italian Senate verbatim reports for all sittings over the last 24 years up to 8 September 2020.

Table 1: The Senate verbatim reports corpus over time, Legislatures and Goverments

Legislature	Elections	Duration in days	Prime Minister	Sittings
XVIII	4 Mar 2018	900 (8 Sept 2020)	Conte II Conte	254 (8 Sept 2020)
XVII	24 - 25 Febr 2013	1,834	Gentiloni Renzi Letta	923
XVI	13 - 14 Apr 2008	1,781	Monti Berlusconi IV	860
XV	9 - 10 Apr 2006	732	Prodi II	283
XIV	13 May 2001	1,794	Berlusconi III Berlusconi II	965
XIII	21 Apr 1996	1,847	Amato II D'Alema II D'Alema Prodi	1,061
		Total		**Total**
		8,888 ≈ 24 years		4,346

The structure of the texts is consistent over time and allowed us to split the single speeches given by orators in turn and build a data frame from the texts corpus.

Figure 1: Sample Italian Senate verbatim reports pages structure



Senato della Repubblica — 5 — XVIII LEGISLATURA
254ª Seduta — ASSEMBLEA - RESOCONTO STENOGRAFICO — 4 Settembre 2020

RESOCONTO STENOGRAFICO

Presidenza del vice presidente CALDEROLI

PRESIDENTE. La seduta è aperta *(ore 11,01)*.
Si dia lettura del processo verbale.

MONTEVECCHI, *segretario, dà lettura del processo verbale della seduta del giorno precedente.*

PRESIDENTE. Non essendovi osservazioni, il processo verbale è approvato.

Comunicazioni della Presidenza

PRESIDENTE. L'elenco dei senatori in congedo e assenti per incarico ricevuto dal Senato, nonché ulteriori comunicazioni all'Assemblea saranno pubblicati nell'allegato B al Resoconto della seduta odierna.

Sull'ordine dei lavori

PRESIDENTE. Informo l'Assemblea che all'inizio della seduta il Presidente del Gruppo MoVimento 5 Stelle ha fatto pervenire, ai sensi dell'articolo 113, comma 2, del Regolamento, la richiesta di votazione con procedimento elettronico per tutte le votazioni da effettuare nel corso della seduta. La richiesta è accolta ai sensi dell'articolo 113, comma 2, del Regolamento.

Governo, composizione

PRESIDENTE. Comunico che, in data 3 settembre 2020, il Presidente del Consiglio dei ministri ha inviato la seguente lettera:

«Onorevole Presidente,

informo la Signoria Vostra che con decreti del Presidente della Repubblica in data odierna, adottati su mia proposta, previa approvazione da parte del Consiglio dei Ministri, a norma dell'articolo 10, comma 3, della legge 23 agosto 1988, n. 400, della delega di funzioni conferita dal Ministro dell'economia e delle finanze, è stato attribuito il titolo di Vice Ministro ai

Senato della Repubblica — 6 — XVIII LEGISLATURA
254ª Seduta — ASSEMBLEA - RESOCONTO STENOGRAFICO — 4 Settembre 2020

Sottosegretari di Stato presso il medesimo Dicastero on. dott.ssa Laura CASTELLI e sen. dott. Antonio MISIANI.

Con viva cordialità,

f.to Giuseppe CONTE».

Seguito della discussione del disegno di legge:
(1883) *Conversione in legge del decreto-legge 16 luglio 2020, n. 76, recante misure urgenti per la semplificazione e l'innovazione digitale (Relazione orale)* **(ore 11,05)**

Discussione e approvazione della questione di fiducia

Approvazione, con modificazioni, con il seguente titolo: *Conversione in legge, con modificazioni, del decreto-legge 16 luglio 2020, n. 76, recante misure urgenti per la semplificazione e l'innovazione digitale*

PRESIDENTE. L'ordine del giorno reca il seguito della discussione del disegno di legge n. 1883.
Ricordo che nella seduta di ieri ha avuto luogo la discussione generale.
Ha facoltà di parlare la relatrice, senatrice Sudano.

SUDANO, *relatrice*. Signor Presidente, ci avviamo alla votazione del disegno di legge di conversione dopo un'attenta discussione, che ha visto davvero un lavoro intenso, il confronto fra tutti i Gruppi parlamentari e il dialogo con il mondo produttivo, accademico, sindacale, delle professioni e dell'associazionismo. Abbiamo infatti acquisito, anche in relazione alla complessità e alla varietà dei temi trattati, circa 90 contributi scritti da parte di tutte le categorie di esperti, che abbiamo audito nelle Commissioni congiunte 1ª ed 8ª, portando avanti il lavoro con il contributo attivo e fattivo del Governo e il supporto costante degli Uffici del Senato e dei Gruppi parlamentari, cui rinnovo il ringraziamento.
Nonostante si sia dovuto procedere un po' a tappe forzate, per la coincidenza dell'esame di altri provvedimenti, le scadenze connaturate al procedimento di conversione e l'ampia estensione dei temi trattati, credo possiamo affermare con soddisfazione di aver svolto un esame attento, che non solo ha contribuito a migliorare significativamente il testo, ma che ha fornito a tutti noi elementi di conoscenza e di riflessione per il prosieguo dell'attività legislativa. L'importanza e l'unità del lavoro svolto sono state riconosciute, pur nell'inevitabile differenza di accenti e di opinioni, anche da molti degli interventi dei colleghi, sia di maggioranza che di opposizione. Come ho già avuto modo di dire in sede di relazione, era chiaro fin dall'inizio che il novero delle problematiche da affrontare era ampio e reso ancora più urgente dalla complessa situazione che stiamo vivendo. Era evidente - lo dico ad alcuni colleghi, che pure hanno richiamato diverse questioni merite-

After a standard cleaning (punctuation and stop-words removal, ...) and stemming - i.e. reduction of each inflected word to its base form - we aggregated the speeches on a daily (sitting) basis and then re-arranged the corpus as a *Document Feature Matrix* (DFM) with the *features* being the stemmed words placed on columns.

To check topic model sensitivity to the dictionary magnitude, we selected the top 20, 40, 60, 80 and 100 percent of most relevant words in terms of their *Term Frequency - Inverse Document Frequency* (TF-IDF). Details on these DFMs are given in the following table.

Table 2: Aggregated corpus DFM information details

Document-Feature Matrices Statistics			Features Counts Statistics											
Dimensions		Sparsity	Across Corpus						Across Vocabulary					
Corpus 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

3 Methodology

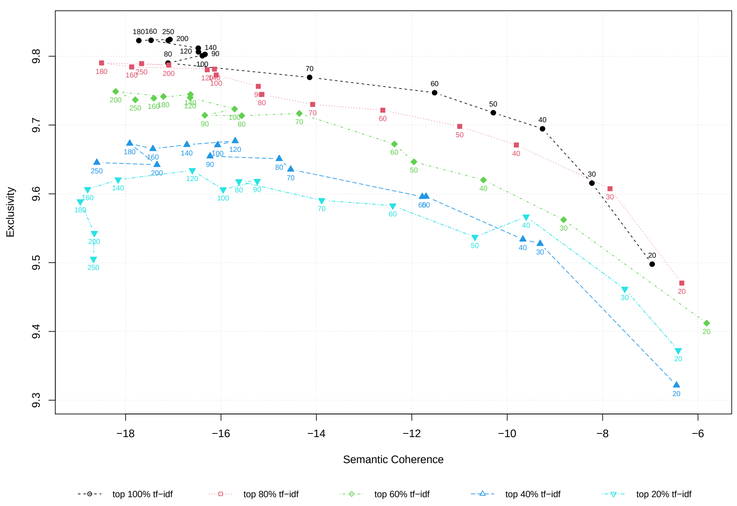
3.1 The Topics

For each DFM we estimated many *Correlated Topic Models* (Blei and Lafferty, 2007) in order to choose a proper number of topics, K - the only parameter to be chosen.

The topics were evaluated by considering their *exclusivity* - i.e. the fact that they are made of words not appearing in other topics - and *semantic coherence* - i.e. how easily they are interpreted by human readers. As shown in the following graph, there is a trade-off between the two measures.

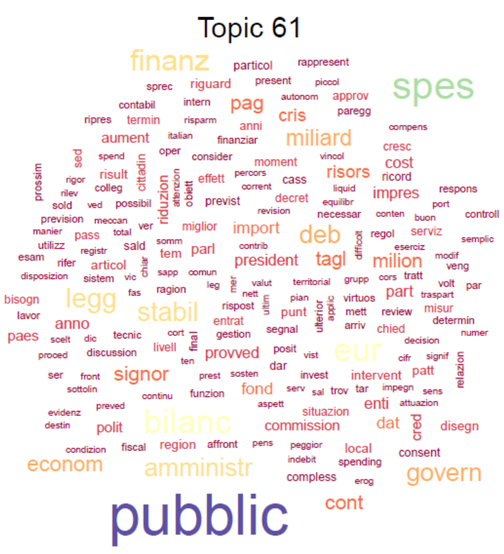
Here we only focus on the 100 topics CTM built on the 100% TF-IDF vocabulary.

Figure 2: Correlated topic todel diagnostics by number of topics and corpus



Topics are essentially probability distributions over the words used in the texts collection. In the following graph, we report a word-cloud representation of the estimated topic n. 61, which seems associated with public expenditure. Biggest words are more frequent and highly associated with such topic.

Figure 3: 100% TF-IDF corpus $K = 100$ CTM topic 61



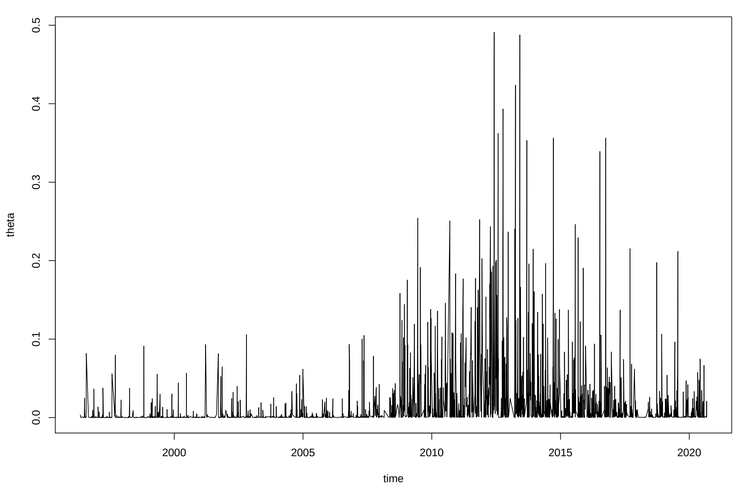
Topic models also estimate the $T \times K$ matrix of topic proportions per document, Θ, where T is the number of documents and K the number of topics.

$$\underset{(T\times K)}{\mathbf{\Theta}} = \begin{bmatrix} \vartheta_{1,1} & \vartheta_{1,2} & \dots & \vartheta_{1,K} \\ \vartheta_{2,1} & \vartheta_{2,2} & \dots & \vartheta_{2,K} \\ \dots & \dots & \dots & \dots \\ \vartheta_{T,1} & \vartheta_{T,2} & \dots & \vartheta_{T,K} \end{bmatrix} \tag{1}$$

As our documents are the daily sitting reports, we can see how the discussion of each topic evolved over time.

As shown in the next graph, proportions of topic n. 61 - public expenditure - peaked in years 2011-2012 during the sovereign debts crisis.

Figure 4: 100% TF-IDF corpus $K = 100$ CTM topic 61 daily proportions over time



3.2 The (Text-based) Economic Indicators

To measure the impact of the estimated topics on economy we worked with the ISTAT time series of the Italian Output, Imports, Consumption, Government Expenditure, Investments, Exports, Wages and Taxation, all measured at *current prices* and adjusted for seasonality and calendar effect by the source.

As the economic series were on quarterly basis (from 1996-Q2 to 2020-Q3), we aggregated the daily topic proportion time series accordingly. To ensure stationarity, we considered the standardized year-on-year logarithmic differences of all series (Larsen and Thorsrud, 2019).

We modelled each economic variable with autoregressive models of order 1, AR, and autoregressive models with exogenous variables, ARX - one for each topic. We also considered the time-varying parameters versions of such models - $TVAR$ and $TVARX$.

This allowed us to estimate the contribution, β, of each topic in explaining the economic variable.

The estimated βs - together with the ϑs - are the core of the text based indicators proposed and detailed in table 3.

Table 3: Text-based economic indicators

Indicator	ARX-AR based	TVARX-TVAR based
I_0	$\sum_{i=1}^{K} \beta_i \vartheta_{i,t-1}$	$\sum_{i=1}^{K} \beta_{i,t} \vartheta_{i,t-1}$
I_1	$\sum_{i=1}^{K} w_i \beta_i \vartheta_{i,t-1}$	$\sum_{i=1}^{K} w_i \beta_{i,t} \vartheta_{i,t-1}$
I_2	$\sum_{i=1}^{K} v_t \beta_i \vartheta_{i,t-1}$	$\sum_{i=1}^{K} v_t \beta_{i,t} \vartheta_{i,t-1}$
I_3	$\sum_{i=1}^{K} w_i v_t \beta_i \vartheta_{i,t-1}$	$\sum_{i=1}^{K} w_i v_t \beta_{i,t} \vartheta_{i,t-1}$

$$0 \leq w_i, v_t \leq 1; \qquad i = 1, ..., K; \qquad t = 1, ..., T$$

$$w_i = \frac{\tilde{w}_i - min(\tilde{w}_i)}{max(\tilde{w}_i) - min(\tilde{w}_i)}; \quad \tilde{w}_i = \frac{R^2_{i(TV)ARX}}{R^2_{i(TV)AR}}$$

$$v_t = \frac{\tilde{v}_t - min(\tilde{v}_t)}{max(\tilde{v}_t) - min(\tilde{v}_t)}; \qquad \tilde{v}_t = \frac{\sum_{i=1}^{K} \tilde{u}^2_{i,t}}{\tilde{u}^2_{i,t}}; \qquad \tilde{u}^2_{i,t} = \frac{\hat{u}^2_{i,t(TV)ARX}}{\hat{u}^2_{i,t(TV)AR}}$$

Here we only focus on the *Output*, defined as the *gross domestic product at market prices*.

The following two graphs show the built indicators (empty dots coloured lines) versus the *Output* series (black full dots line).

Figure 5: 100% TF-IDF corpus $K = 100$ GDP ARX approach text-based economic indicators over time

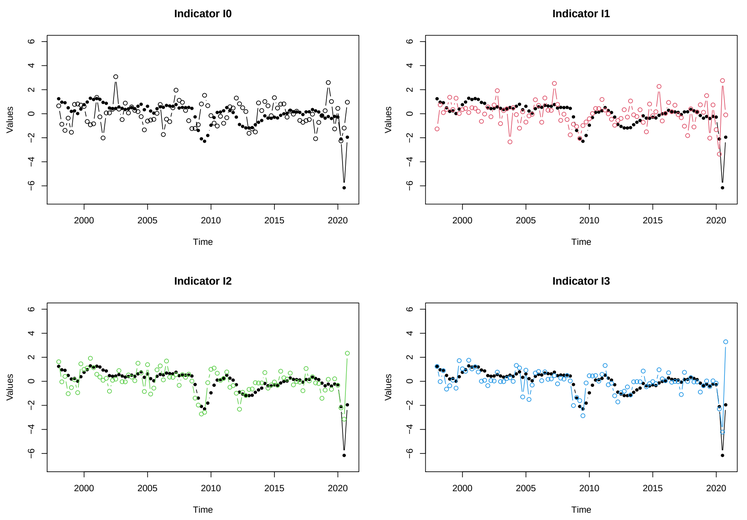
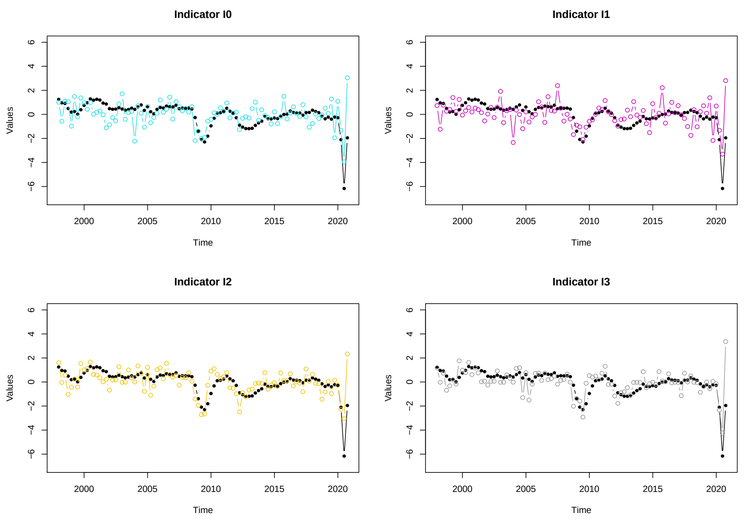


Figure 6: 100% TF-IDF corpus $K = 100$ GDP TVARX approach text-based economic indicators over time



We also used distance measures and performed a model confidence set to find the best indicator in terms of in-sample predictions. The indicator that outperformed the others was the ARX-AR based I_3.

4 Conclusions and Further Research

In this work we showed a way parliamentary debate may be used to obtain new indicators that closely mimic standard economic measures with promising results.

Our research is still ongoing, with the natural next step being forecasting future values of economic variable via our text based indicators.

References

Blei, D. M. and J. D. Lafferty (2007, 06). A correlated topic model of science. *Ann. Appl. Stat. 1*(1), 17–35.

Larsen, V. H. and L. A. Thorsrud (2019). The value of news for economic developments. *Journal of Econometrics 210*(1), 203 – 218. Annals Issue in Honor of John Geweke "Complexity and Big Data in Economics and Finance: Recent Developments from a Bayesian Perspective".