University of Salerno
Department of Economics and Statistics

Ph.D. Programme in
Economics and Policies of Markets and Firms
Curriculum: *Statistical Methods*
Cycle XXXIV

Doctoral Thesis in

Statistical Procedures to Predict Economic Indicators through Political Discourse Text Analysis

Author
Dr. Alessandro Grimaldi
Alessandro Grimaldi

Supervisor
Prof. Alessandra Amendola

Coordinator
Prof. Alessandra Amendola
Amendola

A.Y. 2020/2021

Declaration of Authorship

I, Dr. Alessandro GRIMALDI, declare that this thesis titled, "Statistical Procedures to Predict Economic Indicators through Political Discourse Text Analysis" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Acknowledgements

This thesis is the result of the study and research activity conducted during my PhD, an amazing and challenging experience which allowed me to grow not only technically and professionally but also as a person.

I will admit, the results presented in what follows could not have been achieved without the guidance, advice and support of several people I met on my path whom I feel obliged to and willing to sincerely thank.

First and foremost, I would like to express my deep gratitude to my supervisor and course coordinator, Professor Alessandra Amendola. This thesis would not exist at all if she had not introduced me to the fascinating research area this work belongs to in the first place. Despite the distance due to the pandemic restrictions first and my job later, her presence was constant and punctual throughout my entire journey. With patience, she always encouraged me to persevere even in the most difficult times. I benefited a lot from her deep knowledge, continuous feedback and advice. I will always be grateful to her for believing in me.

Then, my deep appreciation goes to Professor Walter Distaso for his interest, fruitful discussions and comments. I hope our cooperation together with Professor Amendola may continue in the future.

I gratefully acknowledge the two referees, Professors Paola Cerchiello and Giancarlo Ragozini, for their constructive comments and suggestions, which significantly improved the quality of the thesis. I would also like to express my appreciation to the three committee members, Professors Camilla Mastromarco, Marcella Niglio and Giorgia Rivieccio, for their time, interest and helpful comments.

I would express my gratitude to the Director of the Department of Economics and Statistics (DISES), Professor Sergio Pietro Destefanis, for providing the academic support, facilities and infrastructure to accomplish my research activity. I express my gratitude to all professors at DISES for their contribution to my professional and personal life. I sincerely thank the technical members of the Department, specifically Maria Rizzo, Fabio Penna and Pasquale Sabatino, who were always available and willing to provide their kind assistance.

Part of my research was conducted during a visiting period at the Department of Statistics and Econometrics at the Faculty of Economics and Business Administration at the Justus-Liebig University Giessen. I wish to express my gratitude to Professor Peter Winker and the members of his group, with particular regard to Doctor David Lenz, who was also a friend. They provide me with all facilities during my stay. Together we discussed some of the research ideas developed in this thesis. I gained a lot from their supervision and kind assistance in the first steps of my analysis.

Last but not the least, I would like to thank my friends for the enjoyable time spent together, which help me a lot in restoring my spirit during stressful times. I owe a special thanks to Sara, Giuseppe and Antonio, who contributed to my personal life at university. I appreciated their advice and understanding of my feelings throughout this journey, the same they completed during the last years. I wish you to realise your ambitions.

Finally, I would have achieved no result without the unconditional love, patience and support of all my family, especially my parents, Giovanna and Giovanni, who encouraged and helped me every moment of my personal, academic and professional life. To all of you, I owe who I am.

Alessandro Grimaldi

28 September 2022

Contents

List of Figures

List of Tables

To my loved ones

Chapter 1

Introduction

The aim of this thesis is to contribute to the field of textual analysis in economics by proposing text-based indices to quantify the impact of political debate on the economy and improve the forecast accuracy of traditional quantitative measures used to assess the state of the economy.

The task is crucial for private agents and policymakers as the former need to take investment decisions or assess market risk, while the latter need to design the most effective policy strategies for changes in the economic environment. In both cases, accuracy and rapidity are essential. Nonetheless, as noticed by Thorsrud (2020), while financial data are often collected at very high frequency – e.g. infra-daily – on the other hand, official Institutes of statistics compile the principal measures of economic activity at very low frequencies – e.g. monthly, for unemployment data; quarterly or yearly, for GDP and other National Account statistics. In addition, publishing of national statistics may occur with months of delay. Moreover, even when more timely indicators are available, their relationship with variables like the GDP growth rate is unstable and current models struggle to adapt to rapid changes in economic conditions (Thorsrud, 2020).

For such reasons, this thesis employs a source of information somewhat disregarded in the economic literature thus far: parliamentary debates transcripts. Political scientists, instead, are more prone to use such types of texts (Grimmer and Stewart, 2013) as they are often interested in detecting political slants, attitudes of specific (party) groups towards certain subjects or even conflict between them. Recently, Abercrombie and Batista-Navarro (2020) provided an extensive review of text analysis literature concerning this type of text.

However, parliamentary debates transcripts may be of interest also in empirical economic analysis as in legislative assemblies the elected representatives react to current events and try to design future scenarios. Discussions about new regulations, benefits and taxes, public expenditure or debt, incentives towards certain technologies, and job market rules are only a few examples of what may condition the expectations of firms, consumers, employees, their preferences and choices, hence the entire economy. Purely quantitative measures like the national accounts statistics cannot incorporate such qualitative information. Conversely, text-based measures – specifically, textual indices constructed from politicians' speeches in Parliament – might do so.

Such research hypothesis is corroborated by numerous pieces of evidence. In fact, in recent literature, the interest in *text-as-data* among scholars of political science, economics and finance is dramatically increasing (Gentzkow, Kelly, and Taddy, 2019). Several empirical studies experiment with statistical methods from computational linguistics previously employed only in non-economic applications (e.g. spam detection, authorship attribution, text forecasting, spelling mistakes recognition and correction). The reason for this lies in the advantages of text analysis in terms of costs (Shapiro, Sudhof, and Wilson, 2020) and timeliness (Thorsrud, 2020). Digital texts act as a rich complement to usual structured, quantitative data as they provide more qualitative information concerning all kinds of human interaction, communication and culture (Gentzkow, Kelly, and Taddy, 2019).

For instance, text analysis is used in finance to predict asset prices (Tetlock, 2007), quantify financial news impacts on volatility (Manela and Moreira, 2017) or to improve default forecasts (Cerchiello and Scaramozzino, 2020). Recently, textual analysis has been used to evaluate the financial impact of the Covid-19 pandemic (Ahelegbey, Cerchiello, and Scaramozzino, 2022). Loughran and McDonald (2016) provide an extensive survey of applications and methods in accounting and finance. In macroeconomics, texts improve forecasts of inflation or unemployment while, in industrial organisation and marketing, textual information helps understand consumer decisions and, in political economy, text analysis focuses on political slant and agendas (Gentzkow, Kelly, and Taddy, 2019). Furthermore, also central banks are interested in the potential of textual data (Angelico et al., 2022; Aprigliano et al., 2022; Bholat et al., 2015), while central banks' communication itself is an object of analysis too (Hansen and McMahon, 2016; Hansen, McMahon, and Prat, 2017). Finally, in the law field, some authors proposed indicators to evaluate the impacts of a Country's legislation complexity on economic activity (Bholat et al., 2015).

But, what exactly is *textual analysis* in the first place? To answer this question, it is helpful recalling one of the most frequent terms in the field is *Natural Language Processing* (NLP). It refers to the sub-field of linguistics, computer science and artificial intelligence dealing with how computers interact with human language. The purpose of NLP is to extract meaningful information from texts - i.e. "*figuring out who did what to whom, when, where, how and why*" (Kao and Poteet, 2007, p.1). Natural language understanding aside, NLP also involves speech recognition and natural language generation. To address such challenges, NLP applies linguistic concepts as grammatical structure, rules and part-of-speech or lexicons of words and so forth (Kao and Poteet, 2007). NLP techniques are often used in *text analysis* or *text mining*. These two terms are used as synonyms and refer to the content analysis techniques which support the process to construct valid inferences from texts (Krippendorff, 2018; Stieglitz et al., 2014). More explicitly, such techniques allow "*the discovery and extraction of interesting, non-trivial knowledge from free or unstructured text*" (Kao and Poteet, 2007, p.1). Text mining is a type of data mining and typically deals with text classification (Sebastiani, 2002), clustering and association (Jo, 2019).

As noticed by Gentzkow, Kelly, and Taddy (2019), from a statistical point of view, textual data are challenging because of their intrinsically high dimensionality, which is also one of the aspects why texts are different from other types of data used in economics. Therefore, statistical and machine learning techniques are the usual methods employed in the field. Moreover, due to their recent advances, deep learning techniques – specifically, neural networks – have become prominent in many textual analysis tasks. For instance, as recalled by Algaba et al. (2020), for "*word embedding*" – i.e. the mapping of words or entire phrases of non-annotated corpora to numerical vectors preserving semantic and syntactic structure – the popular *Bidirectional Encoder Representations from Transformers* (BERT) introduced by Devlin et al. (2019) and its variants outperformed the earliest algorithms like the *Word2Vec* of Mikolov et al. (2013) or the *GloVe* of Pennington, Socher, and Manning (2014). Specifically, neural networks are attractive for acting as "*universal approximators*" thanks to "*their ability to mimic general, smooth nonlinear associations*" (Gentzkow, Kelly, and Taddy, 2019). More in detail, *Deep Neural Networks* (DNN) are easy to implement than other deep learning models and require less training time, *Convolutional Neural Networks* (CNN) provide accuracy and fast training, *Recurrent Neural Networks* (RNN) can capture sequential data for classification. However, such architectures often suffer from over-fitting, difficulties in tuning and require a large amount of data to be effectively trained. Moreover, even when training is relatively faster they are very complex models resulting essentially in "*black boxes*" (Birjali, Kasri, and Beni-Hssane, 2021).

On the other hand, statistical learning approaches, although less precise (Birjali, Kasri, and Beni-Hssane, 2021), can provide more interpretable outputs while being less computationally expensive. For this reason such approaches were preferred. More in detail, the techniques used in this thesis fall in two important sub-fields of text analysis: topic modelling and sentiment analysis.

Topic modelling refers to a set of probabilistic models which use Bayesian inference to extract latent variables from (usually) large sets of data. As recalled by Vayansky and Kumar (2020), such models are "*particularly well suited for use with text data*" as these are intrinsically high dimensional and unstructured data. Despite their use in other fields such as bioinformatics, social media analysis, environmental analysis and computer vision, topic models originated in text analysis to extrapolate the latent semantic structure of text collections. Chapter 2 provides a review of the main topic modelling methods and literature.

Sentiment analysis, also called *opinion mining* (on this, see Pang and Lee, 2008), is the sub-field of text mining concerned with the task of discovering and summarising people's opinions, attitudes and emotions towards events, topics or issues. The term, actually, also refers to the more specific task of determining the *sentiment polarity* of a text usually identifying *positive* and *negative* terms (sometimes there are also a *neutral*, or more, polarity categories). In their survey, Yadollahi, Shahraki, and Zaiane (2017) list different sentiment analysis tasks and distinguish between opinion-mining and emotion-mining. According to the authors, "*opinion polarity classification is a sub-task of opinion mining, where opinion mining, in turn, is a sub-task* *of sentiment analysis*" (Yadollahi, Shahraki, and Zaiane, 2017). Chapter 3 provides a review of the main sentiment analysis methods and literature.

Chapter 2 and Chapter 3 are the main two parts of this thesis and constitute two separate studies. Indeed, the thesis is designed and organised as a collection of two autonomous papers.

Both chapters analyse the texts from the verbatim reports of the Italian Senate of the Republic, one of the two Houses of the Italian Parliament. The procedure described is entirely replicable for any language and Country. The choice to investigate the Italian case is motivated by the following considerations.

First, this Country presents a rather peculiar and complex political system which often makes it difficult for Governments to last longer than two years on average, likely resulting in uncertainty about economic policies. For instance, from 1948 to the current day, 67 Governments succeeded one another. In several cases, when no alternative majorities existed, the Legislature ended before the term of 5 years prescribed in the Constitution, resulting in new elections (see Table 3.1).

Second, although applying text analysis to economics and politics is becoming a frequent practice, relatively a few studies employing texts in languages other than English have been conducted thus far. Abercrombie and Batista-Navarro (2020), whose focus is on political science applications, made a list of works using parliamentary transcripts in languages different from English. Among the works based on Italian texts, recent studies include: Angelico et al. (2022) and Cerchiello and Giudici (2016) who analysed twits to predict inflation and assess systemic risk in the banking sector, respectively; Aprigliano et al. (2022) who used newspapers for predicting economic activity; Catelli, Pelosi, and Esposito (2022) who made a comparison between lexicon-based and neural network-based models; Pota et al. (2021a) and Pota et al. (2021b) also who used twits for sentiment analysis. No similar studies utilising Italian parliamentary verbatim reports are available.

One significant difference between the two studies in this thesis is the sample size. Chapter 2 utilises transcripts from 1996 to 2020, while the analysis in Chapter 3 bases on the entire collection of transcripts as available at the end of 2020 – i.e. 1948–2020. Moreover, the analysis in Chapter 2 is conducted on a quarterly basis, whereas Chapter 3 develops on a yearly one. Other differences concern the methodologies and results.

More in detail, Chapter 2 aims to construct *Textual Political Debate Indices* (TPDI). As mentioned, the text analysis technique adopted in this chapter is topic modelling, which allows to estimate the major themes discussed during parliamentary sessions. Then, time series of proportions of such topics are used to explore the relationship between political discourse and a set of National Account statistics such as GDP, Government Expenditure, Investments and so forth. Moreover, topic proportions constitute the core of the proposed indices in the fashion of Larsen and Thorsrud (2019) – i.e. using autoregressive models with exogenous variables to construct them. The choice to analyse texts through topic models is motivated by the following considerations. First, as usual with unsupervised learning methods, they allow the discovery of themes from texts and are well suited when there is no prior knowledge of what the text is about – i.e. there are no labelled data already available to be used for training as in supervised methods. On the other hand, the output provided is interpretable and can be tuned. In addition, the chapter presents a sensitivity analysis of different parameters. Hence, the contribution in Chapter 2 is that, first, as assumed, a relationship between the content of the parliamentary discussion and the Country's economy exists and, second, that it can be quantified and used in time series models.

In Chapter 3 the focus of the analysis shifts from the content of political debates to their tone or sentiment. A first contribution of the study is the proposal of a new method to compute words polarities which does not require the use of predefined affective lexicons – i.e. lists of words annotated with their positive or negative sentiment – nor deep learning algorithms which would be computationally expensive and not as effective as for English or Chinese. This is since languages like Italian are classified as "*low-resource languages*" – i.e. languages with no large data-sets available to train the models beforehand (Catelli, Pelosi, and Esposito, 2022). The result of the procedure are several *Textual Political Polarity Indices* (TPPI). Some of these indices aim to capture the tone of debates. Some other indices, instead, aim to capture another appealing feature of parliamentary discussion: the tone divergence, which can be assumed to be a proxy of the agreement or disagreement between political parties. The indices are called *polarity indices* because what the proposed procedure estimates is not the sentiment as it is usually defined – i.e. the emotion associated with a specific word or phrase. The philosophy behind these indices is different: the words used – specifically, their frequencies – and the economic variable at hand are seen as a realisation of a unique probabilistic mechanism which generates both. In this view, the polarity of each word is related not only to its meaning in the language analysed but also to the economic variable it is associated with. More concretely, the idea underlying the proposed procedure is that if some words tend to co-occur with positive or negative economic outcomes, they can be assumed to incorporate some positive or negative information: the word polarity. Such an assumption, which may seem at least naive, appears to be confirmed by reached results. Even after a long time, words polarity in the explained sense does maintain a certain explanatory power on the considered economic variable. Hence, the key contributions of Chapter 3 are the proposal of a new method for polarity computation, and the demonstration that, especially political tones divergence does have a statistically significant impact on economic dynamics.

Finally, Chapter 4 ends the thesis by presenting the main results and conclusions of the two previous chapters. It focuses on the possible drawbacks of the methodologies adopted and proposes possible remedies depicting possible future developments.

Chapter 2

TPDI: Textual Political Debate Indices

2.1 Topic modelling in economics

This section provides a brief overview of research applying topic modelling in the economic-related literature without claiming to be a complete survey as the publishing of new works on the subject is constantly growing.

Before describing its applications in economics and finance, topic modelling needs to be first defined.

Topic models are "*probabilistic generative models*" (Liu et al., 2016) which fall in the vast area of machine learning algorithms and whose aim is the automatic discovery of the themes – the topics, indeed – present in a text collection through "*a hierarchical Bayesian analysis*" (Blei and Lafferty, 2009, p. 72). Specifically, without any prior information about the themes, through these algorithms, documents are modelled "*as arising from multiple topics, where a topic is defined to be a distribution over a fixed vocabulary of terms*" (Blei and Lafferty, 2009, p. 73) under the statistical assumption that a certain number of topics is "*associated with a collection, and that each document exhibits these topics with different proportions*" (Blei and Lafferty, 2009, p. 73). Hence, in a topic model, the input data are the observed documents' words, while the topics themselves and their within documents proportions are "*the hidden variables [. . .] the latent topical structure [. . .]. Given a collection, the posterior distribution of the hidden variables given the observed documents determines a hidden topical decomposition of the collection*" (Blei and Lafferty, 2009, p. 73). The observed documents and the hidden topic structure interact through "*the imaginary random process that is assumed to have produced the observed data*" (Blei and Lafferty, 2009, p. 73).

Topic models originated as a methodology for dimensionality reduction in the field of information retrieval (Vayansky and Kumar, 2020). The first probabilistic – therefore "*authentic*" (Liu et al., 2016) – topic model was the *Probabilistic Latent Semantic Analysis* (PLSA) developed by Hofmann (2001). The technique actually represented an enhancement of the *Latent Semantic Indexing* (LSI) proposed by Deerwester et al. (1990). Although not being a probabilistic model, LSI succeeded previous methods based on the *tf-idf* reduction scheme (Vayansky and Kumar, 2020) of *term-document matrices* [1]. In fact, LSI utilises singular value decomposition (SVD) to factorise such (*tf-idf* reduced) matrices [2]. However, it was only with the *Latent Dirichlet Allocation* (LDA) of Blei, Ng, and Jordan (2003) that topic models became "*even more complete*" (Liu et al., 2016) and famous among researchers. The majority of topic models available build on LDA. Among the most relevant developments, there is the *Correlated Topic Model* (CTM) of Blei and Lafferty (2007) who sample the topic mixture from a logistic-normal instead of a Dirichlet distribution, hence allowing for correlation among topics with the result of a better fitting. CTM aside, many other topic models variations have been proposed as, for instance, the *Structural Topic Model* (STM) of Roberts et al. (2014). See Liu et al. (2016) and Vayansky and Kumar (2020) for extensive and detailed surveys. Besides, many works deal with topic models evaluation such as Airoldi and Bischof (2016), among others.

In the economic field, thus far topic models have been applied in the following common areas:

- analysis of the evolution of the economic literature over time;
- prediction of stock prices, returns and their volatility;
- analysis of the effects of central banks communication;
- development of indices reflecting various aspects of the economy.

These four common areas are not the only ones, and they are not rigid categories. Many works study more subjects or perform different tasks making overlapping quite common.

Concerning the first point, recent examples are the works of Lüdering and Winker (2016) and Wehrheim (2019). Specifically, the former work investigates whether the economic literature can anticipate new trends in the economy or merely looks at it from an ex-post perspective, while the latter analyses articles from the Journal of Economic History between 1941 and 2016. These authors also provide overviews of topic modelling in the field.

The second of the areas listed above is certainly one of the most productive. Among other works, recently, topic modelling has been used by Cerchiello and Nicola (2018) to assess news contagion in finance. Besides, Larsen and Thorsrud (2017) and Adämmer and Schüssler (2020) applied the technique to study the effects of media or news, respectively, on asset returns or equity premium, while Li (2020) did the same for commodity prices and Lüdering and Tillmann (2020) analysed the effects of monetary policy on asset prices.

Regarding the third point, among the first and most relevant studies on central banks communication, there are those of Hansen and McMahon (2016) and Hansen,

[1] A term-document matrix is the matrix representation of words (rows) and documents (columns) of the analysed texts.

[2] Succeeding topic models typically use the transpose of such matrix – i.e. a *document-term* matrix – to represent texts as will be shown in the following.

McMahon, and Prat (2017) who analyse "*information released by the FOMC*" (Federal Open Market Committee) "*on the state of economic conditions, as well as the guidance the FOMC provides about future monetary policy decisions*". Besides, recently, Baerg and Lowe (2020) used topic modelling combined with "*scaling methods to estimate central bank preferences*". Instead, Carboni, Farina, and Previati (2020) "*investigate the differences in the topics contained in European Central Bank (ECB) and Federal Reserve (FED) Governors' speeches over the period 2007–2019*".

Finally, using topic modelling to construct text-based economic measures (which is also the aim of this work, as described in the next sections) finds precedents in many works: Dybowski and Adämmer (2018) who focus on presidential tax communication; Larsen and Thorsrud (2019) and Thorsrud (2020) who build coincident indices of economic indicators (GDP, government expenditure and others), and business cycle, respectively; Angelico et al. (2022) and Larsen, Thorsrud, and Zhulanova (2020) who study inflation expectations.

2.2 Data description

This section describes both types of data employed to build the proposed *Textual Political Debate Indices* (*TPDI*): transcripts of Italian parliamentary debates and quarterly time-series of Italian economic indicators.

In this work, time is measured in quarters. Specifically, $t = 1, \ldots, T = 98$ is the sample period and corresponds to calendar quarters 1996.Q2–2020.Q3.

2.2.1 The Italian Senate verbatim reports corpus

The corpus analysed is the collection of the *Italian Senate of the Republic* parliamentary verbatim reports. It consists of more than 4,300 transcripts from 09 May 1996 to 08 September 2020 which are converted from raw *.pdf* format into a *Document Term Matrix* (DTM).

The about 24 years period considered spans 6 Legislatures – starting from the XIII to the (still ongoing) XVIII – and 14 Governments.

The *Senate of the Republic* – shortly, the Senate – is one of the two Houses of the Italian Parliament. As recalled on the Senate website (Senato della Repubblica, n.d.), "*according to the principle of full bicameralism, the two houses perform identical functions. [...] Both Houses are elected every five years. The only differences between them lie in the membership and the rules for the election of their members. The [...] deputies, who must be at least 25 years of age, are elected by all Italian citizens over 18 years of age. The [...] elected senators must be at least 40 years of age, and their electors must be over 25. In addition to elected members, the Senate also includes life senators – who are appointed by the President of the Republic "for outstanding merits in the social, scientific, artistic or literary field" – and the former Presidents of the Republic, who are ex officio life senators*". Moreover, by the Italian Constitution, senators are half the number of deputies (the members of the other House) and are elected on a regional basis.

Because of these peculiarities, Senate reports are preferable over those of the other House. They convey the same information about the parliamentary debate with relatively fewer records. Hence they are likely to be less noisy and require less computational effort to process.

Coming to more descriptive information about the texts, these are freely accessible in *.pdf* format via the Senate website as, usually, Parliament sittings are open to the public.

Further information – such as Legislatures' exact periods or the Government in charge at each time – is available in Table 2.1.

In the Italian system, a Prime Minister no longer supported by the majority coalition can be ousted with a vote of no confidence. If this is the case, the President of the Republic either appoints a new Prime Minister capable of forming a Government supported by the Parliament, or, dissolves (even just one of) the two Houses calling for elections. For this reason, as visible in Table 2.1, even more than one Government, also supported different Majorities, may happen to be in charge during the same Legislature period, a characteristic which makes the Italian case an appealing to analyse compared to other Countries.

2.2.2 The economic variables

Similarly to Larsen and Thorsrud (2019), to evaluate the eventual link between parliamentary debate and economy dynamics, a set of economic variables (or aggregates) is used in addition to texts. Furthermore, such measures allow the comparison of conventional indices against the proposed text-based ones.

Therefore, time series of the *output* (Y), *wages* (W), *consumption* (C), *investments* (I), *government expenditure* (G), *taxation* (T), *exports* (X) and *imports* (M) are retrieved from the Italian National Institute of Statistics (ISTAT) national accounts statistics.

Specifically, Y is defined as the *gross domestic product at market prices*; W is the *domestic compensation of employees*; C is defined as the *final consumption expenditure of households and non-profit institutions serving households*; I is defined as the *gross fixed capital formation*; G as the *consumption of general government*; T is defined as the *taxes less subsidies on production and imports* X and M are defined as *exports/imports of goods (FOB) and services* [3].

All aggregates are measured in *current prices* millions of euro and span the same period as the texts: 1996–2020. Moreover, all the time series are quarterly based and seasonally adjusted from the source via the Tramo-Seats procedure. As specified by ISTAT on its website, "*seasonally adjusted data are to be intended as seasonal and calendar-adjusted wherever calendar effects are present*" (Istat, n.d.).

[3] In international commercial law, *FOB* stands for *free on board*. It means that the price of the goods is not comprehensive of any other costs or expenses needed to make them arrive at destination

TABLE 2.1: Italian Senate verbatim reports details

Legislatures	**Election Dates**	**Legislature Dates**	**Legislature Duration (Days)**	**Governments (Prime Ministers)**	**Government Dates**	**Senate Sittings (or Reports)**
XVIII	4 Mar 2018	23 Mar 2018, ongoing [8 Sept 2020]	[900]	Conte II	5 Sept 2019, ongoing	[254]
				Conte	1 June2018, 5 Sept 2019	
XVII	24-25 Febr 2013	15 Mar 2013, 22 Mar 2018	1834	Gentiloni	12 Dec 2016, 1 June2018	923
				Renzi	22 Feb 2014, 12 Dec 2016	
				Letta	28 Apr 2013, 22 Feb 2014	
XVI	13-14 Apr 2008	29 Apr 2008, 14 Mar 2013	1781	Monti	16 Nov 2011, 28 Apr 2013	860
				Berlusconi IV	8 May 2008, 16 Nov 2011	
XV	9-10 Apr 2006	28 Apr 2006, 28 Apr 2008	732	Prodi II	17 May 2006, 8 May 2008	283
XIV	13 May2001	30 May 2001, 27 Apr 2006	1794	Berlusconi III	23 Apr 2005, 17 May 2006	965
				Berlusconi II	11 June2001, 23 Apr 2005	
XIII	21 Apr 1996	9 May 1996, 29 May 2001	1847	Amato II	26 Apr 2000, 11 June 2001	1061
				D'Alema II	22 Dec 1999, 26 Apr 2000	
				D'Alema	21 Oct 1998, 22 Dec 1999	
				Prodi	18 May 1996, 21 Oct 1998	
			Total			**Total**
			8888 ($\approx$ 24 years)			4346

2.3 Text processing

2.3.1 Cleaning

Estimating topic models requires a *document-term matrix* (DTM) as input – i.e. a corpus representation where documents are arranged on rows and terms on columns while the inner elements of the resulting matrix are the frequencies each word appears in a specific document.

The process to obtain such DTM for the chosen corpus begins with a preliminary inspection of the Senate reports understanding their general structure and peculiarities (e.g. the presence of regular patterns in the text – also called *regex(es)* in NLP). Then, after checking for *.pdf* files stored as images, optical character recognition (OCR) was performed when needed.

Therefore, all reports were converted from *.pdf* to *.txt* format and standard pre-processing and cleaning (Banks et al., 2018; Benoit et al., 2018; Denny and Spirling, 2018; Welbers, Atteveldt, and Benoit, 2017) were applied. Text pre-processing and cleaning are crucial in NLP as they allow data to be more tractable and reduce noise. After removing the front-page, index, summaries and attachments (usually the bills presented and discussed during the session), by mean of *regular expressions* (REGEX) [4], all texts are associated with the sitting day and then split into orators' speeches. The following additional steps allowed to obtain more "cleaned" texts:

- the *removal of names and surnames*, as usually – in a Parliamentary sitting – these elements appear when the chairing President makes the roll call during nominal votes;
- the *removal of Roman numerals* as well as *session timing and hours indication*;
- the *removal of the content in parenthesis* as, in the context at hand, it usually describes actual circumstances during the session (people shouting or clapping hands, the President calling the Assembly to order, and similar).

The resulting (partially cleaned) speeches form the rows of a data-frame composed by a column with the pre-processed texts and other columns containing the *meta-data* associated with every single speech – namely the indication of date, legislature, Government, session and speaker name (and the number of single words per speech).

After completing pre-processing, the data frame counted about 850,000 raw speeches. Although these texts are not properly raw because of the removal of some words, a deeper cleaning is required to guarantee more efficiency during the topic-modelling phase (Banks et al., 2018; Benoit et al., 2018; Denny and Spirling, 2018; Welbers, Atteveldt, and Benoit, 2017).

Therefore, the following steps were taken:

- the *removal of numbers, punctuation and isolated characters*;
- the *lower-casing* of the entire corpus;
- the *removal of sequences of the same character* (e.g. aaa, bb, and similar);
- the *removal of words with less than 3 characters*, this to speed up the following stemming step and to exclude terms so short that after such procedure would become not recognisable or intelligible anymore.

[4] In the original reports, each front page shows the sitting day, while the orator's surname (or surname and name in case of homonyms) in capital letters precedes every speech. Orators' names are double-checked with a list of all senators retrieved from the Senate website, which is extended by adding the names and surnames of all governments members (President, ministers, . . .) as also they can speak during sessions.

As a result, the length of each speech slightly reduces, and some very brief texts completely disappear as their number of words drops down to zero. At the end of the stages so far described, the total number of cleaned speeches in the data frame was about 770,000. Another considerable action taken in the cleaning phase was the *stop-words removal* Grimmer, 2010; Quinn et al., 2010 – i.e. the exclusion from the analysis of those words which do not add actual content to the texts, for instance, articles, pronouns, conjunctions, adverbs and similar. The complete list of removed stopwords is available in Appendix A. Consequently, the number of words per speech slightly decreases while the total number of cleaned speeches in the data frame reduces to around 760,000.

Finally, in order to decrease the vocabulary size, all the words in each text are reduced to their "base (or root) form" through *stemming* [5]. Being the verbatim reports at hand in Italian, the applied algorithm for stemming is the *Snowball algorithm* (see Porter (2001) and *Snowball* n.d.) that also supports this language among others.

In conclusion, the corpus created and used in the analysis consists of around 760,000 cleaned and stemmed speeches of different lengths, though each composed of words – better, stems – whose minimum length is 3 characters.

2.3.2 Document-Term Matrices

As previously mentioned, estimating topic models requires the corpus to be rearranged in a DTM. However, the content of such matrices, hence their dimensions, can be defined in several ways.

About its content, a DTM might contain simple counts or relative frequencies of a specific word in a determined document. It is also possible to have the counts weighted according to some function (weighted DTM). One of the most interesting and widely adopted scheme to do that is the *term frequency-inverse document frequency* weighting, shortly *tf-idf* (see Blei and Lafferty (2009), Jurafsky and Martin (2020), and Manning, Raghavan, and Schütze (2008) among others). According to this scheme, each word receives a weight which – via the *term frequency* – increases proportionally to the number of times the word appears in the document, but – via the *inverse document frequency* – decreases as the number of times the word appears in the entire collection increases. Therefore, the *tf-idf* tends to allocate more weight to those words that appear more often in the document but are relatively less frequent in the corpus overall – i.e. words with greater discriminative power across documents.

About DTM dimensions, also such choice is subjective and analysis dependent. Removing or aggregating words into multi-word expressions affects the matrix number of columns – i.e. the total number of words composing the vocabulary (the *vocabulary size*). On the other hand, removing or aggregating texts affects the DTM number of rows – i.e. the actual number of documents in the corpus (the *corpus size*).

[5] For instance, stemming would map the English words *connection, connections, connective, connected* and *connecting* to the same form *connect* (*Snowball* n.d.).

TABLE 2.2: Reduced vocabulary DTM information details

Document-Term Matrices Statistics			**Terms Counts Statistics**											
Dimensions		**Sparsity**	**Across Corpus**						**Across Vocabulary**					
Corpus size	*Vocabulary size*	*(%)*	*Min*	*25%*	*Median*	*Mean*	*75%*	*Max*	*Min*	*25%*	*Median*	*Mean*	*75%*	*Max*
762,928	74,050 (100%)	99.96	1	3	8	48.10	27	14,604	1	1	3	495.60	17	507,668
762,928	59,240 (80%)	99.94	1	3	8	48.08	27	14,596	1	2	5	619.20	31	507,668
762,928	44,418 (60%)	99.93	1	3	8	48.05	27	14,593	1	3	10	825.30	65	507,668
762,928	29,620 (40%)	99.89	1	3	8	47.98	27	14,584	1	10	31	1,236.00	174	507,668
762,928	14,810 (20%)	99.78	1	3	8	47.68	27	14,513	1	63	174	2,456.00	732	507,668

Here, addressing both issues serves two purposes: increasing efficiency and speeding up estimation; allowing for sensitivity analysis in the topic modelling phase.

In detail, a total of 5 *CTM* are estimated, one for each of the DTM constructed by keeping constant the number of rows (corpus size) while letting the number of columns (vocabulary size) vary.

More in detail, the original (100% vocabulary) of about $760,000 \times 74,000$ DTM is *tf-idf* weighted, and its columns reduced so to show only the top 80th to 20th words percentiles – i.e. the top 80 to 20 per cent words in terms of their *tf-idf* weight. As reported in Table 2.2, this does not really affect the values of the DTM *sparsity* (i.e. the "emptiness", or the number of zeroes in the matrix) and "across corpus" descriptive statistics. However, the operations described impact the word counts statistics ("across vocabulary" statistics). The average number of times each word appeared in the entire corpus increases from a value of about 500 for the complete matrix to an average of around 2,500 if considering the top *tf-idf* 20% DTM, as the average document length decreases because of the vocabulary size reduction. Finally, the 20% jumps for the thresholds (i.e. $20, 40, 60, 80, 100$) are arbitrarily chosen as – although reducing the vocabulary size according to the *tf-idf* is common in practice (see, for instance, Blei and Lafferty (2009) and Larsen and Thorsrud (2019) among others) – there are no fix rules for choosing "cutting points".

Before estimating the CTM, also DTM rows need some manipulation. Very brief speeches delivered by participants during the Assembly are removed from the original corpus (better, the five corpora) and therefore disregarded to reduce noise in the data and increase efficiency during the estimation procedure. Similarly to the vocabulary thresholds, also the minimum 10 words value for the document length is chosen arbitrarily, considering that the median value of the number of words per speech was equal to 8 (see Table 2.2). The main goal here is disregarding very brief, uninformative "documents" (on the point, see also Adämmer and Schüssler, 2020, p. 5 who reason in terms of upper/lower quantiles instead). In doing so, as reported in Table 2.3, not only the number of texts (corpus size) decreases from about 760,000 to about 360,000 (357,582 – or, 357,430 for the top *tf-idf* 20% DTM), but also the vocabulary size slightly decreases if compared with that in Table 2.2: some words (actually very common/frequent ones) only appear in the removed speeches.

Finally, in each of the five (cleaned, stemmed and reduced) DTM, texts are aggregated daily – i.e. the speeches delivered in the same Senate session (or, equivalently,

TABLE 2.3: Reduced corpus DTM information details

Document-Term Matrices Statistics			Terms Counts Statistics												
Dimensions		Sparsity	Across Corpus						Across Vocabulary						
Corpus size	*Vocabulary size*	*(%)*	*Min*	*25%*	*Median*	*Mean*	*75%*	*Max*	*Min*	*25%*	*Median*	*Mean*	*75%*	*Max*
357,582	73,638 (100%)	99.91	10	19	31	98.40	97	14,604	1	1	3	477.8	17	459,095
357,582	58,828 (80%)	99.88	10	19	31	98.36	97	14,596	1	2	5	597.9	31	459,095
357,582	44,006 (60%)	99.85	10	19	31	98.29	97	14,593	1	3	11	798.7	66	459,095
357,582	29,208 (40%)	99.78	10	19	31	98.14	97	14,584	1	10	32	1,202.0	178	459,095
357,430	14,557 (20%)	99.55	10	19	31	97.53	97	14,513	1	65	179	2,395.0	749	459,047

TABLE 2.4: Aggregated corpus DTM information details

Document-Term Matrices Statistics			Terms 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	477.8	17	459,095
2,858	58,828	96.23	38	7,388	11,816	12,306	16,803	68,531	1	2	5	597.9	31	459,095
2,858	44,006	94.98	38	7,387	11,802	12,298	16,787	68,524	1	3	11	798.7	66	459,095
2,858	29,208	92.49	38	7,376	11,779	12,279	16,748	68,496	1	10	32	1,202.0	178	459,095
2,858	14,557	85.35	38	7,331	11,686	12,198	16,654	68,304	1	65	179	2,395.0	749	459,047

the same day) are combined to form single rows in the document-word matrix. The resulting five DTM were those employed for the topic model estimation. They were all composed of about 2,800 rows. See Table 2.4 for further details.

As a concluding remark, handling the corpus and creating the DTM is implemented in the statistical software R (R Core Team, 2020) also by mean of the library *quanteda* (Benoit et al., 2018).

2.4 Topic modelling

2.4.1 The Correlated Topic Model

As previously mentioned, the "*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. For this reason, to extract the themes discussed in the Senate sessions, the Correlated Topic Model of Blei and Lafferty (2007) is used.

The consideration of Vayansky and Kumar (2020) that CTM "*would work well with sets which are expected to have strongly correlating topics, such as the [...] articles within a single journal*" seems to support the choice for such model here. Similarly to a journal collection, all analysed texts come from documents belonging to the same corpus of parliamentary debate sessions.

In the following, the notation and a brief description of the model are detailed.

Figure 2.1 shows the *plate notation* of a CTM model. All nodes represent random variables that may be observable (shaded nodes) or latent (unshaded nodes). Each edge indicates a dependence, while plates represent the replicated variables.

Let V be the *vocabulary size* – i.e. the total number of words in the texts collection – and let K be the (user) specified *number of topics* in the model. Each V-dimensional

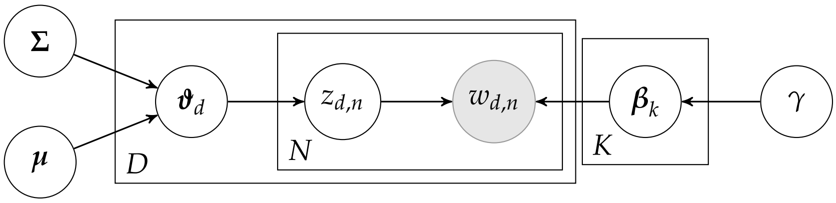


FIGURE 2.1: Plate Diagram of CTM based on Blei and Lafferty (2007)

vector $\boldsymbol{\beta}_k$, where $k \in \{1,\dots,K\}$, is a *topic* – i.e. a distribution over the entire vocabulary. Specifically, $\boldsymbol{\beta}_k \sim Dir_V(\gamma)$, where $Dir_V(\gamma)$ is a Dirichlet distribution with parameter γ.

Let D be the *corpus size* – i.e. the total number of documents in the collection of texts – and let N_d be the the number of words in each document $d \in \{1,\dots,D\}$. Each word $w_{d,n}$, where $n \in \{1,\dots,N_d\}$, is the n-th word in the d-th document and it is assumed to be drawn from one of the K topics $\boldsymbol{\beta}_k$. Specifically, $w_{d,n} \sim Mult_V(\boldsymbol{\beta}_{z_{d,n}})$, where $Mult_V(\boldsymbol{\beta}_{z_{d,n}})$ is a Multinomial distribution with parameter $\boldsymbol{\beta}_{z_{d,n}}$. The integer $z_{d,n} \in \{1,\dots,K\}$ is the *topic assignment* of $w_{d,n}$ – i.e. the identity of the topic of the n-th word in the d-th document. Each topic assignment is assumed to come from a Multinomial distribution of parameter $\boldsymbol{\vartheta}_d$ - i.e. $z_{d,n} \sim Mult_K(\boldsymbol{\vartheta}_d)$. Each $\boldsymbol{\vartheta}_d$ is a distribution over topics and is obtained as a function of $\boldsymbol{\eta}_d$, which is the (typically considered) natural parameterization (Blei and Lafferty, 2007) of the distribution $Mult_K()$. In formulae, $\vartheta_i = f(\eta_i) = \frac{\exp \eta_i}{\sum_j \exp \eta_j}$ is the function that maps the natural parametrization $\boldsymbol{\eta}_d$ of topic proportions to the mean parameterization $\boldsymbol{\vartheta}_d$ (Blei and Lafferty, 2007). Real values $\boldsymbol{\eta}_d$ are drawn from a multivariate Normal distribution – i.e. $\boldsymbol{\eta}_d \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ – whose parameters are the K-dimensional mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

The following steps describe the generative, random process underlying the documents and words:

1. For each $k \in \{1,\dots,K\}$:
   (a) Draw $\boldsymbol{\beta}_k \sim Dir_V(\gamma)$.
2. For each $d \in \{1,\dots,D\}$:
   (a) Draw $\boldsymbol{\eta}_d | \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$;
   (b) For $n \in \{1,\dots,N_d\}$:
      i. Draw a topic assignment $z_{d,n} | \boldsymbol{\eta}_d$ from $Mult_K(\boldsymbol{\vartheta}_d)$.
      ii. Draw a word $w_{d,n} | \{z_{d,n}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K\}$ from $Mult_V\left(\boldsymbol{\beta}_{z_{d,n}}\right)$.

To understand the output of a topic model is helpful to think at the per-topic term probabilities $\boldsymbol{\beta}_k$ and the per-document topic proportions $\boldsymbol{\theta}_d$ as the rows of two matrices – $\boldsymbol{B}$ and $\boldsymbol{\Theta}$ – which result from the decomposition of the original DTM. $\boldsymbol{B}$ is a topic-term matrix where each row is a distribution over words. $\boldsymbol{\Theta}$ is a document-topic matrix whose rows are distributions over topics.

As for corpus handling and DTM creation, also the correlated CTM estimation was performed using the R software (R Core Team, 2020). Specifically, applying the functions available in the library *stm* by Roberts, Stewart, and Tingley (2019).

2.4.2 Optimal number of topics

A key aspect in topic modelling is determining the (optimal) number of topics, K, to estimate. Independently on the model (here a CTM), this information is crucial as needed to initialise the algorithm, and its choice is up to the researcher (on the point, see also Grimmer and Stewart (2013) among others).

Therefore, to address the issue, the following data-driven strategy is adopted. Similarly to Adämmer and Schüssler (2020) – in order to choose a reasonable K – various CTM are run and compared through diagnostic statistics.

In detail, 16 CTM are estimated for each of the 5 DTM previously built: one per each number of topics $K = 10, 20, \ldots, 90, 100, 120, \ldots, 180, 200, 250$. The following considerations motivate the choice of such values for K in making comparisons.

First, K should be large enough to predict the topics' proportions within the corpus. Second, the estimated topics also need to be semantically coherent – i.e. intelligible and recognisable by human readers with relative ease. However, as pointed out by Chang et al. (2009), increasing K over a certain point makes the estimated topics less comprehensible and hard for humans to interpret.

Second, setting K in the interval 10–250 is consistent with the evidence in previous literature. For instance, Adämmer and Schüssler (2020) set the number of topics equal to 100 as they find that, over that level, there are no significant improvements in the diagnostic statistics. Instead, Larsen and Thorsrud (2019) set $K = 80$. Moreover, in accordance with the "*principle of parsimony*", Banks et al. (2018, p. 455) recommend starting with few topics (up to 100) and increasing K only if needed to obtain better results.

Finally, as estimation time dramatically increases with the number of topics, setting high K may prevent any use of topic models in practice (Adämmer and Schüssler, 2020, p. 6).

To compare estimated models, four diagnostic statistics are considered:

- *held-out likelihood* (Wallach et al., 2009);
- *model residuals* (Taddy, 2012);
- *semantic coherence* (Mimno et al., 2011; Roberts et al., 2014);
- *exclusivity* (Roberts et al., 2014).

Tables 2.5 and 2.6 report their values and summary, respectively. The first two statistics pertain to model accuracy, while the other two relate to topics quality.

More precisely, high values of held-out likelihood are associated with high probabilities for a model to produce correct predictions for unseen documents (the held-out documents). This diagnostic statistic is also named *document-completion held-out likelihood method* (Roberts et al., 2014; Wallach et al., 2009) and it is concerned only with models predictive power. As shown in figure 2.2, the held-out likelihood seems to become stable after $K = 180$ for all the five CTM versions.

Concerning models residuals, the lower their values and, more importantly, their (over)dispersion (Taddy, 2012) the better the model.

Notably, reaching high performances in terms of held-out likelihood and residuals usually requires high values of K. However, these are at odds with reasonable computation time and topic quality.

Semantic coherence is the statistic used to measure topic quality. Its values are high when the most probable words in the estimated topics co-occur within the documents. In such a case, the topics are considered internally consistent (Mimno et al., 2011).

Lastly, Roberts et al., 2014 propose to look at exclusivity to account for the circumstance that different topics may have words in common. The value of this measure increases if words assigned a high probability in a specific topic also receive low probability in all other topics.

There is a trade-off between the last two measures. Topics that are both semantically coherent and exclusive are also more likely to be high quality overall (Roberts et al., 2014) hence useful for the analysis. Such trade-off is clearly evident in figures 2.2 and 2.3 for the five DTM employed. Semantic coherence constantly decreases as K increases, although not so rapidly due to a change in slope around $K = 100$. Exclusivity rapidly increases with the number of topics at least up to 100. After that point, its curve becomes pretty flat at each K. Therefore, for each of the 5 DTM used as inputs, the 100 topics CTM are chosen as baseline models.

2.4.3 CTM results

As previously mentioned, topic models produce two main outputs: the per-topic term probabilities, β_k, and the per-document topic proportions, θ_d. Figures 2.4 and 2.5 show a possible representation of both outputs for topic 61 based on the 100 topic CTM on the 100% tf-idf DTM (i.e. the complete vocabulary). Topic numbering has no meaning. Instead, the choice for showing this topic is that it probably relates to public expenditure, a subject that received considerable attention recently. Finally, an overview of all estimated topics and their expected proportions in the collection is given in Figures 2.6 and 2.7.

TABLE 2.5: Correlated Topic Model(s) Diagnostic Values by Number of Topics and Corpus 2

		CTM(s) Diagnostic Values																			
		Exclusivity					*Semantic Coherence*					*Heldout Likelihood*					*Residual*				
	tf-idf Corpus Top Percentiles	100%	80%	60%	40%	20%	100%	80%	60%	40%	20%	100%	80%	60%	40%	20%	100%	80%	60%	40%	20%
Number of Topics (K)	*10*	8.974	8.948	8.857	8.729	8.669	-3.252	-3.183	-3.236	-3.840	-2.873	-7.367	-7.383	-7.370	-7.360	-7.286	5.044	4.944	4.453	3.873	3.282
	20	9.498	9.470	9.412	9.322	9.372	-6.961	-6.340	-5.819	-6.451	-6.414	-7.313	-7.342	-7.336	-7.318	-7.245	3.983	3.865	3.559	3.145	2.736
	30	9.616	9.607	9.562	9.527	9.462	-8.224	-7.845	-8.818	-9.309	-7.535	-7.290	-7.319	-7.307	-7.289	-7.225	3.484	3.367	3.136	2.775	2.448
	40	9.695	9.671	9.620	9.534	9.567	-9.260	-9.809	-10.498	-9.671	-9.603	-7.264	-7.301	-7.292	-7.269	-7.201	3.233	3.072	2.849	2.583	2.273
	50	9.718	9.698	9.646	9.596	9.537	-10.291	-10.996	-11.956	-11.702	-10.678	-7.261	-7.291	-7.278	-7.257	-7.192	2.965	2.844	2.675	2.437	2.155
	60	9.747	9.722	9.672	9.596	9.583	-11.522	-12.608	-12.365	-11.780	-12.400	-7.246	-7.279	-7.269	-7.252	-7.180	2.841	2.692	2.581	2.325	2.055
	70	9.769	9.730	9.717	9.636	9.591	-14.142	-14.079	-14.359	-14.539	-13.890	-7.244	-7.276	-7.260	-7.241	-7.172	2.744	2.562	2.462	2.223	1.988
	80	9.790	9.745	9.714	9.651	9.618	-17.110	-15.142	-15.565	-14.780	-15.625	-7.235	-7.269	-7.255	-7.234	-7.166	2.678	2.473	2.364	2.160	1.930
	90	9.803	9.756	9.714	9.655	9.618	-16.335	-15.219	-16.339	-16.230	-15.241	-7.234	-7.266	-7.252	-7.230	-7.161	2.611	2.441	2.313	2.108	1.878
	100	9.801	9.773	9.723	9.671	9.606	-16.391	-16.101	-15.715	-16.071	-15.955	-7.227	-7.257	-7.250	-7.226	-7.158	2.584	2.385	2.265	2.074	1.836
	120	9.806	9.780	9.740	9.677	9.634	-16.473	-16.289	-16.646	-15.703	-16.603	-7.222	-7.252	-7.241	-7.219	-7.151	2.571	2.360	2.214	2.014	1.775
	140	9.812	9.781	9.745	9.671	9.621	-16.478	-16.139	-16.641	-16.716	-18.158	-7.217	-7.248	-7.237	-7.214	-7.145	2.604	2.361	2.195	1.961	1.734
	160	9.823	9.784	9.739	9.665	9.607	-17.466	-17.873	-17.408	-17.428	-18.797	-7.212	-7.247	-7.234	-7.212	-7.141	2.682	2.376	2.179	1.929	1.695
	180	9.823	9.790	9.741	9.673	9.589	-17.723	-18.504	-17.206	-17.914	-18.950	-7.211	-7.243	-7.232	-7.209	-7.138	2.796	2.448	2.158	1.935	1.673
	200	9.824	9.787	9.749	9.642	9.543	-17.070	-17.095	-18.209	-17.342	-18.656	-7.209	-7.244	-7.230	-7.210	-7.137	3.000	2.521	2.177	1.918	1.658
	250	9.823	9.789	9.737	9.645	9.505	-17.107	-17.665	-17.796	-18.604	-18.674	-7.209	-7.246	-7.231	-7.208	-7.134	3.892	2.966	2.342	1.939	1.648

TABLE 2.6: Correlated Topic Model(s) Diagnostic Values Descriptive Statistics 2

		CTM Diagnostic Values Descriptive Statistics																			
		Exclusivity					*Semantic Coherence*					*Heldout Likelihood*					*Residual*				
	tf-idf Corpus Top Percentiles	100%	80%	60%	40%	20%	100%	80%	60%	40%	20%	100%	80%	60%	40%	20%	100%	80%	60%	40%	20%
Statistic	*Min*	8.974	8.948	8.857	8.729	8.669	-17.723	-18.504	-18.209	-18.604	-18.950	-7.367	-7.383	-7.370	-7.360	-7.286	2.571	2.360	2.158	1.918	1.648
	25%	9.712	9.691	9.640	9.580	9.529	-17.079	-16.491	-16.786	-16.873	-18.282	-7.262	-7.293	-7.281	-7.260	-7.195	2.661	2.427	2.209	1.955	1.724
	Median	9.795	9.750	9.715	9.644	9.586	-16.363	-15.181	-15.640	-15.242	-15.433	-7.234	-7.268	-7.254	-7.232	-7.164	2.818	2.541	2.353	2.134	1.904
	75%	9.815	9.782	9.739	9.667	9.609	-10.033	-10.699	-11.591	-11.194	-10.409	-7.216	-7.248	-7.236	-7.214	-7.144	3.295	2.992	2.718	2.474	2.185
	Max	9.824	9.790	9.749	9.677	9.634	-3.252	-3.183	-3.236	-3.840	-2.873	-7.209	-7.243	-7.230	-7.208	-7.134	5.044	4.944	4.453	3.873	3.282
	Mean	9.708	9.677	9.631	9.556	9.508	-13.488	-13.430	-13.661	-13.630	-13.753	-7.248	-7.279	-7.267	-7.247	-7.177	3.107	2.855	2.620	2.337	2.048
	St. Dev.	0.215	0.212	0.225	0.238	0.234	4.580	4.562	4.516	4.388	5.021	0.044	0.040	0.041	0.044	0.043	0.685	0.700	0.627	0.538	0.452

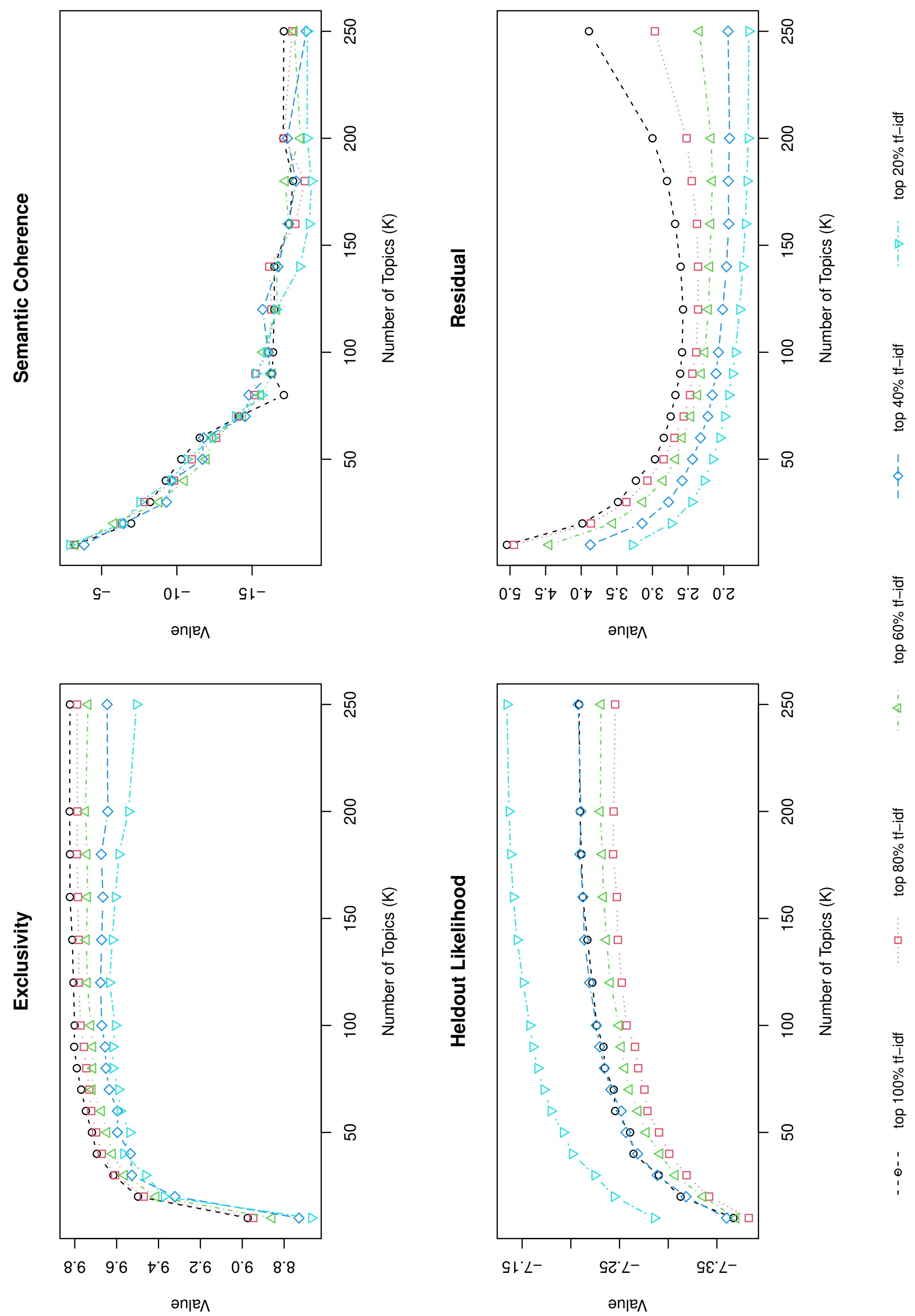


FIGURE 2.2: Correlated Topic Model Diagnostic Values by Number of Topics and Corpus 2

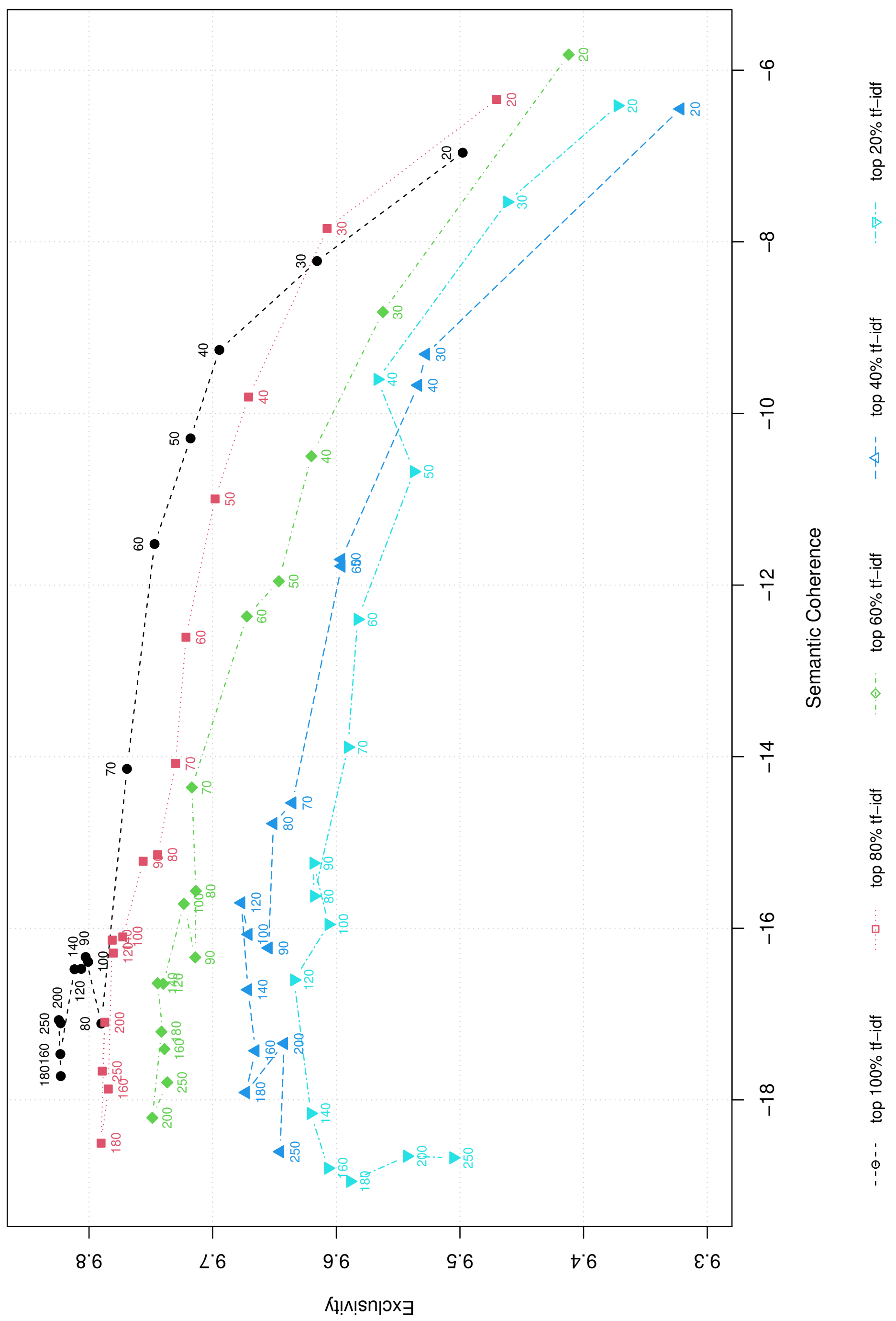


FIGURE 2.3: Correlated Topic Model Exclusivity vs Semantic Coherence by Number of Topics and Corpus

FIGURE 2.4: 100% tf-idf Corpus $K = 100$ CTM Sample Word-Cloud

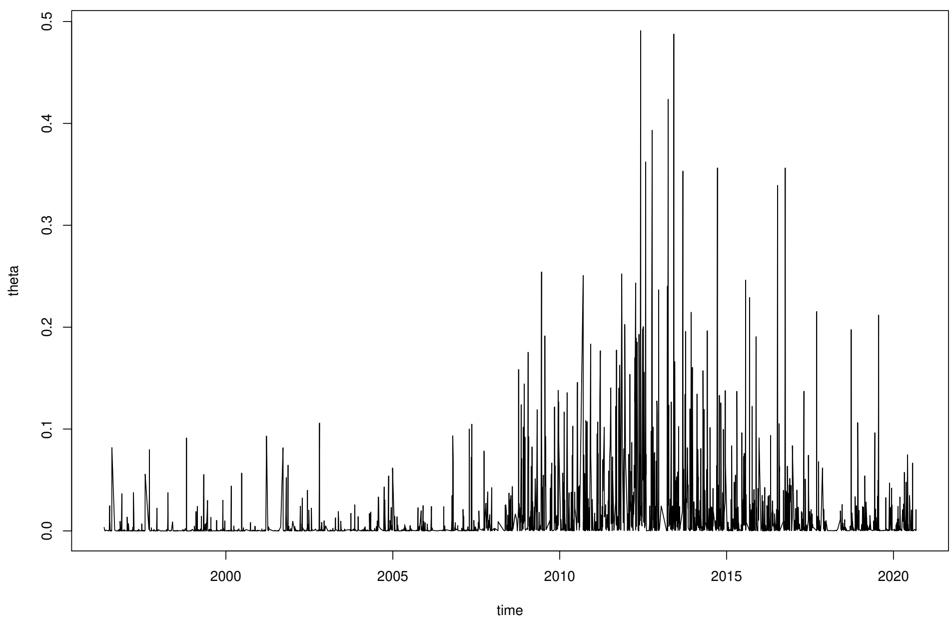


FIGURE 2.5: 100% tf-idf Corpus - $K = 100$ CTM Topic N. 61.
Daily proportions over time

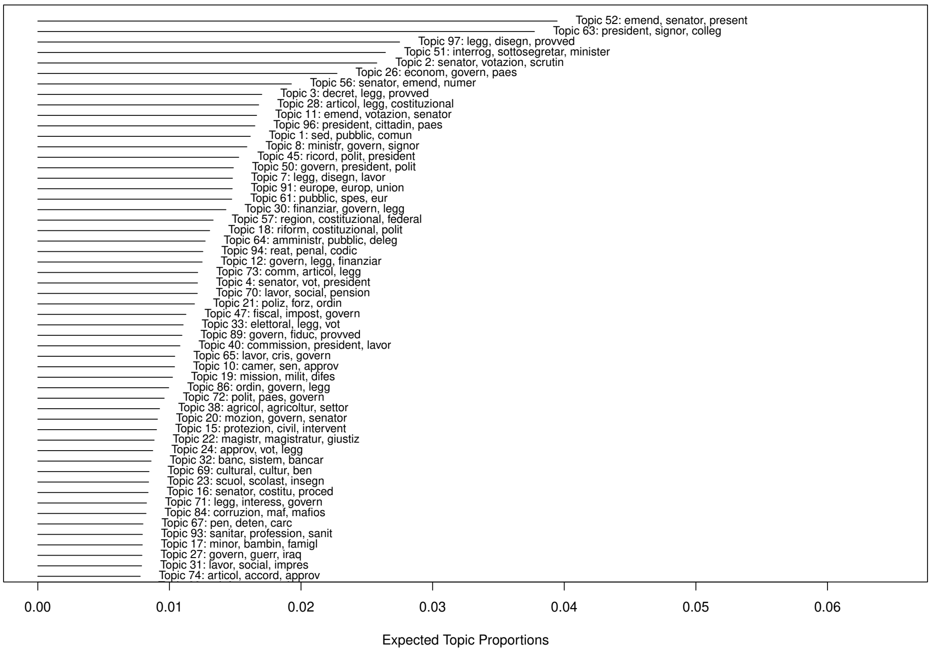


FIGURE 2.6: 100% tf-idf Corpus - $K = 100$ CTM: Expected Topic Proportions - I

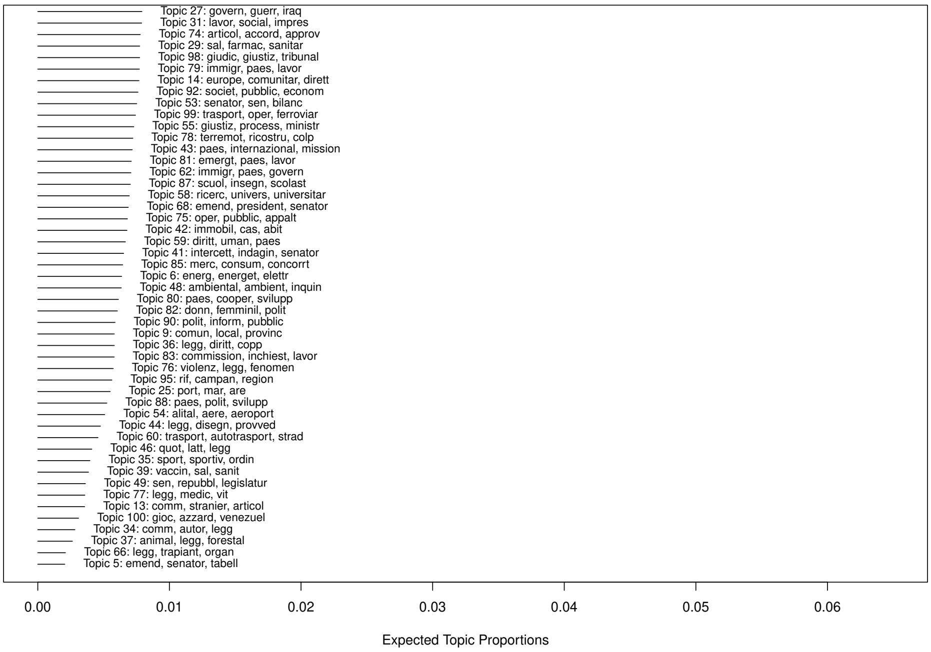


FIGURE 2.7: 100% tf-idf Corpus - $K = 100$ CTM: Expected Topic Proportions - II

2.5 Connecting themes with economic variables

This section describes the construction of the proposed text-based economic indices. Some exploratory data analysis precedes the procedure description, and it aims to corroborate the underlying hypothesis that political debate – specifically, the parliamentary one – is related to the economic dynamics. Concerning the modelling of such connection, the approach here adopted somehow resembles the one followed by Larsen and Thorsrud (2019) although with the substantial methodological difference that those authors operate within the Bayesian framework.

2.5.1 Exploring the relation between the estimated topics and the economic variables

The connection between political debate and economic dynamics is explored by looking at the correlations between the estimated topic proportions of the baseline model (i.e. the 100 topics CTM) and the time series of each considered National Account aggregate. The *topic proportions*, ϑ, are one of the outputs of the topic model. Their values lie in the continuous interval $[0,1]$ and are daily based as the CTM estimation relies on senators' speeches aggregated at a day level. To make significant comparisons, following Larsen and Thorsrud (2019), the frequency of such ϑ is aligned to that of the economic variables.

Precisely, let $t = 1, \dots, T$ be the index for each quarter, where T is the total number of quarters in the sample. Let $s_t = 1, \dots, S_t$ be the index for each day in quarter t, where S_t is the total number of days in t. Then, $\dot{\vartheta}_t = \frac{1}{s_t} \sum_{s=1}^{S_t} \vartheta_{i,s}$ with $i = 1, \dots, K$. Clearly, as all ϑ_s lie in $[0,1]$ also their quarterly averages $\dot{\vartheta}$ do so.

Let $Z \in \{Y, W, C, I, G, T, X, M\}$ be the generic economic aggregate time series and let $\dot{z} = \frac{Z - \min Z}{\max Z - \min Z}$, be the normalised series in the interval $[0,1]$ associated with it. Then, the correlation at lag $0 \leq h \leq T$ between the (normalised) economic variable $\dot{z}$ and the (quarterly) topic proportions $\dot{\vartheta}$ is indicated by $corr_{\dot{z},\dot{\vartheta}}(h)$.

Figures 2.8 to 2.10 present the correlations between each of the 100 topic proportions and the 8 economic aggregates for $h = 0, 1, 2$ to account for the circumstance that political and economic dynamics and their relation may be different in certain phases. As shown in Figures 2.11 to 2.13 which focus on the top 10% strongest correlations – on average – the link between topic proportions and economic variables is persistent. The topics with very high/low correlation when $h = 0$, also exhibit a high/low correlation when $h = 1, 2$. In all graphs, red and blue shades indicate positive and negative bonds, respectively. Darker colours are associated with stronger relations – i.e. very high/low correlations. The graphs also present dendrograms on its borders as both topics and economic variables are ordered and clustered according to the bond linking them.

Figures 2.14 and 2.15, instead, show the per topic word distributions for each of the 18 highly correlated topics in Figures 2.11 to 2.13 and give some insights about the what themes are highly linked to all the economic variables considered. For instance, public expenditure (Topic 61) or the contrast to mafia and corruption (Topic 84) or government stability (Topic 89) are among the issues positively associated with the economic variables growth rates. On the other hand, the discussion about financing resources problems (Topic 12) or internal security (Topic 21) or social issues connected to labour and related contracts and welfare (Topic 31) is negatively associated with growth rates in the economy.

The existence of persistent weak and strong, positive and negative correlations between the values of the quarterly topic proportions and (all the) economic variables confirms not only that the connection between political debate and economy is not a mere hypothesis but also that such intrinsically qualitative relation can be measured, therefore modelled. Furthermore, the possibility to estimate daily topic proportions makes concrete the idea of integrating high-frequency information into the standard quarterly or monthly economic indicators. Such additional information would be more qualitative as provided by texts and very up to date, which may cast some light on the opinions and expectations that different social actors – as firms, employees and others – form about the future state of the economy.

2.5.2 Constructing TPDI

As mentioned, the idea underlying the construction of the text-based indices proposed in this work somehow resembles the approach of Larsen and Thorsrud (2019). In fact, also the procedure described here utilises the (quarterly aggregated) estimated topic proportions, $\dot{\vartheta}$, within common *auto-regressive models*, AR, (Box and Jenkins, 1970; Box et al., 2015; Whittle, 1951, among others). To be precise, not only simple AR models are utilised but also *auto-regressive with exogenous inputs models*, ARX, (Hannan, 1976, for instance) as well as *time varying parameters auto-regressive models* (Cai, 2007; Hastie and Tibshirani, 1993; Robinson, 1989; Robinson, 1991, among others), either with exogenous terms (TVARX) or without (TVAR).

Such regressions are estimated with the software R (R Core Team, 2020), through the libraries *dynlm* (Zeileis, 2019) and *tvReg* (Casas and Fernandez-Casal, 2019).

Before detailing the Textual Political Debate Indices (TPDI) constructing procedure, introducing some notation is needed.

Let $\ln Z_t - \ln Z_{t-4}$, with $t = 1, \ldots, T = 98$, be the year-on-year logarithmic differences (or growth rates) associated with Z and let z_t be the corresponding standardised values. Similarly, let $\ln \dot{\vartheta}_{i,t} - \ln \dot{\vartheta}_{i,t-4}$ be the growth rates of topic $i = 1, \ldots, K$ and let $\widetilde{\vartheta}_{i,t}$ be the corresponding standardised values.

The equations of the $AR(1)$, $TVAR(1)$, $ARX(1)$ and $TVARX(1)$ are given by:

$$z_t = \varphi z_{t-1} + u_t \tag{2.1}$$

$$z_t = \varphi_t z_{t-1} + u_t \tag{2.2}$$

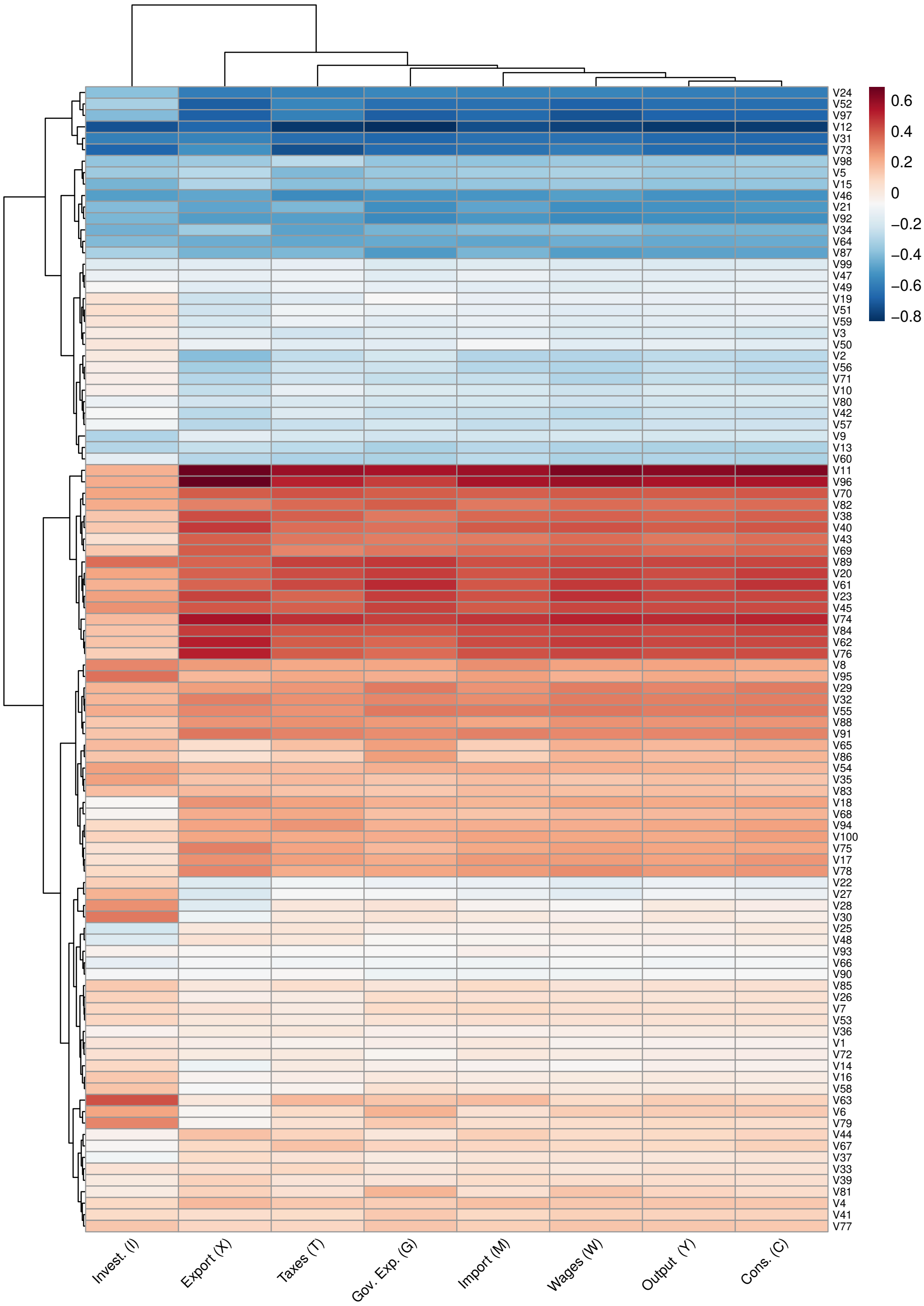


FIGURE 2.8: Correlation between quarterly topic proportions at Lag 0 of the 100 topics CTM (100% tf-idf corpus) and quarterly economic aggregates

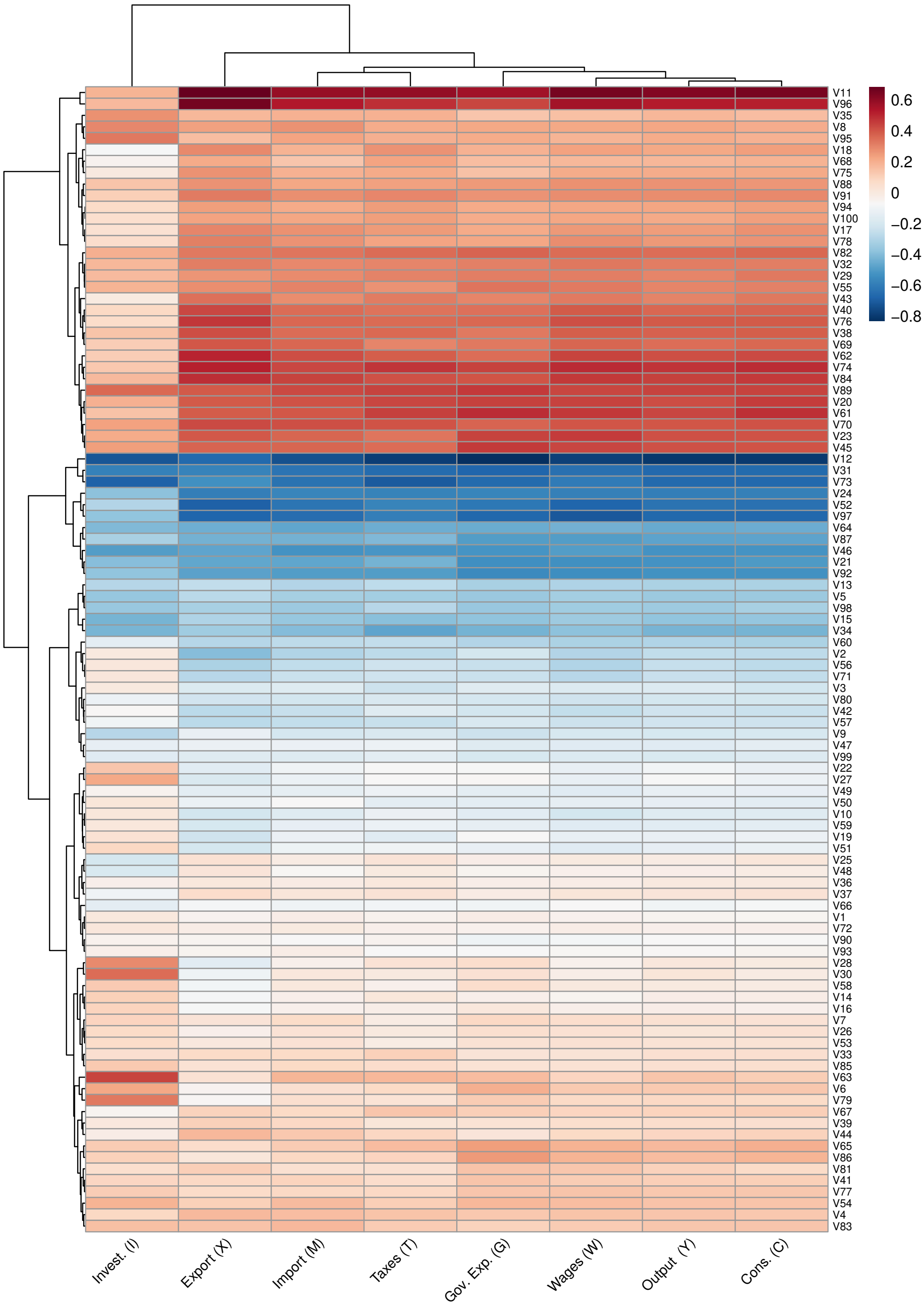


FIGURE 2.9: Correlation between quarterly topic proportions at Lag 1 of the 100 topics CTM (100% tf-idf corpus) and quarterly economic aggregates

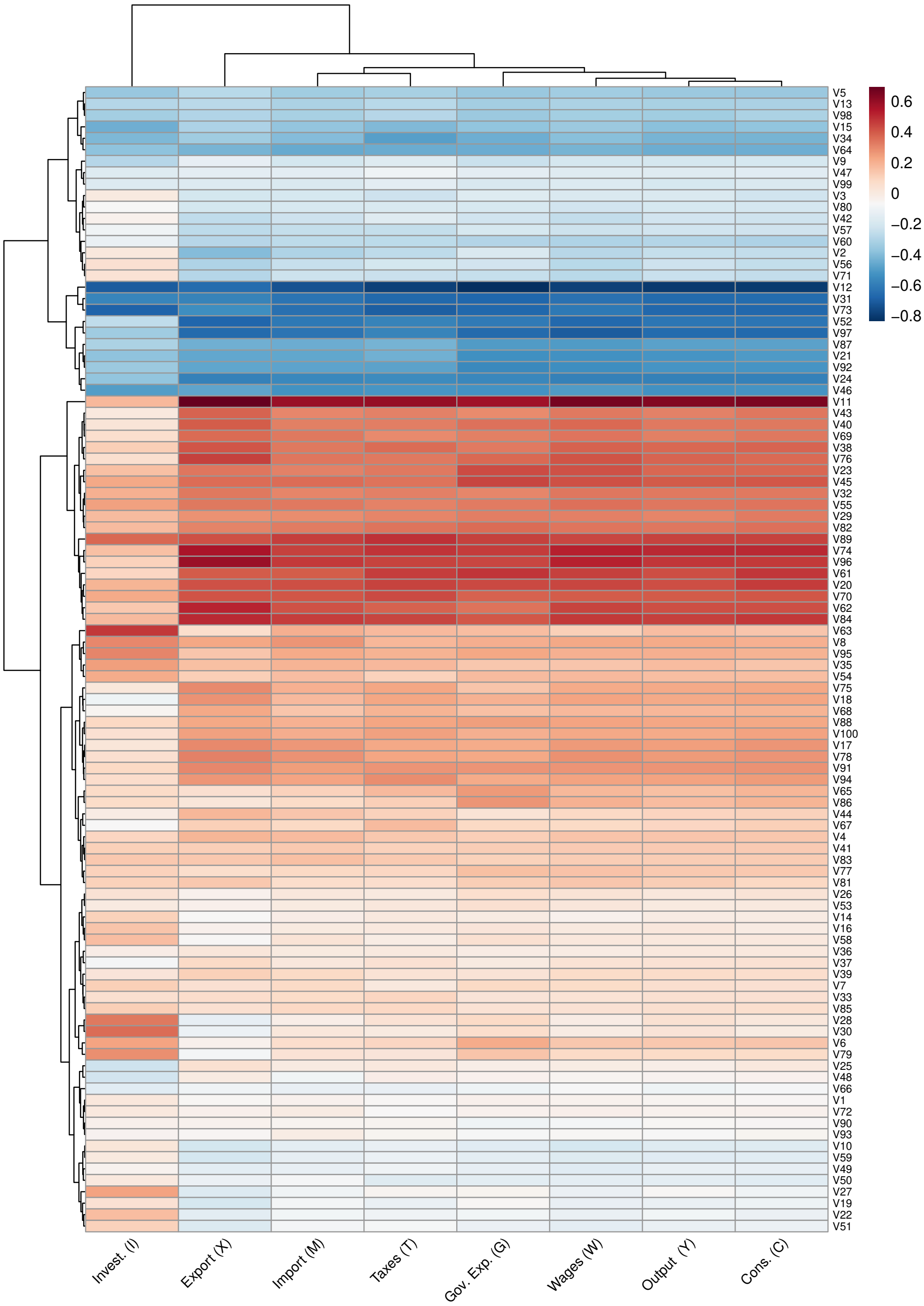


FIGURE 2.10: Correlation between quarterly topic proportions at Lag 2 of the 100 topics CTM (100% tf-idf corpus) and quarterly economic aggregates

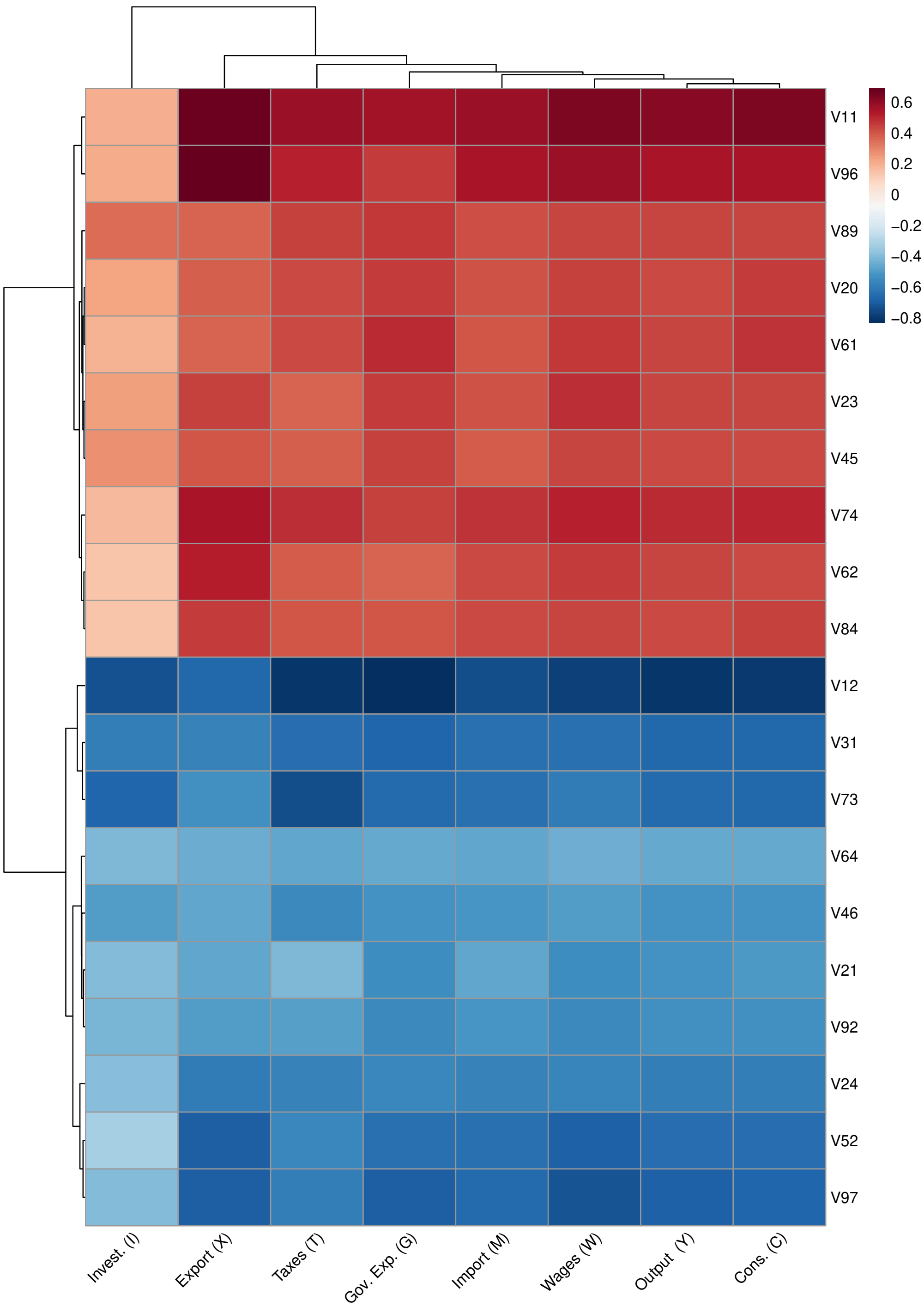


FIGURE 2.11: Correlation between quarterly topic proportions at Lag 0 of the 100 topics CTM (100% tf-idf corpus) and quarterly economic aggregates. Detail of top 10% strongest values on average

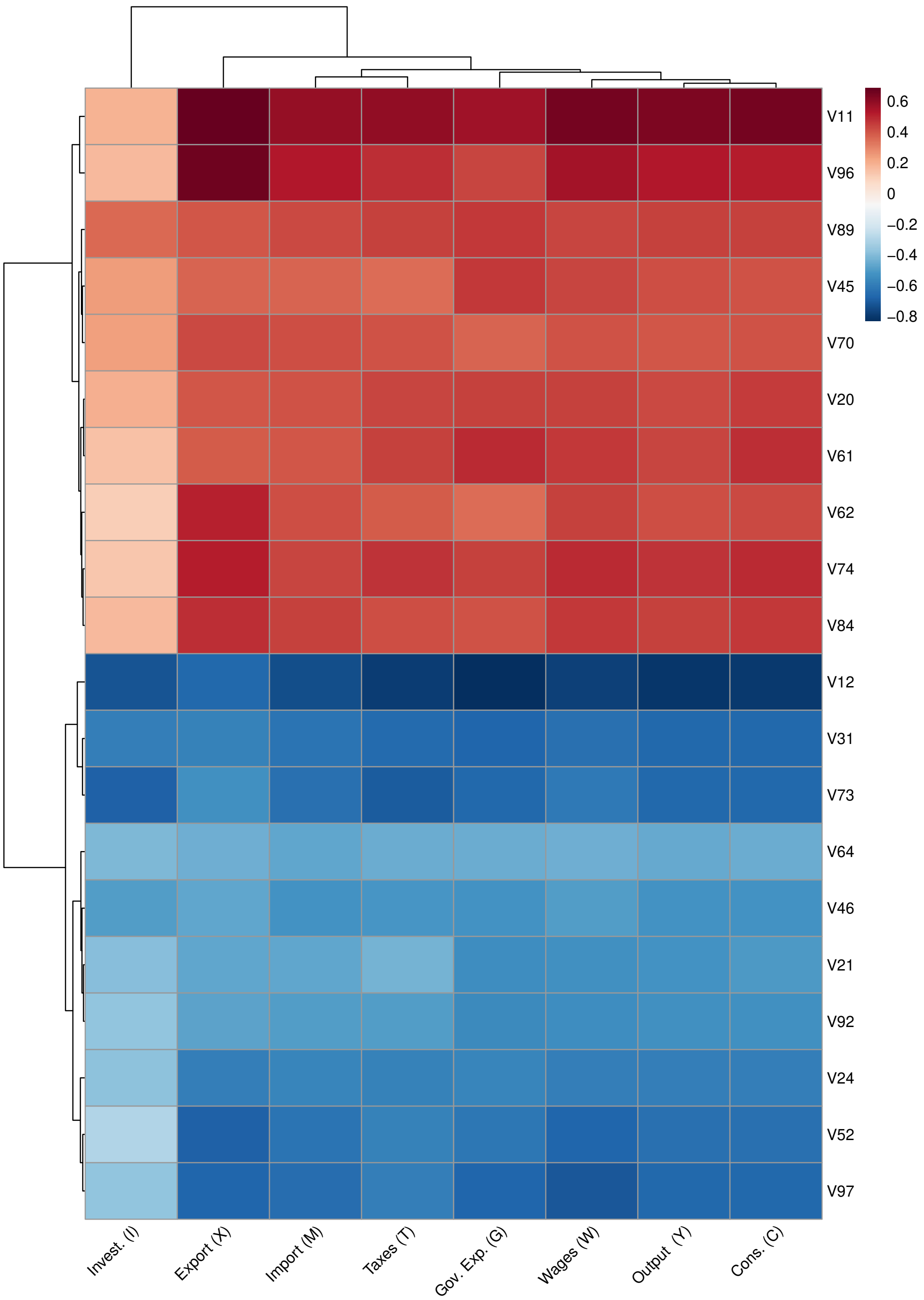


FIGURE 2.12: Correlation between quarterly topic proportions at Lag 1 of the 100 topics CTM (100% tf-idf corpus) and quarterly economic aggregates. Detail of top 10% strongest values on average

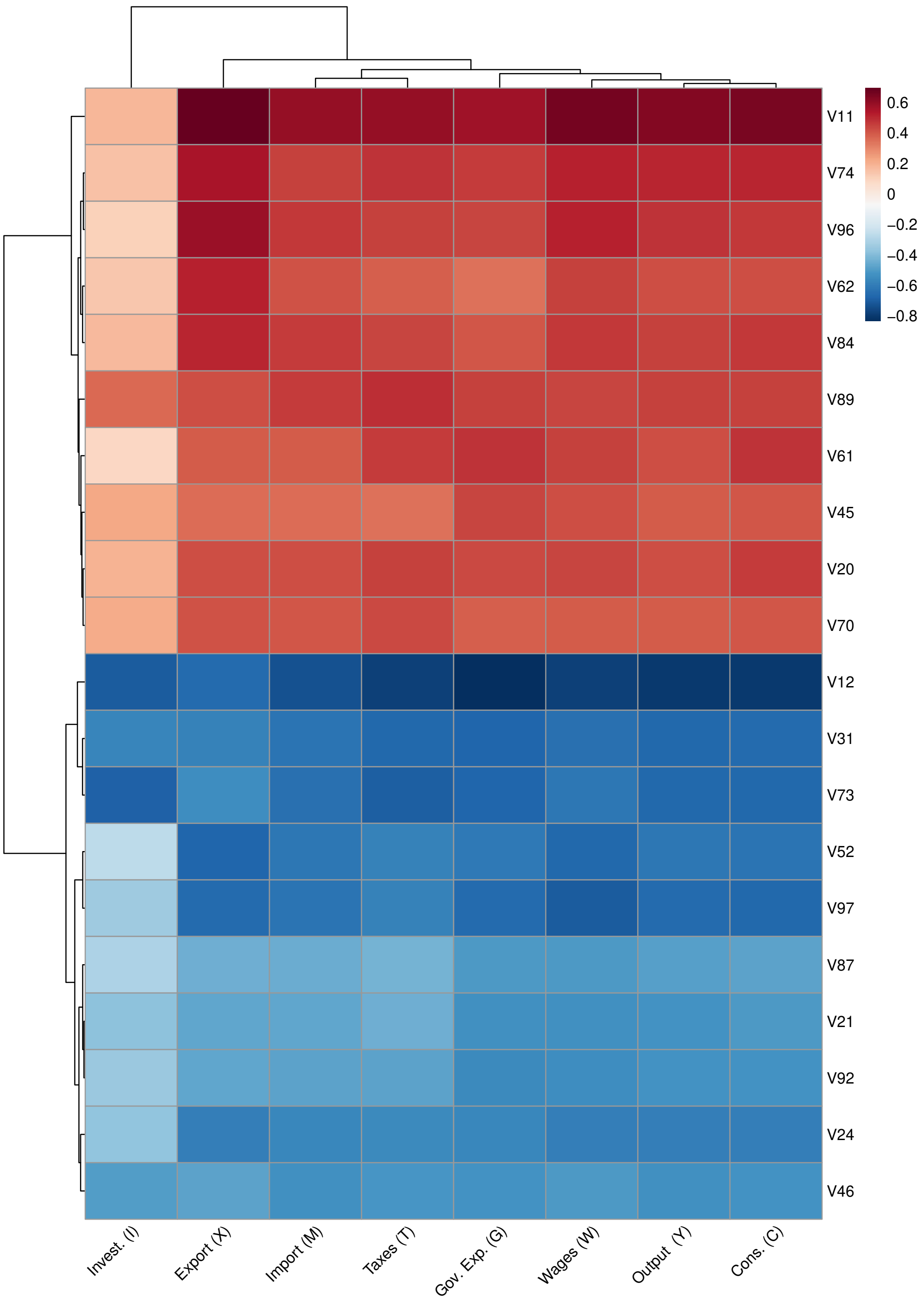


FIGURE 2.13: Correlation between quarterly topic proportions at Lag 2 of the 100 topics CTM (100% tf-idf corpus) and quarterly economic aggregates. Detail of top 10% strongest values on average

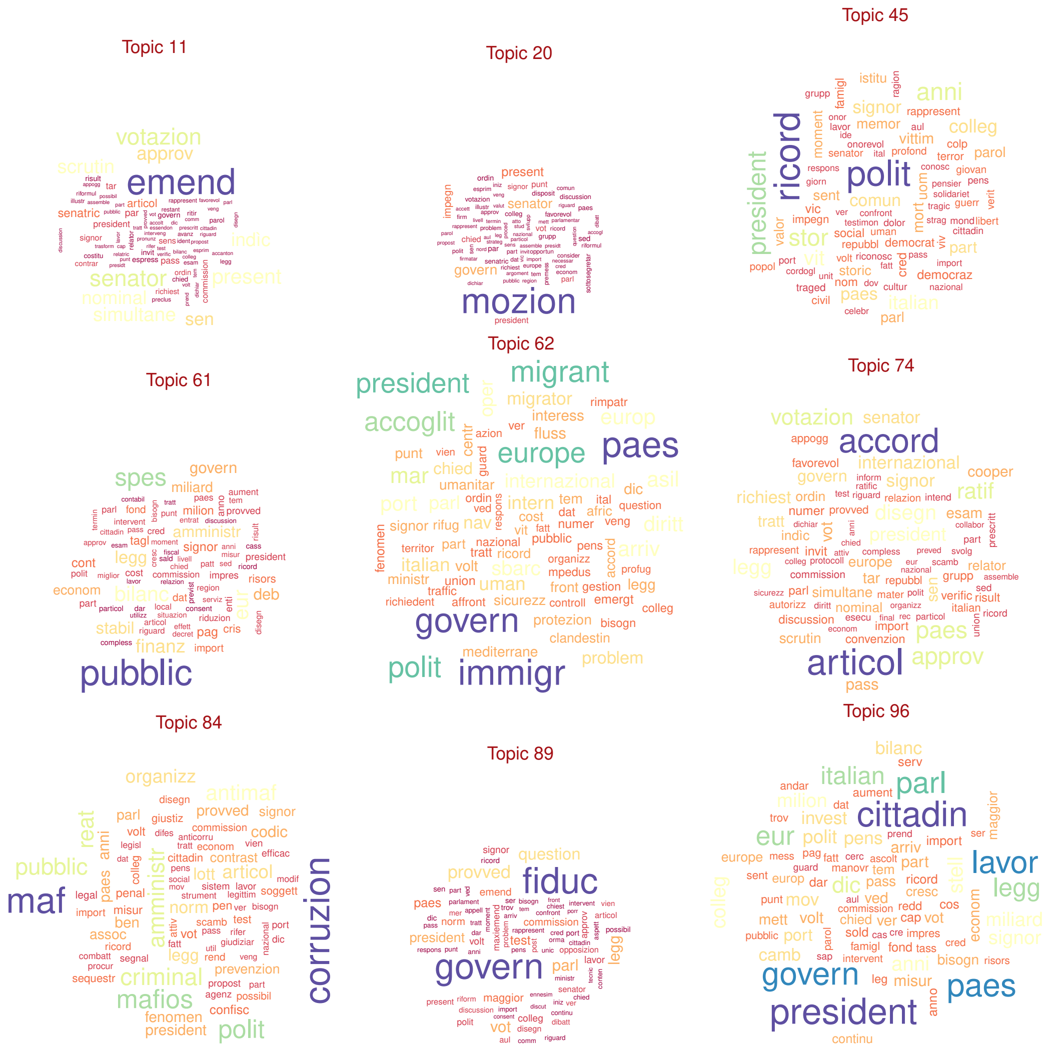


FIGURE 2.14: 100% tf-idf Corpus, $K = 100$ CTM: per topic word distribution of top 10% topics showing the strongest positive correlations with the economic aggregates at Lags 0, 1, 2

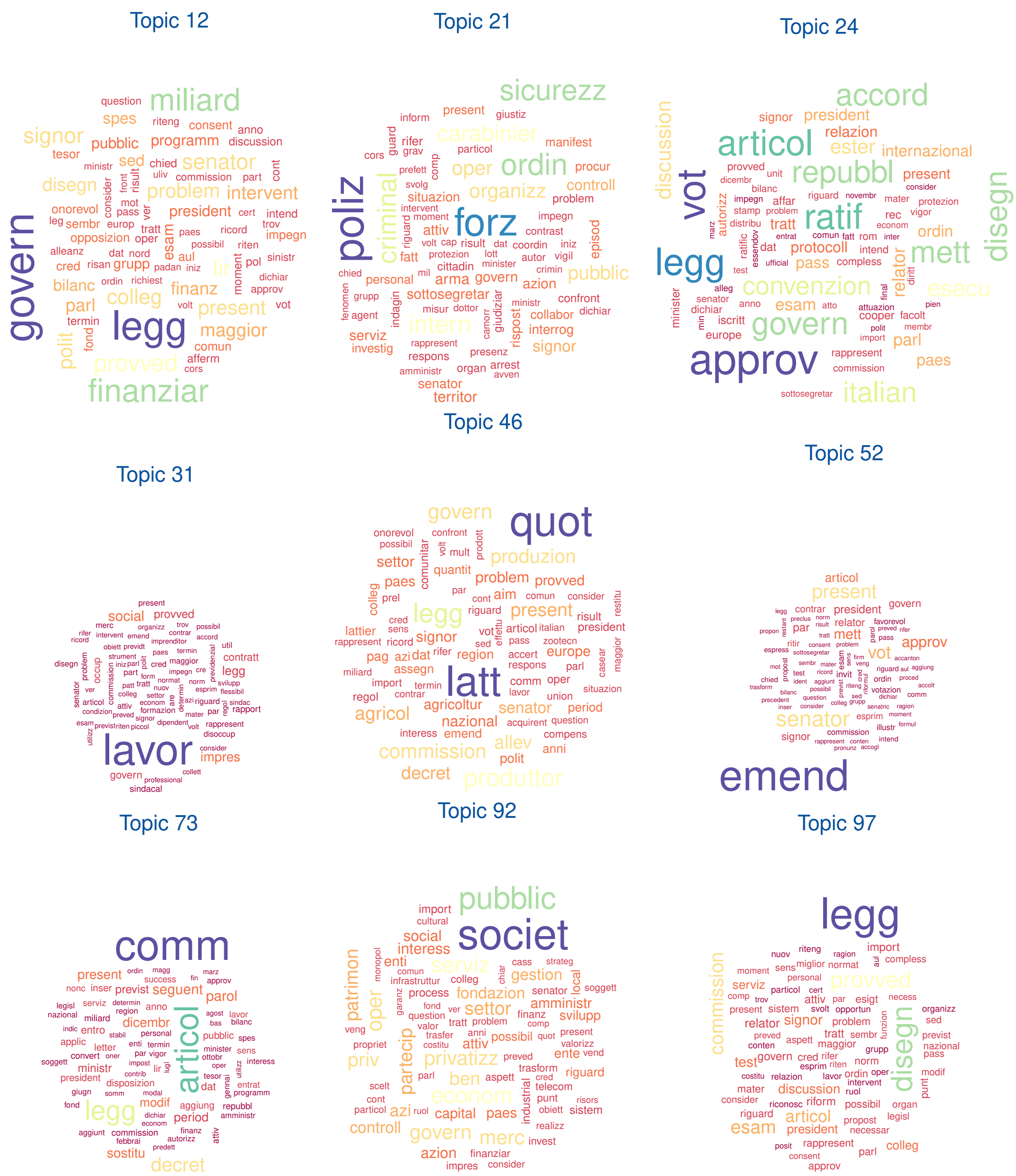


FIGURE 2.15: 100% tf-idf Corpus, $K = 100$ CTM: per topic word distribution of top 10% topics showing the strongest negative correlations with the economic aggregates at Lags 0, 1, 2

$$z_t = \varphi_i z_{t-1} + b_i \widetilde{\vartheta}_{i,t-1} + u_{i,t} \tag{2.3}$$

$$z_t = \varphi_{i,t} z_{t-1} + b_{i,t} \widetilde{\vartheta}_{i,t-1} + u_{i,t} \tag{2.4}$$

where φ and b are the model coefficients for the economic variable and the topic proportion, respectively, while u are independent and identically distributed stochastic error terms with zero mean and constant variance – i.e. $u \sim IID(0, \sigma^2)$.

Considering growth rates and standardising allows focusing on the dynamic of the variables over time and, importantly, ensures stationarity (Larsen and Thorsrud, 2019).

In what follows, when referring to equations from 2.1 to 2.2, the order of the auto-regressions is omitted, for a more agile notation: AR, ARX, $TVAR$, $TVARX$. Moreover, to avoid repeating the same expression twice, quantities calculated starting from both constant and time-varying parameters models are referred to by putting a "TV" in parenthesis, e.g. $(TV)AR$. Subscripts i and t are omitted when speaking of variables or model parameters in general.

The following paragraphs present the different versions of the TPDI. Table 2.7 provides a synoptic scheme. Figures 2.16 to 2.27 show the time series of the proposed indices versus the economic measures they are built for.

Indicator I_0

The two main components in all the proposed indices are b and $\widetilde{\vartheta}$. The reason for this is quite intuitive. Textual information is incorporated in the topic proportions $\widetilde{\vartheta}$ which, via the $(TV)ARX$ regressions, are given a specific weight each. In other terms, b represent the additional information provided by each topic (proportion) when this is used as an exogenous variable in the auto-regressions for the economic measure at hand.

As K $(TV)ARX$ regressions are estimated, to obtain a unique value at each time, a simple operation is summing across topics the products between b and $\widetilde{\theta}$.

The resulting indicator is:

$$I_{0,ARX\text{-}AR} = \sum_{i=1}^{K} b_i \widetilde{\vartheta}_{i,t-1} \tag{2.5}$$

A slightly more complex version of this first indicator arises from allowing topic influence to change over time. Therefore, equation 2.5 becomes

$$I_{0,TVARX\text{-}TVAR} = \sum_{i=1}^{K} b_{i,t} \widetilde{\vartheta}_{i,t-1} \tag{2.6}$$

Indicator I_1

The indicators in Equations 2.5 and 2.6 are not able to capture the overall performance of the $(TV)ARX$ regressions each beta comes from. More in detail, among those K $(TV)ARX$ regressions, some would have a better fit than others, meaning that the topic (proportion) included as an exogenous term does have a greater influence on the economic variable.

To account for this circumstance, the following strategy is adopted. For each $(TV)ARX$ regression the R-squared, R^2, is computed and compared with the R^2 of the simpler $(TV)AR$ so to obtain the ratios:

$$\widetilde{w}_i = \frac{R^2_{i,(TV)ARX}}{R^2_{i,(TV)AR}} \tag{2.7}$$

which, being by construction either greater or lower than 1, are then normalized in the interval $[0, 1]$ trhough the min-max transformation:

$$w_i = \frac{\widetilde{w}_i - min(\widetilde{w}_i)}{max(\widetilde{w}_i) - min(\widetilde{w}_i)} \tag{2.8}$$

In this way, the $(TV)ARX$ model with the best overall fitting to the economic variable (compared to the $(TV)AR$) is assigned a normalised weight equal to 1. Conversely, the $(TV)ARX$ model that exhibits the worst fitting receives a normalised weight equal to 0.

Therefore, the previous I_0 indicators become, respectively:

$$I_{1,ARX\text{-}AR} = \sum_{i=1}^{K} w_i b_i \widetilde{\vartheta}_{i,t-1} \tag{2.9}$$

$$I_{1,TVARX\text{-}TVAR} = \sum_{i=1}^{K} w_i b_{i,t} \widetilde{\vartheta}_{i,t-1} \tag{2.10}$$

Indicator I_2

Overall goodness-of-fit is not the only measure for models quality. A pointwise evaluation is also possible. In fact, for $t = 1, \ldots, T$, each of the K $(TV)ARX$ regressions gives an estimate of the macro-economic variable value at that time. Again, some auto-regressions perform better than others, as the specific topic they include does have more explaining power. Therefore, one may assign such better models a greater weight to emphasize topics being more relevant at some times rather than others. Hence, the following strategy is adopted. Moving from the consideration that a good point estimate also implies a low value in the corresponding model residual, for each of the $(TV)ARX$ regressions the following square residuals ratios are calculated:

$$\widetilde{u}_{i,t} = \frac{\widehat{u}^2_{i,t_{(TV)ARX}}}{\widehat{u}^2_{i,t_{(TV)AR}}} \tag{2.11}$$

Equation 2.11 allows the comparison between the squared residuals $\widehat{u}^2_{i,t}$ from each of the K $(TV)ARX$ regressions and the corresponding squared residual from the $(TV)AR$ regressions at the same instant t.

To obtain single models weights, the ratios between the sum over topics $i = 1, \ldots, K$ of the computed squared residuals ratios over the *i-th* is also calculated:

$$\widetilde{v}_{i,t} = \frac{\sum_{i=1}^{K} \widetilde{u}^2_{i,t}}{\widetilde{u}^2_{i,t}} \tag{2.12}$$

Similarly to equation 2.8, the obtained weights are then normalised in $[0, 1]$:

$$v_t = \frac{\widetilde{v}_{i,t} - min(\widetilde{v}_{i,t})}{max(\widetilde{v}_{i,t}) - min(\widetilde{v}_{i,t})} \tag{2.13}$$

As a result, the two I_2 indicators are respectively given by:

$$I_{2,ARX\text{-}AR} = \sum_{i=1}^{K} v_{i,t} b_i \widetilde{\vartheta}_{i,t-1} \tag{2.14}$$

$$I_{2,TVARX\text{-}TVAR} = \sum_{i=1}^{K} v_{i,t} b_{i,t} \widetilde{\vartheta}_{i,t-1} \tag{2.15}$$

Indicator I_3

Finally, to allow for both the intuitions underlying I_1 and I_2 – namely overall and point-wise goodness of fit – two more indicators are computed:

$$I_{3,ARX\text{-}AR} = \sum_{i=1}^{K} w_i v_{i,t} b_i \widetilde{\vartheta}_{i,t-1} \tag{2.16}$$

$$I_{3,TVARX\text{-}TVAR} = \sum_{i=1}^{K} w_i v_{i,t} b_{i,t} \widetilde{\vartheta}_{i,t-1} \tag{2.17}$$

In Equations 2.16 and 2.17 the quarterly topic proportions $\widetilde{\vartheta}_{i,t-1}$ are weighted by their specific regression parameter b, by the normalised *inter-topics* weights w which put more emphasis on the best topic overall, and by the normalised *infra-time* weights v which, instead, emphasise the best topic at each time.

As a final remark, similarly to Larsen and Thorsrud (2019), the topic proportions are indexed by $t - 1$ both in the regressions and indices to allow for delayed effects of the parliamentary debate on the economic variable at hand.

TABLE 2.7: TPDI Synoptic Scheme

Indicator	ARX-AR based	TVARX-TVAR based
I_0	$\sum_{i=1}^{K} b_i \widetilde{\vartheta}_{i,t-1}$	$\sum_{i=1}^{K} b_{i,t} \widetilde{\vartheta}_{i,t-1}$
I_1	$\sum_{i=1}^{K} w_i b_i \widetilde{\vartheta}_{i,t-1}$	$\sum_{i=1}^{K} w_i b_{i,t} \widetilde{\vartheta}_{i,t-1}$
I_2	$\sum_{i=1}^{K} v_{i,t} b_i \widetilde{\vartheta}_{i,t-1}$	$\sum_{i=1}^{K} v_{i,t} b_{i,t} \widetilde{\vartheta}_{i,t-1}$
I_3	$\sum_{i=1}^{K} w_i v_{i,t} b_i \widetilde{\vartheta}_{i,t-1}$	$\sum_{i=1}^{K} w_i v_{i,t} b_{i,t} \widetilde{\vartheta}_{i,t-1}$

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

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

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

2.5.3 TPDI Evaluation

After constructing the 4 TPDI versions for each of the 8 economic variables, indices' ability to adapt to the corresponding economic series is both measured by utilising the Root Mean Squared Error (RMSE) as a distance measure, and also tested through the Model Confidence Set (MCS) procedure of Hansen, Lunde, and Nason (2011). Specifically, for both approaches, the Squared Error (SE) function is calculated as:

$$SE_t^{(z)} = \left(z_t - TPDI_t^{(z)}\right)^2 \tag{2.18}$$

where $TPDI \in \{I0, I1, I2, I3\}$ is the index version at hand.

The SE is used as a distance function in each of the 7 MCS performed: one per $K \in \{20, 40, 60, 80, 100, 120, 140\}$ to check the sensitivity of each index version to the number of topics. Each MCS is computed with a 0.99 confidence level.

Evaluation results for both approaches indicate the $I_{3,ARX\text{-}AR}$, $I_{2,ARX\text{-}AR}$ as the best TPDI versions. More details are given in Table 2.8, which reports the best indicators from the MCS. Figures 2.32 to 2.39 show the sensitivity of the RMSE to the number of topics K in the CTM indices are based on.

TABLE 2.8: Best TPDI per K and Economic Aggregate

Number of Topics	**Economic Variables**							
K	Y	W	C	I	G	T	X	M
20	$I_{2,ARX\text{-}AR}$	$I_{2,TVARX\text{-}TVAR}$	$I_{2,TVARX\text{-}TVAR}$	$I_{2,ARX\text{-}AR}$	$I_{2,ARX\text{-}AR}$	$I_{2,TVARX\text{-}TVAR}$	$I_{3,ARX\text{-}AR}$	$I_{2,ARX\text{-}AR}$
40	$I_{2,ARX\text{-}AR}$	$I_{2,TVARX\text{-}TVAR}$	$I_{3,ARX\text{-}AR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{2,ARX\text{-}AR}$	$I_{2,ARX\text{-}AR}$	$I_{3,ARX\text{-}AR}$	$I_{3,ARX\text{-}AR}$
60	$I_{2,ARX\text{-}AR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{3,ARX\text{-}AR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{2,ARX\text{-}AR}$	$I_{3,ARX\text{-}AR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{3,TVARX\text{-}TVAR}$
80	$I_{3,TVARX\text{-}TVAR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{2,ARX\text{-}AR}$	$I_{3,ARX\text{-}AR}$	$I_{2,TVARX\text{-}TVAR}$	$I_{2,ARX\text{-}AR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{3,ARX\text{-}AR}$
100	$I_{3,ARX\text{-}AR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{2,ARX\text{-}AR}$	$I_{2,ARX\text{-}AR}$	$I_{3,ARX\text{-}AR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{3,ARX\text{-}AR}$
120	$I_{2,ARX\text{-}AR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{3,TVARX\text{-}TVAR}$	$I_{2,TVARX\text{-}TVAR}$	$I_{3,ARX\text{-}AR}$	$I_{3,ARX\text{-}AR}$	$I_{2,TVARX\text{-}TVAR}$	$I_{2,ARX\text{-}AR}$
140	$I_{3,ARX\text{-}AR}$	$I_{3,ARX\text{-}AR}$	$I_{3,ARX\text{-}AR}$	$I_{2,TVARX\text{-}TVAR}$	$I_{3,ARX\text{-}AR}$	$I_{2,ARX\text{-}AR}$	$I_{3,ARX\text{-}AR}$	$I_{3,TVARX\text{-}TVAR}$

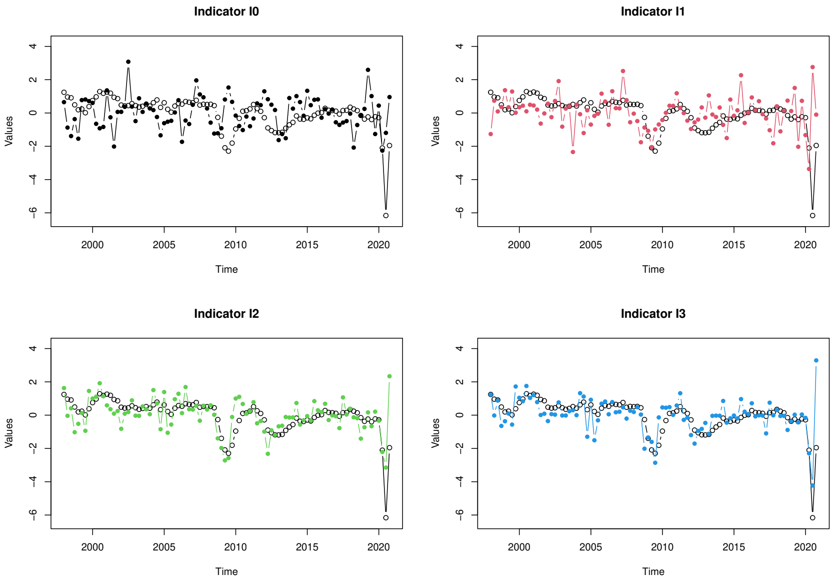


FIGURE 2.16: 100% TF-IDF corpus $K = 100$ GDP ARX based TPDI over time

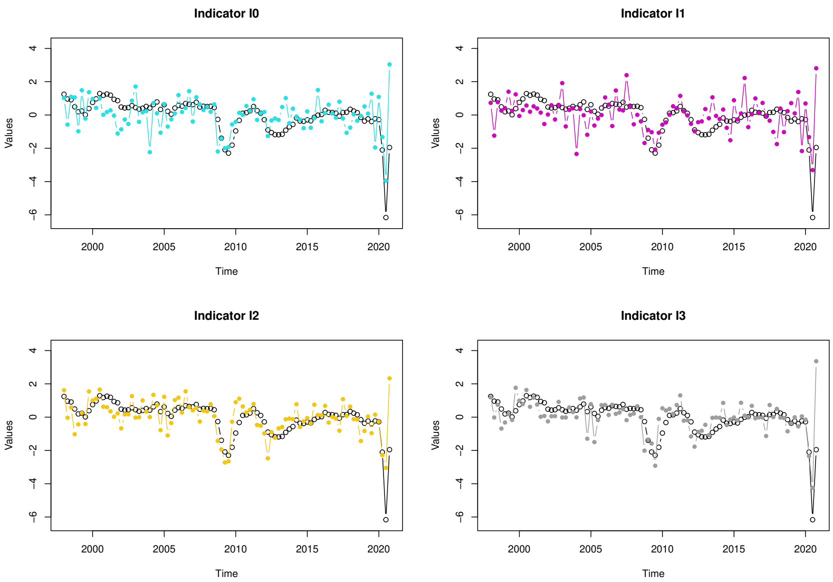


FIGURE 2.17: 100% TF-IDF corpus $K = 100$ GDP TVARX based TPDI over time

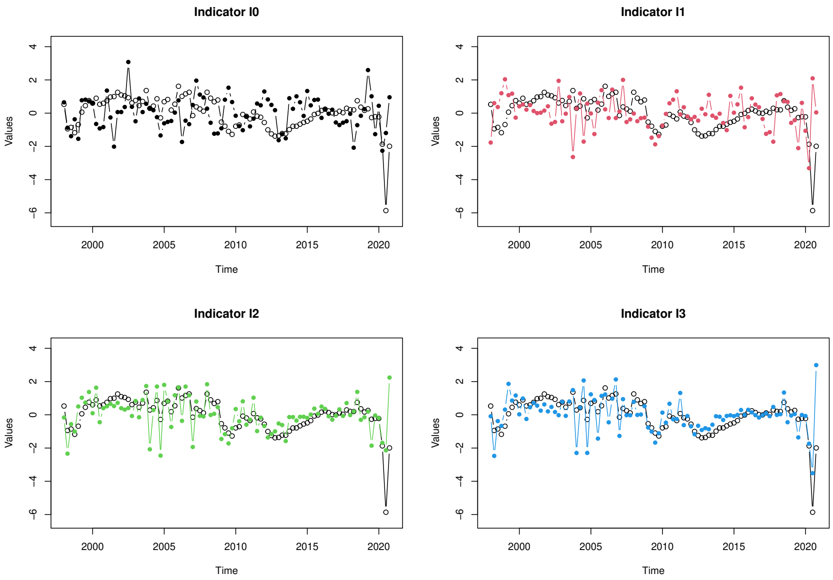


FIGURE 2.18: 100% TF-IDF corpus $K = 100$ Wages ARX based TPDI over time

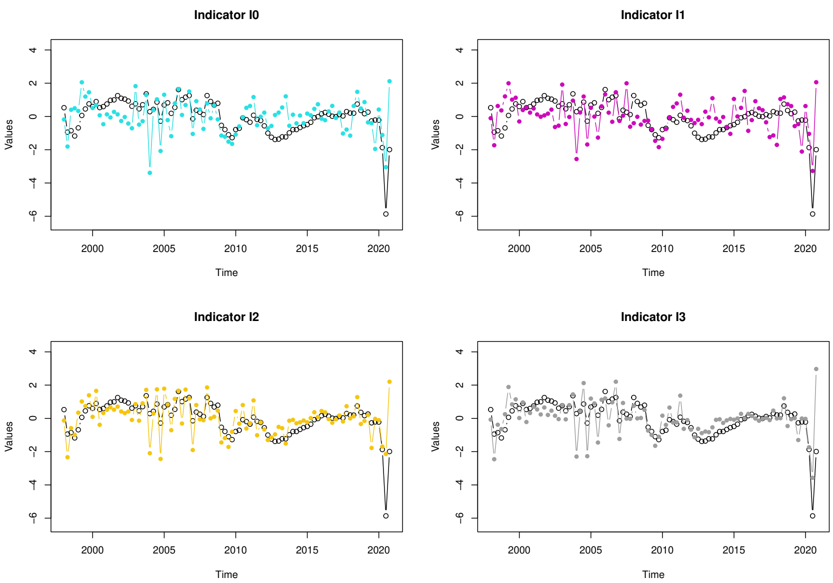


FIGURE 2.19: 100% TF-IDF corpus $K = 100$ Wages TVARX based TPDI over time

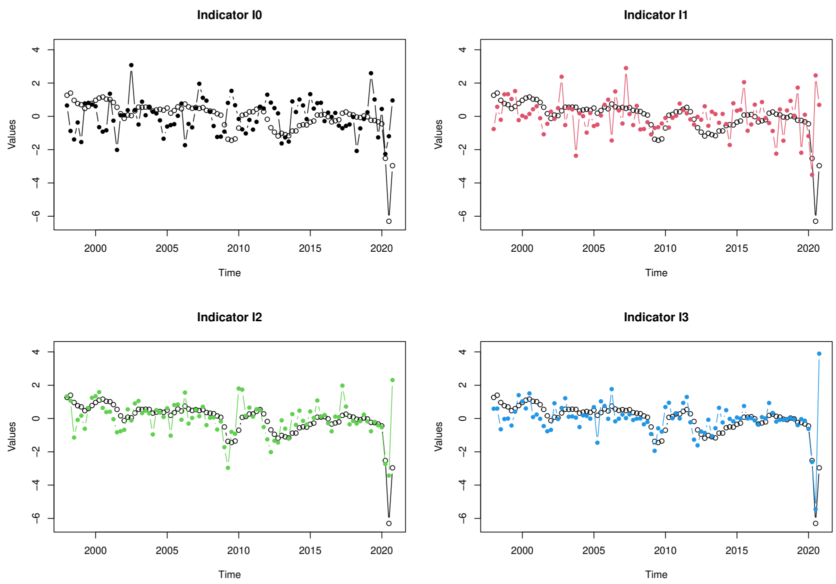


FIGURE 2.20: 100% TF-IDF corpus $K = 100$ Private Consumption ARX based TPDI over time

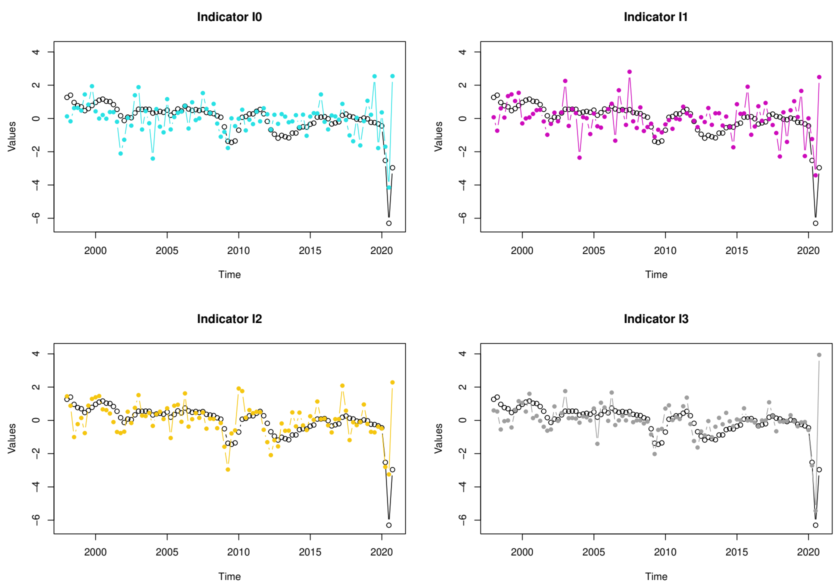


FIGURE 2.21: 100% TF-IDF corpus $K = 100$ Private Consumption TVARX based TPDI over time

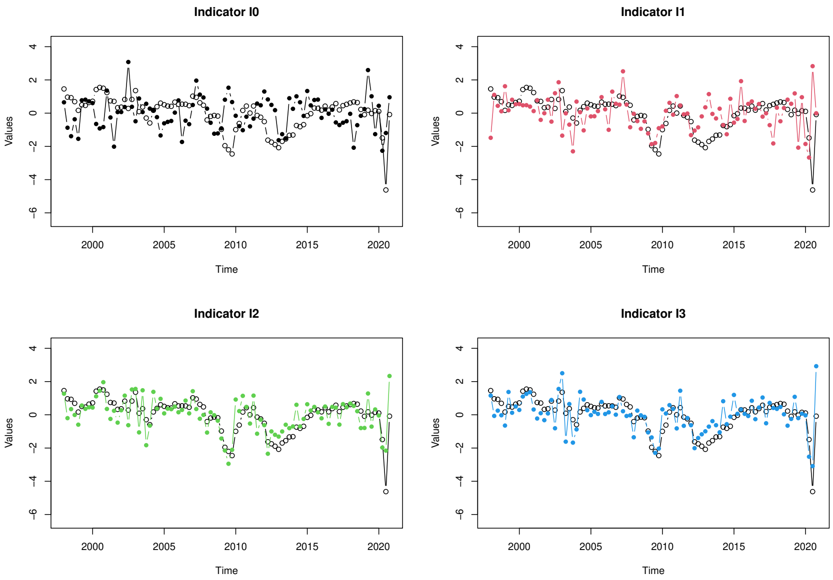


FIGURE 2.22: 100% TF-IDF corpus $K = 100$ Investments ARX based TPDI over time

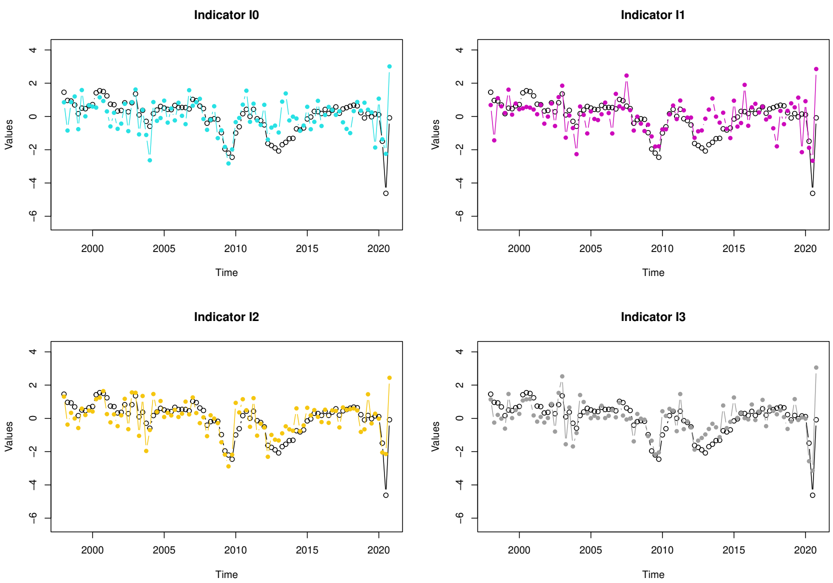


FIGURE 2.23: 100% TF-IDF corpus $K = 100$ Investments TVARX based TPDI over time

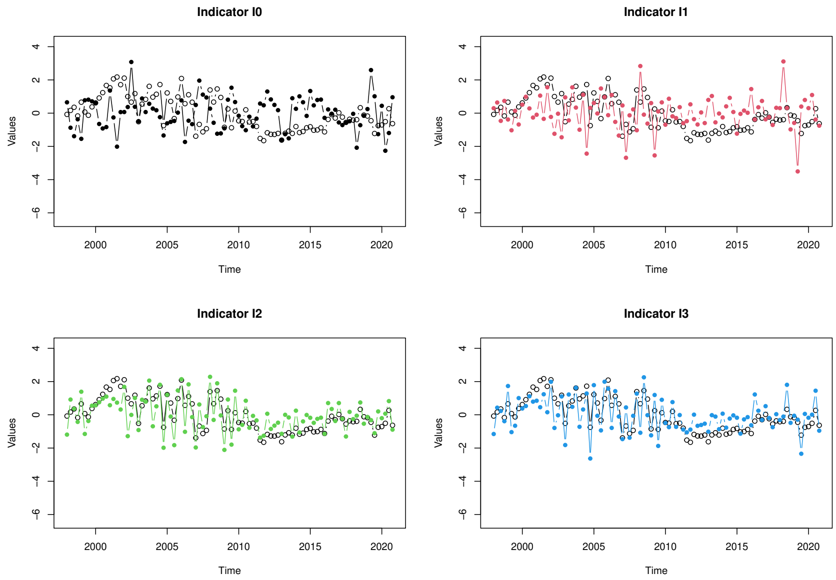


FIGURE 2.24: 100% TF-IDF corpus $K = 100$ Public Expenditure ARX based TPDI over time

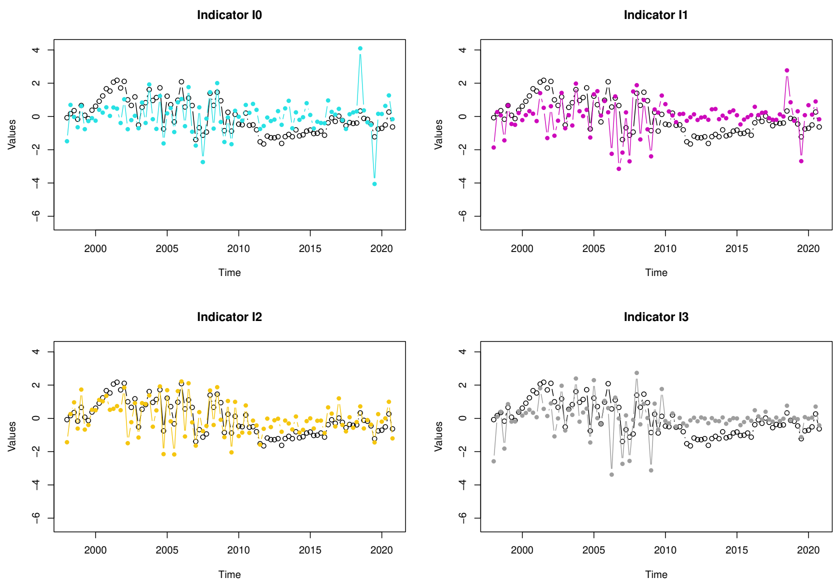


FIGURE 2.25: 100% TF-IDF corpus $K = 100$ Public Expenditure TVARX based TPDI over time

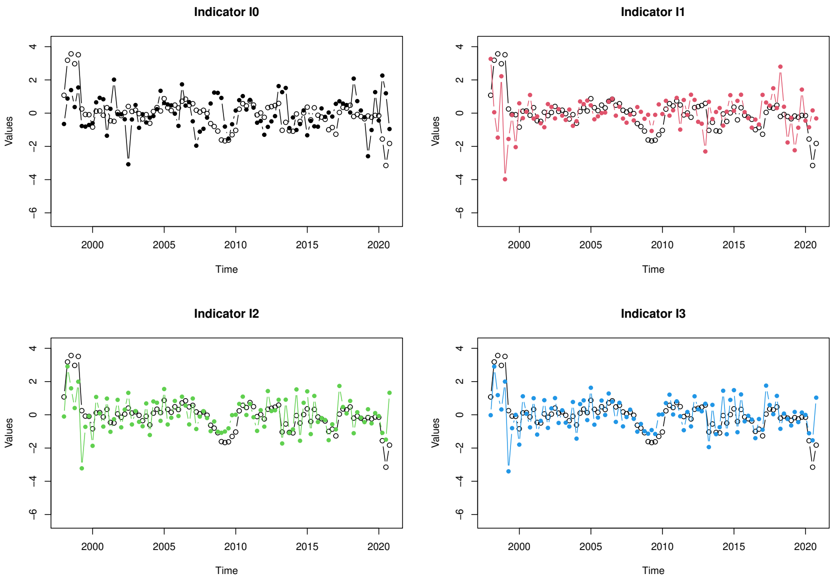


FIGURE 2.26: 100% TF-IDF corpus $K = 100$ Taxation ARX based TPDI over time

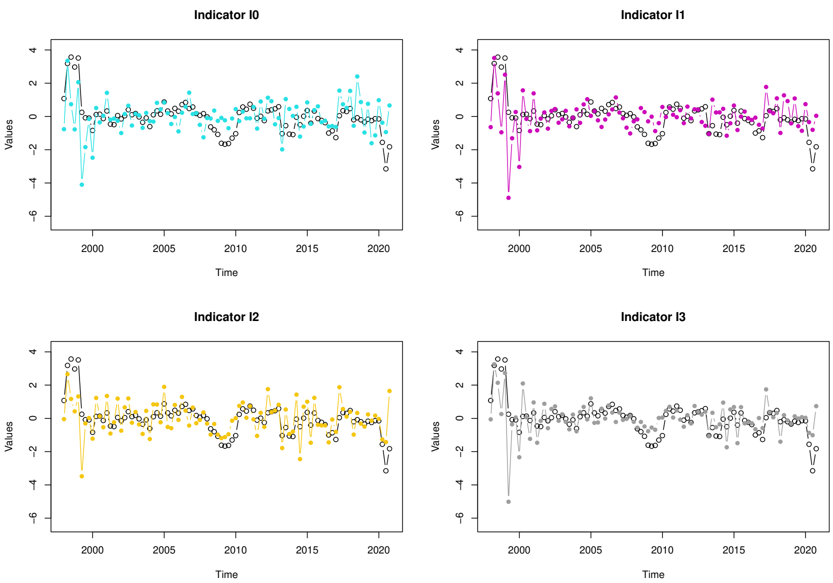


FIGURE 2.27: 100% TF-IDF corpus $K = 100$ Taxation TVARX based TPDI over time

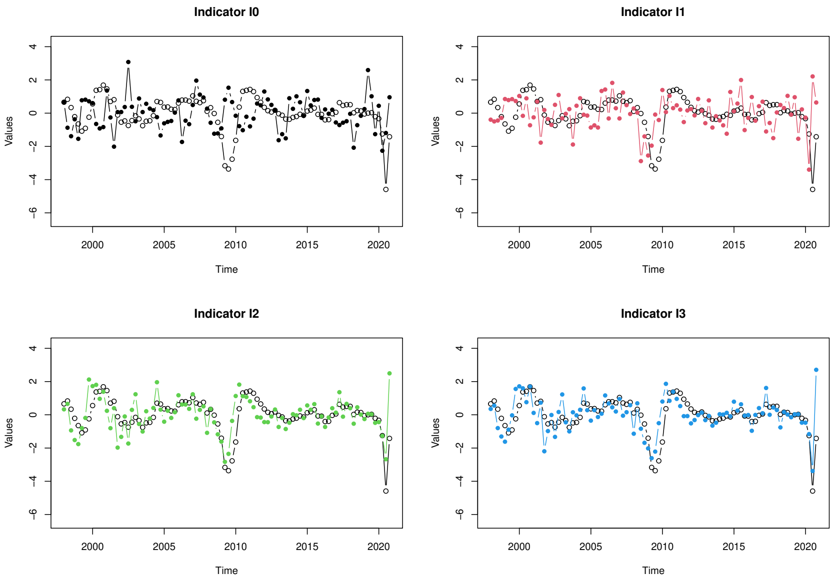


FIGURE 2.28: 100% TF-IDF corpus $K = 100$ Exports ARX based TPDI over time

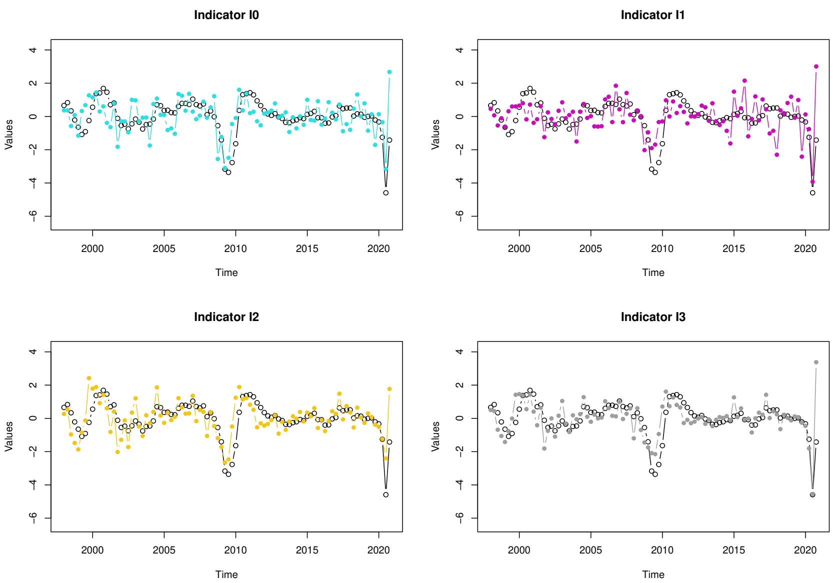


FIGURE 2.29: 100% TF-IDF corpus $K = 100$ Exports TVARX based TPDI over time

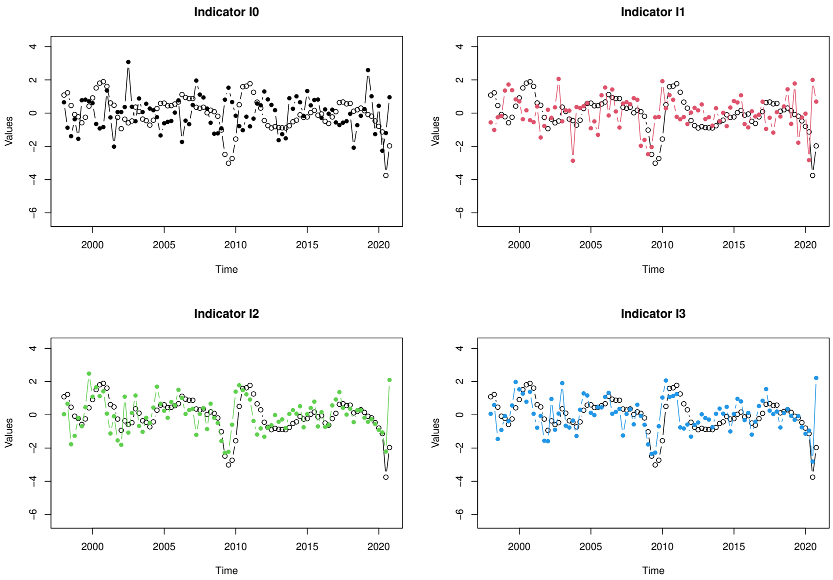


FIGURE 2.30: 100% TF-IDF corpus $K = 100$ Import ARX based TPDI over time

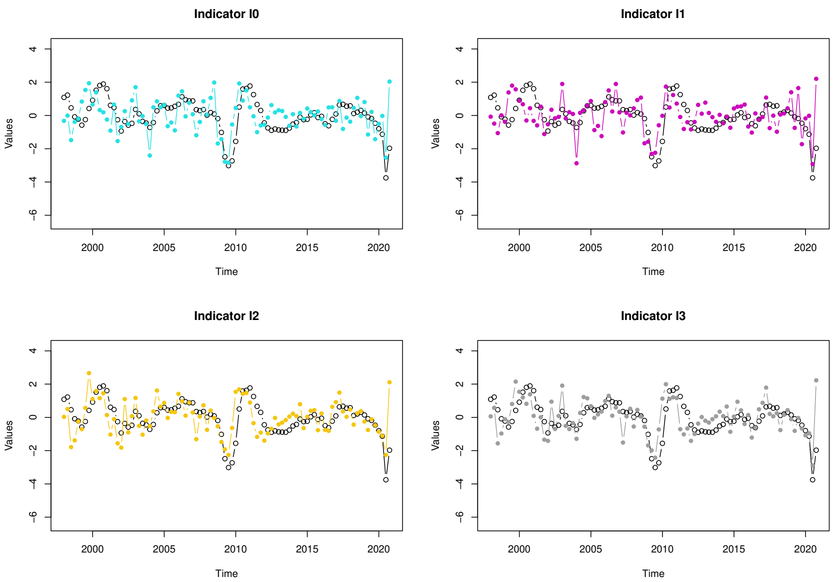


FIGURE 2.31: 100% TF-IDF corpus $K = 100$ Import TVARX based TPDI over time

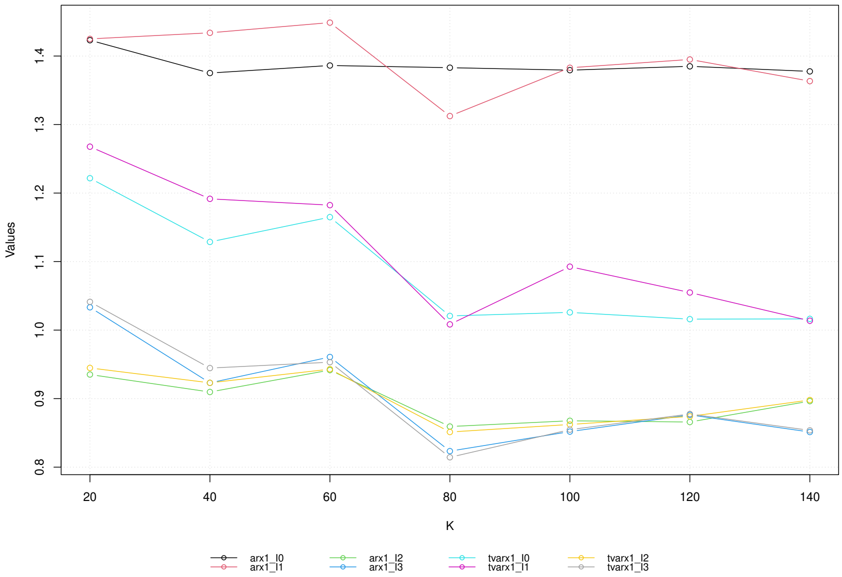


FIGURE 2.32: 100% TF-IDF Corpus RMSE Sensitivity to K for the Output TPDI

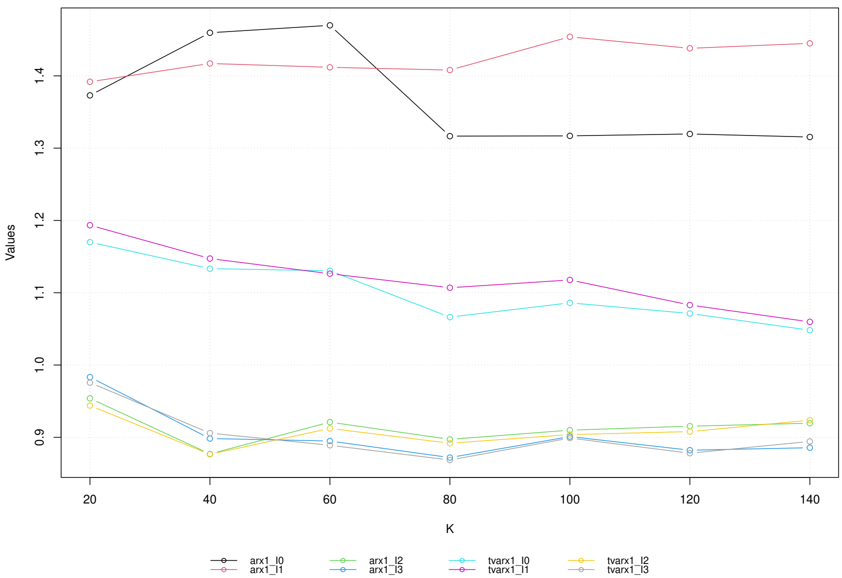


FIGURE 2.33: 100% TF-IDF Corpus RMSE Sensitivity to K for the Wages TPDI

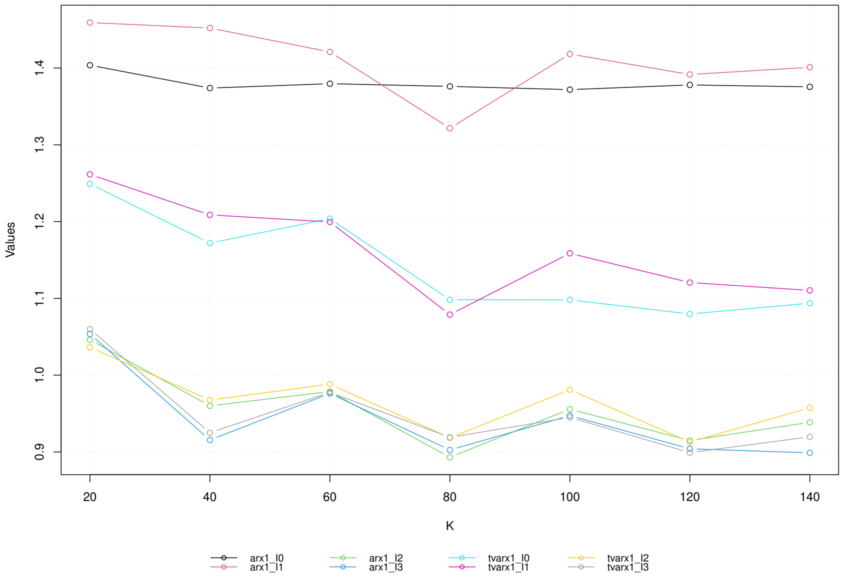


FIGURE 2.34: 100% TF-IDF Corpus RMSE Sensitivity to K for the Consumptions TPDI

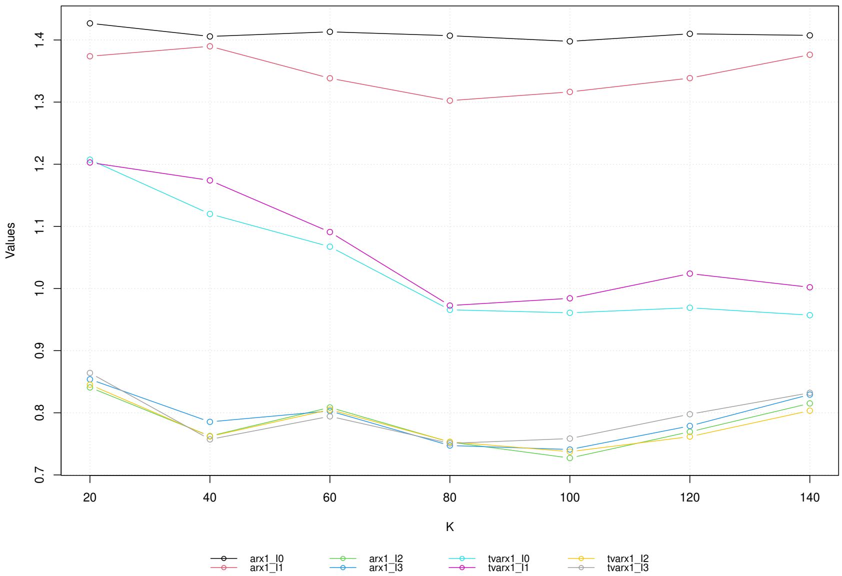


FIGURE 2.35: 100% TF-IDF Corpus RMSE Sensitivity to K for the Investments TPDI

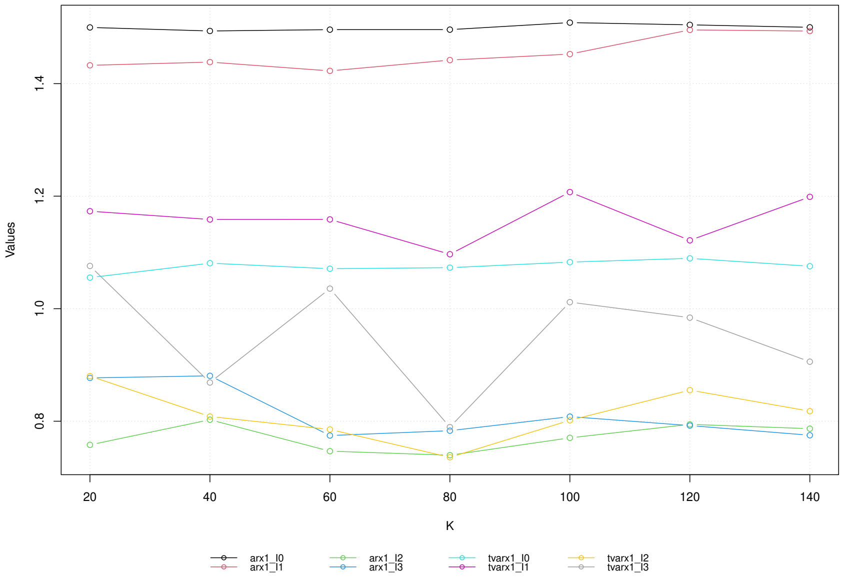


FIGURE 2.36: 100% TF-IDF Corpus RMSE Sensitivity to K for the Gov. Exp. TPDI

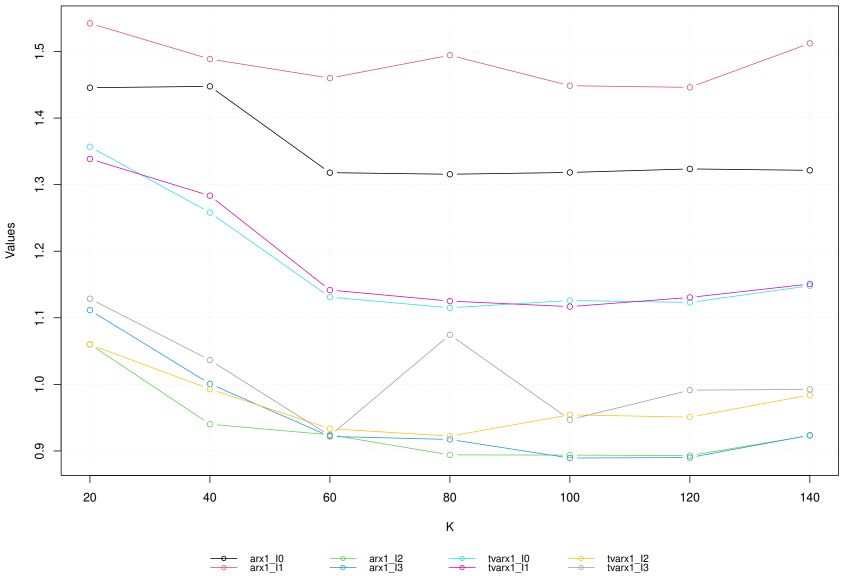


FIGURE 2.37: 100% TF-IDF Corpus RMSE Sensitivity to K for the Taxes TPDI

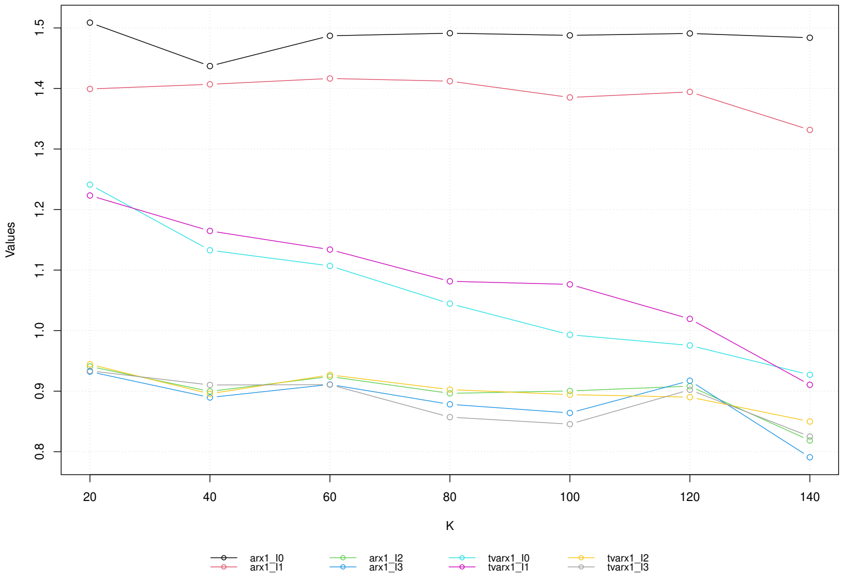


FIGURE 2.38: 100% TF-IDF Corpus RMSE Sensitivity to *K* for the Exports TPDI

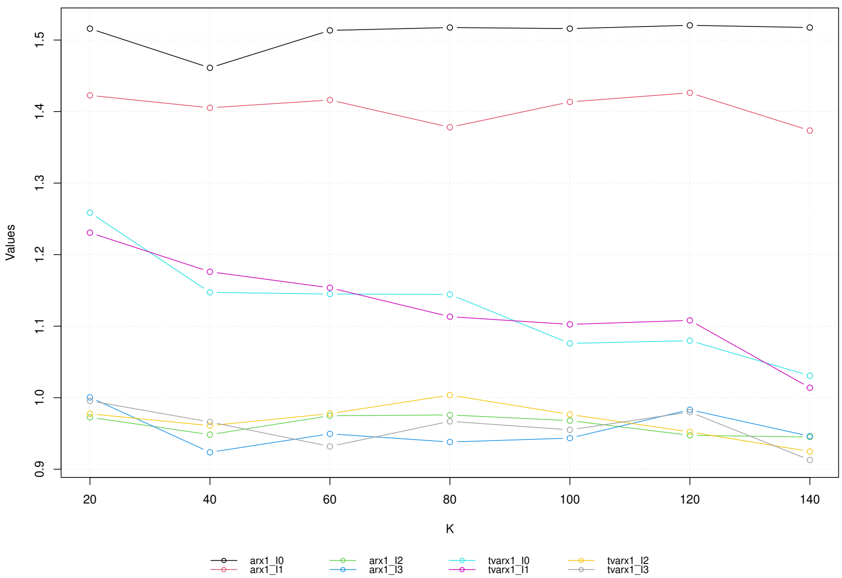


FIGURE 2.39: 100% TF-IDF Corpus RMSE Sensitivity to *K* for the Imports TPDI

Chapter 3

TPPI: Textual Political Polarity Indices. The case of Italian GDP

3.1 Introduction

This work proposes new text-based polarity indices to measure the qualitative information conveyed by parliamentary debates and improve the predictions of traditional quantitative macroeconomic indicators.

The need for exploiting qualitative information from texts rises from some drawbacks of conventional National Statistics. For instance, recently, Thorsrud (2020) underlined their lack of "*timeliness*" and capability to incorporate high-frequency information reflecting economic fluctuations, two characteristics that are crucial especially for policymakers who attempt to react to (or, possibly, lead) changes in the state of the economy.

In this framework, textual data might make a substantive difference as they are intrinsically high dimensional and therefore, especially if collected at a lower frequency than standard quantitative data, might unveil some of those (high frequency) information about the economic dynamics that the state-of-the-art models cannot currently take into account.

Fortunately, nowadays, there are plenty of content analysis techniques developed in Natural Language Processing (NLP) and text analysis fields that allow extracting information from different types of text.

3.2 Related literature

This section provides a brief overview of research applying sentiment analysis in the economic-related literature and political science. Such a survey does not claim to be complete as new works are constantly published.

In economics, thus far sentiment analysis has been applied in the following main areas:

- prediction of stock prices, returns and their volatility;
- analysis of the effects of central banks communication;

- development of indices reflecting various aspects of the economy.

Such areas are not intended as rigid categories as many works analyse several aspects performing different tasks. Thus overlapping is quite common. A recent survey of application of sentiment analysis in economics is provided by Algaba et al. (2020).

Concerning the first point, most of the works focusing on financial variables such as stock prices, returns and volatility either apply sentiment analysis on twits or newspapers articles. The reason for that is straightforward: financial data are usually high frequency. Hence, they are assumed to likely co-move with the information conveyed by daily (newspapers articles) or infra-daily (twits) texts sources.

Among the first works in the field, Bollen, Mao, and Zeng (2011) focused on the Dow Jones Industrial Average (DJIA) ad found out that mood measurements based on Twitter feeds significantly reduce prediction error. More recent works in the field focused on: systemic risk and contagion effects (Cerchiello, Giudici, and Nicola, 2017); Chief Executive Officer (CEO) succession (Leitch and Sherif, 2017); construction of a "*lexicon of words used by online investors*" (Renault, 2017); comparison of co-movement between social media metrics and stock returns for firms having official Twitter accounts and firms without such accounts (Liu et al., 2015).

Moving to researches applying sentiment analysis techniques to newspapers, among the first works, Tetlock (2007) and Tetlock, Saar-Tsechansky, and Macskassy (2008) quantified the "*interactions between the media and the stock market*" (Tetlock, 2007) from Wall Street Journal articles. Besides, more recent works focused on: the effects of news about macroeconomic fundamentals and geopolitical events on crude oil markets (Brandt and Gao, 2019); the impacts of news stories and earnings announcements of the S&P 100 constituents (Caporin and Poli, 2017); Taiwan stock market returns (Guan-Ru Wu, Chieh-Tse Hou, and Lin, 2019); the differences of positive and negative news, as well as daily or weekly aggregation, in improving returns predictions (Heston and Sinha, 2017); the use of existing "*affective lexicons*" such as the Harvard psychological dictionary and Loughran–McDonald sentiment dictionary (Loughran and McDonald, 2011) to construct a sentiment space (Li et al., 2014); the analysis of oil-related news and creation of a specific keywords list (Loughran, McDonald, and Pragidis, 2019); credit spreads, and information asymmetry (Yang, Liu, and Wang, 2020).

Concerning the second research area, recently, the analysis of Central Banks communication has been studied in Ter Ellen, Larsen, and Thorsrud (2022) who, based on Norwegian data, focused on the effects of "*narrative surprise*" of "*central bank communication accompanying interest rate meetings*". Ehrmann and Talmi (2020), instead, studied the effects of semantic similarity in Central Bank communications on market volatility. Besides, Tillmann and Walter (2019) focused on the effects of diverging tones in the public communication ECB and Bundesbank presidents.

Finally, among the works within the third area of research, which is concerned with the application of sentiment analysis to various aspects of the economy, recently, Aprigliano et al. (2022) focused on the Italian economy and proposed a new dictionary of Italian words and used it to construct sentiment and uncertainty indices based on newspaper articles. Based on a dictionary approach (and combined with a clustering technique) is also the *Social Mood on Economy Index* (SMEI) [1] recently constructed by the Italian National Institute of Statistics (ISTAT) in 2016 to measure the sentiment of Italians with respect to the economy.

Besides, Ardia, Bluteau, and Boudt (2019) focused on US industrial production and proposed a "*sentiment engineering framework*" based on elastic net and optimised for forecasting purposes. Bowden, Kwiatkowski, and Rambaccussing (2019) developed "*time series of economic news tonality*" from major British newspapers. Mogaji, Balakrishnan, and Kieu (2021), instead, focused their analysis on the UK energy sector, based on twits. Naderi Semiromi, Lessmann, and Peters (2020), on the other hand, focused on the foreign exchange market and proposed a novel sentiment dictionary for prediction purposes. Shapiro, Sudhof, and Wilson (2020) also develop time series of economic sentiment from newspapers. Specifically, while reviewing the current state-of-the-art sentiment analysis tools, the authors generate their sentiment-scoring model and a new lexicon.

Moving to political science, several works in the literature apply sentiment analysis not only to newspaper articles and social media but also to parliamentary debates. For instance, Young and Soroka (2012) develop their "*Lexicoder Sentiment Dictionary (LSD)*" comparing automated tone detection against human readers' judgement of articles about the economy, environment, crime, and international affairs in the New York Times. Besides, Haselmayer and Jenny (2017) propose a sentiment measurement procedure that applies to languages other than English and build their German dictionary "*for the analyses of party statements and media reports*". Among researches based on parliamentary debates, recently, Proksch et al. (2019) applied a "*multilingual sentiment-based approach*" to "*capture different types of parliamentary conflict*". Rauh (2018) also developed their German dictionary and find "*that positive language is easier to detect than negative language*". Rheault et al. (2016), instead, used an NLP *approach to examine changes in aggregate levels of emotional polarity in the British parliament and to test a hypothesis about the emotional response of politicians to economic recessions*. Finally, Abercrombie and Batista-Navarro (2020) recently provided a detailed and extensive survey of and researches applying sentiment analysis to parliamentary debates in several languages.

Concerning the methodologies for sentiment analysis, most of the studies in the field fall in one of three macro-categories: *Lexicon Based approaches*, *Machine Learning approaches* and *Hybrid approaches* (Birjali, Kasri, and Beni-Hssane, 2021). In the following the former two are briefly described.

Lexicon based approaches are probably the earliest methods due to their relatively easier application in many fields. According to Algaba et al. (2020), "*a lexicon-based computation of sentiment is the most straightforward, efficient, and parsimonious method*". In fact, such approaches use *affective lexicons*: lists of words or phrases annotated with a positive or negative polarity scores. Examples in the

[1] https://www.istat.it/it/archivio/219585

economic/financial and political fields are given by: Picault and Renault (2017) who construct a lexicon specific to European Central Bank communication; Young and Soroka (2012) whose *Lexicoder Sentiment Dictionary* focuses on news about politics; Loughran and McDonald (2011) who create a lexicon specific for finance domain. Birjali, Kasri, and Beni-Hssane (2021), in accordance with previous literature (Algaba et al., 2020) divide the techniques under this framework in three subcategories: *manual approaches*, *dictionary-based approaches* and *corpus-based approaches*. One of the most used hand-crafted dictionary is certainly the *General Inquirer* of Stone and Hunt (1963). "*Dictionary-based approaches [...] start from a list of seed sentiment words with known polarity (often found using the manual approach) and then expands this list*" (Algaba et al., 2020). "*The assumption behind this approach is that synonymous words have the same sentiment polarities, while antonyms words have the opposite polarities*" (Birjali, Kasri, and Beni-Hssane, 2021). Among others, a recent example is the *SentiWordNet* of Baccianella, Esuli, and Sebastiani (2010). Finally, corpus based approach methods "*adapt an existing lexicon using information from a domain-specific corpus*" (Algaba et al., 2020). Specifically, they "*start with a list of seed sentiment words with pre-known orientation and exploit syntactic or co-occurrence patterns to find new sentiment words with their orientation in a large corpus*" (Birjali, Kasri, and Beni-Hssane, 2021) relying on the so called *sentiment consistency* hypothesis.

As noted by Algaba et al. (2020) and Birjali, Kasri, and Beni-Hssane (2021), dictionary or lexicon based methods are attractive because they are not computationally expensive. Although compiling dependency rules – e.g. to consider valence switchers/shifters such as negations and or intensifiers adjective, adverbs or conjunctions – they "*represent a good strategy to easily and quickly build a lexicon with a large number of sentiment words and their orientation*". On the other hand, sentiment scores associated with such lists do not perform so well in context or domain-specific classification tasks.

Moving to the second category, *machine learning* approaches use statistical algorithms to classify words or phrases sentiment based on training and test data (Birjali, Kasri, and Beni-Hssane, 2021). The performances of such methods are usually evaluated in comparison with annotated corpora (Algaba et al., 2020). Following precedent literature, Algaba et al. (2020) divide machine learning methods into two sub-categories: *supervised*, *unsupervised*. Birjali, Kasri, and Beni-Hssane (2021), who provide an extensive survey on sentiment analysis approaches, add also the subcategories of *semi-supervised* and *reinforcement* learning.

Supervised learning requires already labelled texts. They represent the most used algorithms thus far in sentiment analysis and include linear and probabilistic approaches among others (Birjali, Kasri, and Beni-Hssane, 2021). The linear approach includes techniques like Support Vector Machine and Artificial Neural Networks. Given an input, such algorithms predict the most likely polarity – e.g. either positive or negative. On the other hand, probabilistic approaches predict a probability distribution instead, usually basing on Bayes' theorem. They include: Naive Bayes, Bayesian Networks and Maximum Entropy.

Unsupervised learning include hierarchical and partition methods. Unlike supervised learning, it is used to cluster unlabelled data usually basing on "*statistical properties such as word co-occurrence, NLP processes, and existing lexicons with emotional (or) polarised words*" (Birjali, Kasri, and Beni-Hssane, 2021). The main drawback of such algorithms is that they output clusters which are difficult to interpret as it is not straightforward their association with a specific sentiment (Birjali, Kasri, and Beni-Hssane, 2021).

Recently, another vast area has emerged and is becoming prominent also in sentiment analysis: *Deep Learning*. On the topic a recent and rather detailed review is provided by Yadav and Vishwakarma (2020). Specifically, deep learning refers to *Deep Neural Networks* (DNN) – "*artificial neural networks having multiple hidden layers between the input layer and the output layer*" (Yadav and Vishwakarma, 2020) – and its variants: *Convolutional Neural Networks* (CNN), *Recurrent Neural Networks* (RNN) and *Deep Belief Networks* (DBN). The main attractive feature of this models is that *they can learn sophisticated features from the dataset by themselves*. More in detail, DNN (Deep Neural Networks) are easy to implement than other deep learning models and require less training time, CNN (Convolutional Neural Networks) provide accuracy and fast training, RNN (Recurrent Neural Networks) are able to capture sequential data for classification. However, such architectures often suffer from over-fitting, difficulties in tuning and require large amount of data to be effectively trained. Moreover, even when training is relatively faster they are very complex models resulting essentially in "*black boxes*" (Birjali, Kasri, and Beni-Hssane, 2021), as usually it is not possible to know *why* a specific output is produced. Further details about the functioning of each algorithm are provided by Yadav and Vishwakarma (2020).

3.3 Textual data

3.3.1 The Italian Senate verbatim reports

The proposed text-based indices aim at quantifying the effect on the economy of the political debate in Legislative Assemblies. Hence, the speeches of *Representatives* in such Assemblies are essential.

Specifically, in this work the proposed indices are based on the entire collection of the *Italian Senate of the Republic* parliamentary verbatim reports - i.e. the *corpus* or *texts sample* - from their first issue on 08 May 1948 to 10 March 2021 - i.e. the *sample period*.

Reports and Senate sessions are in a *one-to-one* relation. Although the Senate meets on no regular basis, reports are almost daily as, usually, sittings are held three to four times per week over the entire period.

Table 3.1 provides information about the number of reports, the Governments and the Legislatures throughout the entire sample period.

TABLE 3.1: Italian Legislatures, Governments and Senate verbatim reports details

Legislature	*Legislature Dates*	*Duration (days)*	*Prime Minister*	*Government Dates*	*Time in Office (days)*	*Senate Reports*
XVIII	23 Mar 2018 ongoing [a]	1083 [a]	Draghi [a]	13 Feb 2021 ongoing [a]	26 [a]	305 [a]
			Conte II	5 Sep 2019 13 Feb 2021	527	
			Conte I	1 June 2018 5 Sep 2019	461	
XVII	15 Mar 2013 22 Mar 2018	1834	Gentiloni	12 Dec 2016 1 June 2018	536	923
			Renzi	22 Feb 2014 12 Dec 2016	1024	
			Letta	28 Apr 2013 22 Feb 2014	300	
XVI	29 Apr 2008 14 Mar 2013	1781	Monti	16 Nov 2011 28 Apr 2013	529	860
			Berlusconi IV	8 May 2008 16 Nov 2011	1287	
XV	28 Apr 2006 28 Apr 2008	732	Prodi II	17 May 2006 8 May 2008	722	283
XIV	30 May 2001 27 Apr 2006	1794	Berlusconi III	23 Apr 2005 17 May 2006	389	965
			Berlusconi II	11 June 2001 23 Apr 2005	1412	
XIII	9 May 1996 29 May 2001	1847	Amato II	26 Apr 2000 11 June 2001	411	1061
			D'Alema II	22 Dec 1999 26 Apr 2000	126	
			D'Alema I	21 Oct 1998 22 Dec 1999	427	
			Prodi I	18 May 1996 21 Oct 1998	886	
XII	15 Apr 1994 8 May 1996	755	Dini	17 Jan 1995 18 May 1996	487	310
			Berlusconi I	11 May 1994 17 Jan 1995	251	
XI	23 Apr 1992 14 Apr 1994	722	Ciampi	29 Apr 1993 11 May 1994	377	287
			Amato I	28 June 1992 29 Apr 1993	305	

continued on next page . . .

Table 3.1: *. . . continued from previous page*

Legislature	*Legislature Dates*	*Duration (days)*	*Prime Minister*	*Government Dates*	*Time in Office (days)*	*Senate Reports*
X	2 Jul 1987 22 Apr 1992	1757	Andreotti VII	13 Apr 1991 28 June 1992	442	665
			Andreotti VI	23 Jul 1989 13 Apr 1991	629	
			De Mita	13 Apr 1988 23 Jul 1989	466	
			Goria	29 Jul 1987 13 Apr 1988	259	
IX	12 Jul 1983 1 Jul 1987	1451	Fanfani VI	18 Apr 1987 29 Jul 1987	102	597
			Craxi II	1 Aug 1986 18 Apr 1987	260	
			Craxi I	4 Aug 1983 1 Aug 1986	1093	
VIII	20 June 1979 11 Jul 1983	1483	Fanfani V	1 Dec 1982 4 Aug 1983	246	617
			Spadolini II	23 Aug 1982 1 Dec 1982	100	
			Spadolini I	28 June 1981 23 Aug 1982	421	
			Forlani	18 Oct 1980 28 June 1981	253	
			Cossiga II	4 Apr 1980 18 Oct 1980	197	
			Cossiga I	5 Aug 1979 4 Apr 1980	243	
VII	5 Jul 1976 19 June 1979	1080	Andreotti V	21 Mar 1979 5 Aug 1979	137	395
			Andreotti IV	13 Mar 1978 21 Mar 1979	373	
			Andreotti III	30 Jul 1976 13 Mar 1978	591	
VI	25 May 1972 4 Jul 1976	1502	Moro V	12 Feb 1976 30 Jul 1976	169	572
			Moro IV	23 Nov 1974 12 Feb 1976	446	
			Rumor V	15 Mar 1974 23 Nov 1974	253	
			Rumor IV	8 Jul 1973 15 Mar 1974	250	
			Andreotti II	26 June 1972 8 Jul 1973	377	

continued on next page . . .

Table 3.1: …*continued from previous page*

Legislature	*Legislature Dates*	*Duration (days)*	*Prime Minister*	*Government Dates*	*Time in Office (days)*	*Senate Reports*
V	5 June 1968 24 May 1972	1450	Andreotti I	18 Feb 1972 26 June 1972	129	597
			Colombo	6 Aug 1970 18 Feb 1972	561	
			Rumor III	28 Mar 1970 6 Aug 1970	131	
			Rumor II	6 Aug 1969 28 Mar 1970	234	
			Rumor I	13 Dec 1968 6 Aug 1969	236	
			Leone II	25 June 1968 13 Dec 1968	171	
IV	16 May 1963 4 June 1968	1847	Moro III	24 Feb 1966 25 June 1968	852	804
			Moro II	23 Jul 1964 24 Feb 1966	581	
			Moro I	5 Dec 1963 23 Jul 1964	231	
			Leone I	22 June 1963 5 Dec 1963	166	
III	12 June 1958 15 May 1963	1799	Fanfani IV	22 Feb 1962 22 June 1963	485	697
			Fanfani III	27 Jul 1960 22 Feb 1962	575	
			Tambroni	26 Mar 1960 27 Jul 1960	123	
			Segni II	16 Feb 1959 26 Mar 1960	404	
			Fanfani II	2 Jul 1958 16 Feb 1959	229	
II	25 June 1953 11 June 1958	1813	Zoli	20 May 1957 2 Jul 1958	408	653
			Segni I	6 Jul 1955 20 May 1957	684	
			Scelba	10 Feb 1954 6 Jul 1955	511	
			Fanfani I	19 Jan 1954 10 Feb 1954	22	
			Pella	17 Aug 1953 19 Jan 1954	155	
			De Gasperi VIII	16 Jul 1953 17 Aug 1953	32	

continued on next page …

Table 3.1: …*continued from previous page*

Legislature	*Legislature Dates*	*Duration (days)*	*Prime Minister*	*Government Dates*	*Time in Office (days)*	*Senate Reports*
I	8 May 1948 24 June 1953	1874	De Gasperi VII	26 Jul 1951 16 Jul 1953	721	984
			De Gasperi VI	27 Jan 1950 26 Jul 1951	545	
			De Gasperi V	24 May 1948 27 Jan 1950	613	
		Total				*Total*
		26819				11575

[a] As at 10 Mar 2021

3.3.2 Texts pre-processing and data-frame creation

Analysing texts requires them to be arranged in a suitable format as, for instance, a data frame.

This section describes the steps to convert and organise the raw .pdf verbatim reports into the data frame used for the analysis, starting from the following initial set of information:

- the verbatim reports pdfs;
- a list of all legislatures and sessions/reports dates;
- a complete list of Senators and Government members' surnames and names throughout the sample period;
- a complete list of all Senators' political party belonging with the indication of the date each Senator changed their party affiliation;
- a list of all Governments with the respective supporting or opposing parties at each time of their office;
- a list of the most common male and female Italian names.

Information other than reports is needed to assign each text also a set of additional details called *metadata*.

The first operations performed on the .pdf documents are:

- initial inspection for documents structure understanding;
- creation of three groups of reports with homogeneous typographic characteristics and structure: legislatures 1–9, 10–12, 13–18;
- checking for .pdf files stored as images and performing OCR when needed;

- .pdf files extraction and conversion in .txt format.

Each line of the converted texts usually contains the speech of an orator. However, as this is not always true, some refinements are needed. Specifically, by mean of *regular expressions*, for each one of the three groups of reports, the following operations are performed:

- speakers names and surnames correctness double-checking and correction of accented letters and apostrophes in surnames when needed;
- removal of punctuation and spaces at the beginning of each string;
- removal of all lines formed by just a single character, numbers or punctuation only;
- correction of words divided in single characters (e.g. *g o v e r n m e n t* instead of *government*);
- correction of words "*broken*" at line ends (e.g. *govern- ment* instead of *government*);
- removal of reports title pages, headings, indices, attachments (which, usually, are the texts of the discussed bills and proposals) and similar items;
- lowercase conversion of the titles of regulatory provisions (e.g. *ACT XXX*, *ART. 123*, ...) and abbreviation of national or international institutions (e.g. NATO, EEC, ...) when appearing at the beginning of a line [2] and of all capital characters words other than orator's names and surnames;
- splitting of texts in correspondence of each orator (each of these text fragments constitutes a row in the column named *text* in the raw data-frame);
- inclusion of the following texts metadata (other columns of the raw data-frame): *session date*, *legislature*; *session number*, *text* fragment *number* and *Government* in office at session date.

3.3.3 Data-frame refining

Obtaining the final data frame with all the texts and associated metadata requires the following operations:

- removal of regular expressions for sitting opening and closing,

[2]In this framework, a line always begins with the orator's surname (or surname and name in case of homonyms) in capital letters. The only exception is for the session *PRESIDENT*.

- extraction and correction of the metadata "*orator*". Specifically, the initial part of each text string [3] in the raw data-frame is cleaned of eventual punctuation and compared against surnames and names within the Senate and Government members' list. In case of typographic errors or mismatches due to pdf conversion into plain text, the wrong orator's name is replaced with the correct one. When more than one option is available from the Senate and Government members' list, the correction showing the minimum edit distance [4] is chosen. In the rare cases where more than one minimum edit distance correct suggestion is available, the most frequent option is chosen. Finally, the metadata are added as a column to the data frame.

- association of each orator with the "*Parliamentary Group*" they belong to the day the speech is pronounced [5],

- aggregation of Parliamentary Groups into the four "*political groups*" used for further analysis: *Government*, *Majority*, *Mixed* and *Opposition*. The *Government* group includes all those orators who are Government members [6] at the time they speak. *Majority* and *Opposition* groups aggregate all Senators who belong - at the time they spoke - to Parliamentary Groups (or political parties) which either support the Government or not, respectively. Finally, the *Mixed* group aggregates all orators belonging to the Parliamentary "*Mixed Group*" at the time they spoke. Political groups are added as metadata to the data-frame in a column named "*collocation*".

- removal of all speeches corresponding to the orator "*PRESIDENT*". They are the *Senate President* or the *Vice-President* who is chairing the session, therefore they do not actively participate in the debate as their role is more administrative and procedural (e.g. giving the floor or a roll-call vote, ...)

3.3.4 Text cleaning and stemming

After creating and refining the speeches data frame, to reduce corpus complexity, some standard text cleaning (Denny and Spirling, 2018) is performed. Specifically,

[3]More in detail, the first n characters of each line, with n being the number of letters forming the longest surname (or surname and name) in the Senate and Government members' list (including also the word *PRESIDENT*).

[4]The Levenshtein edit distance, precisely.

[5]In the Italian system, all Senators (as well as Deputies) must join a "*Parliamentary Group*" - which, indeed, represents a political party. Groups must be composed of at least ten members, although some exceptions are allowed. Senators who do not join any Group form to the so-called "*Mixed Group*". Groups composition may vary over time as one, more, or even all members decide to enter (or, also, constitute) another Group or join the "*Mixed Group*". Furthermore, also Groups names may vary over time independently of any changes in their internal composition or even in their decision to support the Government or not.

[6]Actually, the *Government* group also contains some other orators who are not strictly Government members. For instance, it is the case of the *President of the Republic* in rare occasions, or maybe high officers of Ministries illustrating technical matters (this also a rare case).

the texts are lower-cased and tokenized - i.e. texts strings are split in words (tokens) - and all punctuation, symbols, URLs and numbers [7] are removed.

Accounting for the peculiarities of the Senate reports requires some other text processing. In detail, such further steps are:

- removal of all tokens formed by less than three letters, with the exception of "*o*", "*sì*", "*no*" and "*ma*" - i.e. "*or*", "*yes*", "*no/not*" and "*but*", respectively - as such words may change the meaning of a sentence,
- removal of all names and surnames of Senators and Government members as well as Italian common names as listed in Appendix B,
- removal of the "*stop-words*" [7] contained in "*ad hoc*" list in Appendix A,
- removal of words formed by more than twenty characters,
- spell-checking and correction of misspelled words. In detail, each word is compared with the words in an Italian dictionary [8] and is substituted with the closest suggestion in terms of the Levenshtein edit distance [9]. When more suggestions are available, the option chosen is the most frequent in the corpus.

After pre-processing, tokenization and cleaning, in order to be used for further analysis, speeches need a last common NLP routine: text "*stemming*" - i.e. the reduction of each word to its base form (e.g. the words *governance* and *government* both reduce to the stem *government*). The procedure is highly beneficial as it allows significantly reducing the vocabulary size.

3.4 Annual polarity indices

This section describes the procedure to construct the proposed annual [10] polarity indices. The first and most significant step in the procedure is determining the polarity of each word in the processed corpus.

Such a goal is achieved in an innovative way that requires information regarding the state of the economy. For this reason, in addition to texts, there is the need for a macroeconomic indicator time-series.

Differently from common practice in literature, word polarity - also "*tone*" or "*sentiment*" in literature - does not rely on the subjective choice of a so called "*affective lexicon*" - i.e. a lexicon where each word is assigned a positive or negative score depending on its actual meaning in the language at hand. Here word polarity is learned in an entirely data-driven manner.

[7] Numbers and stop-words removal is done with "*padding*", i.e. by replacing the token with a space.

[8] Specifically, the *Hunspell* dictionary (`https://hunspell.github.io/`).

[9] The maximum tolerated edit distance is equal to three.

[10] Depending on the contest and data availability, the same technique can be applied to obtain text-based polarity indices at any desired frequency - e.g. quarterly, monthly. . . .

To this aim, original sample years (S) $t = 1, \dots, T = 73$ texts and macroeconomic variable - i.e. from 1948 to 2020 [11] - are split as follows:

- years $t = 1, \dots, M = 43$ (1948–1990) as "*training sample*" (S_1) for the actual words polarities computation;
- years $t = M + 1, \dots, T$ (1991–2020) as "*test sample*" (S_2) to build and evaluate "*out-of-sample*" time-series of polarity indices.

Sample splitting ensures that no data used for polarity calculations are employed when testing the indices' time-series validity. Training and test samples sizes of respectively 60% and 40% of the original sample size appeared to be a balanced choice providing enough data for both procedures: assigning words polarity and indices construction and evaluation.

3.4.1 The Document Term Matrix

To determine words polarities, the texts in the constructed data frame need to be rearranged in matrix form. Therefore, the following operations are performed on texts in the training sample:

1. Clean speeches delivered by each orator are arranged in chronological order to form the row entries of a *Document Term Matrix* (DTM). On columns - are the single words or terms in the corpus vocabulary. The DTM structure is clarified in Equation 3.1:

$$\underset{D\times V}{\boldsymbol{C}} = \begin{bmatrix} c_{1_1,1} & c_{1_1,2} & \cdots & c_{1_1,V} \\ \vdots & \vdots & c_{d_1,v} & \vdots \\ c_{D_1,1} & c_{D_1,2} & \cdots & c_{D_1,V} \\ c_{1_2,1} & c_{1_2,2} & \cdots & c_{1_2,V} \\ \vdots & \vdots & c_{d_t,v} & \vdots \\ c_{1_M,1} & c_{1_M,2} & \cdots & c_{1_M,V} \\ \vdots & \vdots & c_{d_M,v} & \vdots \\ c_{D_M,1} & c_{D_M,2} & \cdots & c_{D_M,V} \end{bmatrix} \tag{3.1}$$

[11] Even though in 1948 verbatim reports are only available starting from May, 5th, this year is not excluded from the analysis as the corresponding yearly macroeconomic variable data point is available. On the other hand, the text of all 2021 reports is excluded from the analysis as the last yearly economic data point available is for 2020.

where

$$
\begin{aligned}
D &= \textit{total number of documents in the training sample}\\
V &= \textit{total number of words in the training sample}\\
D &= \sum_{t=1}^{M} D_t \qquad D_t = \textit{number of documents at each time } t\\
c_{d_t,v} &= \textit{count of word } v \textit{ in document } d_t \textit{ at time } t\\
d_t &= 1,\dots,D_t \qquad t = 1,\dots,M \qquad v = 1,\dots,V
\end{aligned}
$$

The matrix in Equation 3.1 is built at time M. Therefore it has a fixed number of rows and columns given by the *corpus size* (D) and the *vocabulary size* (V) at that time. Moreover, it is a high dimensional and sparse matrix. As $c_{d_t,v}$ may range from 0 to any positive integer, the DTM is also a full matrix: it contains no missing values. Specifically, $c_{d_t,v} = 0$ when the word v is not used in document d_t. Less commonly, $c_{d_t,v} = 0$ also in case v did not exist (yet) before a specific time $t < M$. An example of such DTM is given in Table 3.2.

TABLE 3.2: Matrix $\boldsymbol{C}$ Example

		v_1	v_2	v_3	v_4	v_5
t_1	d_1	2	0	2	0	0
t_1	d_2	0	0	0	2	0
t_1	d_3	2	0	2	2	0
t_2	d_1	0	0	0	0	0
t_2	d_2	2	0	2	0	0
t_3	d_1	0	2	2	0	2
t_3	d_2	2	0	3	0	3
t_3	d_3	2	0	4	0	0
t_3	d_4	0	0	0	0	0
t_4	d_1	2	0	2	0	0
t_4	d_2	0	0	4	0	0
t_4	d_3	2	2	0	0	0

2. The original DTM in Equation 3.1 is aggregated by summing together the frequencies (counts) of the words appeared at each time $t = 1,\dots,M$. The resulting $\widetilde{\boldsymbol{C}}$ is a matrix with the same number of columns (the vocabulary size) but with a number of rows equal to the number of years within the in-sample period (M). The formula for this aggregated matrix is given in Equation 3.2:

$$
\underset{M\times V}{\widetilde{\boldsymbol{C}}} = \begin{bmatrix} \widetilde{c}_{1,1} & \widetilde{c}_{1,2} & \dots & \widetilde{c}_{1,V} \\ \vdots & \vdots & \widetilde{c}_{t,v} & \vdots \\ \widetilde{c}_{M,1} & \widetilde{c}_{M,2} & \dots & \widetilde{c}_{M,V} \end{bmatrix} \tag{3.2}
$$

where

$$\widetilde{c}_{t,v} = \sum_{d_t=1}^{D_t} c_{d_t,v} \qquad t = 1, \ldots, M \qquad v = 1, \ldots, V$$

Following the Table example in 3.2, the corresponding $\widetilde{\boldsymbol{C}}$ matrix is reported in Table 3.3.

TABLE 3.3: Matrix $\widetilde{\boldsymbol{C}}$ Example

	v_1	v_2	v_3	v_4	v_5
t_1	4	0	4	4	0
t_2	2	0	2	0	0
t_3	4	2	9	0	5
t_4	4	2	6	0	0

3. To put more weight on the most relevant words at each time, the *tf-idf* transformation (*term frequency-inverse document frequency*) is applied to the aggregated frequencies in the $\widetilde{\boldsymbol{C}}$. In detail, let x be a word in a corpus, its *term frequency-inverse document frequency*, *tf-idf*(x), is given by:

$$tf\text{-}idf(x) = tf(x) \times idf(x) \tag{3.3}$$

where

$$\begin{aligned} tf(x) &= \textit{relative frequency of } x \\ idf(x) &= \log \frac{\textit{total number of documents}}{\textit{number of documents containing } x} \\ &= \textit{inverse document frequency of } x \end{aligned}$$

Following the example in Table 3.3, the corresponding *tf*, *idf* matrices are reported in Tables 3.4, 3.5.

TABLE 3.4: Matrix $\widetilde{\boldsymbol{C}}$ *tf* Example

	v_1	v_2	v_3	v_4	v_5
t_1	0.29	0	0.19	1	0
t_2	0.14	0	0.10	0	0
t_3	0.29	0.50	0.43	0	1
t_4	0.29	0.50	0.29	0	0

The resulting matrix $\boldsymbol{F}$ has the same dimensions of the $\widetilde{\boldsymbol{C}}$. In formulae:

TABLE 3.5: Matrix $\widetilde{C}$ *idf* Example

	v_1	v_2	v_3	v_4	v_5
t_1	0	0.3	0	0.6	0.6
t_2	0	0.3	0	0.6	0.6
t_3	0	0.3	0	0.6	0.6
t_4	0	0.3	0	0.6	0.6

$$\underset{M \times V}{\boldsymbol{F}} = \begin{bmatrix} f_{1,1} & f_{1,2} & \dots & f_{1,V} \\ \vdots & \vdots & f_{t,v} & \vdots \\ f_{M,1} & f_{M,2} & \dots & f_{M,V} \end{bmatrix} \tag{3.4}$$

where $f_{t,v}$ is the *weighted frequency of word v in document t*.

Finally, each row in matrix $\boldsymbol{F}$ is normalised (Sebastiani, 2002) with its *L2-norm* so to obtain:

$$\underset{M \times V}{\widetilde{\boldsymbol{F}}} = \begin{bmatrix} \widetilde{f}_{1,1} & \widetilde{f}_{1,2} & \dots & \widetilde{f}_{1,V} \\ \vdots & \vdots & \widetilde{f}_{t,v} & \vdots \\ \widetilde{f}_{M,1} & \widetilde{f}_{M,2} & \dots & \widetilde{f}_{M,V} \end{bmatrix} \tag{3.5}$$

where

$$\widetilde{f}_{t,v} = \frac{\widetilde{f}_{t,v}}{\sqrt{\sum_{v=1}^{V} f_{t,v}^2}} \qquad t = 1, \dots, M \qquad v = 1, \dots, V$$

Following the example in Table 3.3, the corresponding *tf-idf* weighted matrices are reported in Tables 3.6, 3.7.

TABLE 3.6: Matrix $\boldsymbol{F}$ Example

	$v1$	$v2$	$v3$	$v4$	$v5$
$t1$	0	0	0	0.6	0
$t2$	0	0	0	0	0
$t3$	0	0.15	0	0	0.6
$t4$	0	0.15	0	0	0

TABLE 3.7: Matrix $\widetilde{\boldsymbol{F}}$ Example

	$v1$	$v2$	$v3$	$v4$	$v5$
$t1$	0	0	0	1	0
$t2$	0	0	0	0	0
$t3$	0	0.24	0	0	0.97
$t4$	0	1	0	0	0

3.4.2 The macroeconomic time-series

As previously mentioned, determining words polarities also requires information regarding the state of the economy. Traditional economic indicators provide such information. Hence, the following lines describe their processing.

Let $\{y_t\}_{t=0}^{M}$ be a time-series of a macroeconomic variable. To proceed with the polarity calculations detailed in the following sections, y_t values are arranged in a $(M+1)$-dimensional vector $\boldsymbol{y}$ as in Equation 3.6:

$$\underset{(M+1)\times 1}{\boldsymbol{y}} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_M \end{bmatrix} \tag{3.6}$$

For the purposes of this work $\{y_t\}_{t=0}^{M}$ is the Italian yearly gross domestic product (GDP) time-series described in Section 3.4.2 below.

Depending on the contest, any other macroeconomic indicator (public expenditure, taxation, consumption, ...) may be used.

As the main aspect of interest of the state of the economy is its temporal dynamic, and because the chosen GDP time-series is *non-stationary*, the series of its *year-on-year* growth rates $\{\Delta y_t\}_{t=1}^{M}$ is calculated as in Equation 3.7:

$$\underset{M\times 1}{\boldsymbol{\Delta y}} = \begin{bmatrix} \Delta y_1 \\ \vdots \\ \Delta y_M \end{bmatrix} \qquad \Delta y_t = \frac{y_t - y_{t-1}}{y_{t-1}} \qquad t = 1, \ldots, M \tag{3.7}$$

To perform the word polarity assigning procedure requires the derivation, $\{sgn\Delta y_t\}_{t=1}^{M}$, of such growth rates signs as described in Equation 3.8:

$$\underset{M\times 1}{\boldsymbol{sgn\Delta y}} = \begin{bmatrix} sgn\Delta y_1 \\ \vdots \\ sgn\Delta y_M \end{bmatrix} \qquad sgn\Delta y_t = \begin{cases} 1 & if\ \Delta y_t > 0 \\ 0 & if\ \Delta y_t = 0 \\ -1 & if\ \Delta y_t < 0 \end{cases} \tag{3.8}$$

$$t = 1, \ldots, M$$

The Italian yearly GDP time-series

As mentioned above, the Italian GDP (better, its year-on-year growth rate) is the macroeconomic indicator chosen to evaluate the possible link between the Italian parliamentary debate and economic dynamics.

However, as there are no GDP time-series spanning the same period as the verbatim reports collection at hand, a custom Italian GDP time-series is created starting from two already available series.

More in detail, the first of these two existing series is the GDP time-series reconstructed by *ISTAT* in collaboration with *Bank of Italy* [12]. It starts in 1861 (the year of the *Unification of Italy*) and ends in 2017. The second series employed is the GDP time-series as made available by *ISTAT* in its National Accounts database. It starts in the year 1995 and ends in 2020.

Both series are measured in millions of euros at current market prices and are compiled by the sources following the *System of National Accounts, 2008* (*2008 SNA*) as implemented in the *European National Accounts 2010* (ESA 2010) currently in use.

To have a unique series matching the reports period (1948–2020), years 1948–2017 are taken from the first series while the second series is used entirely. For years 1995–2017 where two GDP values are available, a weighted average is calculated at each time as in Equation 3.9:

$$GDP_{average,t} = (1 - w_t)GDP_{BI\text{–}ISTAT,t} + w_t GDP_{ISTAT,t} \tag{3.9}$$

where

$$\begin{aligned} t &= 1995, \ldots, 2017 \\ GDP_{BI\text{–}ISTAT} &= \textit{Bank of Italy and ISTAT Italian GDP time-series} \\ GDP_{ISTAT} &= \textit{ISTAT Italian GDP time-series} \end{aligned}$$

and

$$w_t = \begin{cases} 0 & \text{if } t = 1995 \\ w_{t-1} + \frac{1}{2017-1995+1} & \text{if } t = 1996, \ldots, 2017 \\ 1 & \text{if } t = 2017 \end{cases}$$

Therefore, considering the entire sample period, the final GDP series is composed by three blocks as in Equation 3.10:

$$GDP_t = \begin{cases} GDP_{BI\text{–}ISTAT,t} & \text{if } t = 1948, \ldots, 1994 \\ GDP_{average,t} & \text{if } t = 1995, \ldots, 2016 \\ GDP_{ISTAT,t} & \text{if } t = 2017, \ldots, 2020 \end{cases} \tag{3.10}$$

The final current market prices GDP time-series in Equation 3.10 is then converted in a fixed 2000 market prices time-series using specific multipliers made available by *ISTAT* [13].

[12] The series is available at https://www.bancaditalia.it/statistiche/tematiche/stat-storiche/stat-storiche-economia/NA150_2.0.zip or at http://seriestoriche.istat.it/fileadmin/documenti/Serie%20Storiche%20della%20Contabilit%C3%A0%20nazionale%201861-2017.zip.

[13] See https://www.istat.it/it/archivio/258610.

3.4.3 Determining word polarities

After arranging the speeches and macroeconomic variable in a suitable matrix form, the actual word polarities calculations can be performed.

They are detailed in the following steps:

1. The elements in the *tf-idf* weighted matrix in Equation 3.4 are multiplied by the growth rates signs in Equation 3.8.

   This multiplication may be performed in several ways. For instance, the growth rates sign available at each time t may be assigned to the speeches pronounced at that specific time. In this case a "*contemporaneous lag configuration*" ($B_{0,0}$) would be adopted.

   On the other hand, it is also possible to assign the growth rate sign at each time t to the speeches delivered at time $t-1$. This would be referred to as a "*lag one configuration*" ($B_{1,1}$).

   Another option might well be assigning the growth rate sign at each time t to the texts at both times t and $t-1$. Therefore, a "*lag zero-one configuration*" ($B_{0,1}$) would be adopted [14].

   Of course, these are not the only possible configurations because, as seen, the amount of text to assign the growth rates signs to is a subjective choice.

   Being more specific, determining words polarities multiplying the elements in the $\boldsymbol{F}$ matrix in Equation 3.4 by the growth rates signs in the $\boldsymbol{sgn\Delta y}$ vector in Equation 3.8 requires setting the "*textual window size*", W, defined as follows:

   $$W = \textit{number of consecutive rows in the TF-IDF-S-DFM} \tag{3.11}$$

   After setting W it is possible to define the $W \times V$ "*textual window*" matrix $\boldsymbol{W_t}$. This is a sub-matrix of the $\boldsymbol{F}$ matrix in Equation 3.4. Specifically, it has the same columns but its specific rows depend on the "*lag configuration*" applied.

   More in detail, under the generic textual window configuration $B_{k,h}$ - where h and k are, respectively, the textual window starting and ending lags and $W = h + k + 1$ - the general equation of $\boldsymbol{W_t}$ is given by:

   $$\boldsymbol{W}_t^{(B_{k,h})} = \begin{bmatrix} \widetilde{f}_{t-h,1} & \cdots & \widetilde{f}_{t-h,V} \\ \vdots & \cdots & \vdots \\ \widetilde{f}_{t-k,1} & \cdots & \widetilde{f}_{t-k,V} \end{bmatrix} \tag{3.12}$$

[14] For indicating the textual window configuration the letter "*B*" is chosen for analogy with the *Backshift Operator*, B, such that $B^k X_t = X_{t-k}$ where $X = \{X_1, X_2, \dots\}$ is a given time-series and $t > 1$.

where,

$$t = \begin{cases} h+1,\ldots,M & \textit{if } k = 0 \\ h,\ldots,M-k & \textit{if } k \neq 0 \textit{ and } M \textit{ is even} \\ h-1,\ldots,M-k & \textit{if } k \neq 0 \textit{ and } M \textit{ is odd} \end{cases}$$

$$h \leq k = 0,\ldots,M-1$$

The choice of the textual window size and configuration is crucial for determining words polarities. It allows controlling for the circumstance that it might take some time before what is discussed in the Parliament may impact the Country's economy.

As it is a sensible choice, setting W has to be tuned.

In order to do so, in this work the following *eight* textual windows are considered:

- $B_{0,0}$, $B_{0,1}$, $B_{0,2}$, $B_{0,3}$ which allow determining words polarities using the texts from one year and up to three years, including the contemporaneous one;
- $B_{1,1}$, $B_{1,2}$, $B_{1,3}$, $B_{1,4}$ which, instead, allow determining words polarities using the texts from one year and up to three years, excluding the contemporaneous one.

So, in total, under both configurations category, a maximum of four years is considered.

Limiting the choice only to the mentioned textual windows relies on the assumption that the effect of current Parliament discussion on the economy, although probably delayed, tends to fade over time. Hence what was discussed many years ago presumably has no effect today compared to more recent debate. For this reason, $W = 4$ (with or without the contemporaneous year) appears to be a reasonable choice for the textual window considering four years as a long enough period for an economic cycle duration.

To clarify the procedure, the following equations report the formulae for the $\boldsymbol{sgnW}_t^{(*)} = \boldsymbol{W}_t^{(*)}\boldsymbol{sgn\Delta y}$ products, where $* = \{B_{0,0}, B_{1,1}, B_{0,1}\}$, and $\boldsymbol{sgnW}_t^{(*)}$ is the resulting "*polarised textual window*". Analogous matrices and products correspond to the other five windows configurations.

Under the $B_{0,0}$ configuration the *textual window* is given by:

$$\boldsymbol{W}_t^{(B_{0,0})} = \begin{bmatrix} \widetilde{f}_{t,1} & \ldots & \widetilde{f}_{t,V} \end{bmatrix} \qquad t = 1\ldots,M \tag{3.13}$$

while the corresponding *polarised textual window* is given by:

$$\begin{aligned} \boldsymbol{sgnW}_t^{(B_{0,0})} &= \boldsymbol{W}_t^{(B_{0,0})} \boldsymbol{sgn\Delta y} \\ &= \begin{bmatrix} sgnW_{t,1} & \dots & sgnW_{t,V} \end{bmatrix} \end{aligned} \tag{3.14}$$

where

$$sgnW_{t,v} = \widetilde{f}_{t,v} \times sgn\Delta y_t \qquad t = 1, \dots, M \qquad v = 1, \dots, V$$

Instead, under the $B_{1,1}$ configuration, the *textual window* is given by:

$$\boldsymbol{W}_t^{(B_{1,1})} = \begin{bmatrix} \widetilde{f}_{t,1} & \dots & \widetilde{f}_{t,V} \end{bmatrix} \qquad t = 1 \dots, M-1 \tag{3.15}$$

while the corresponding *polarised textual window* is given by:

$$\begin{aligned} \boldsymbol{sgnW}_t^{(B_{1,1})} &= \boldsymbol{W}_t^{(B_{1,1})} \boldsymbol{sgn\Delta y} \\ &= \begin{bmatrix} sgnW_{t,1} & \dots & sgnW_{t,V} \end{bmatrix} \end{aligned} \tag{3.16}$$

where

$$sgnW_{t,v} = \widetilde{f}_{t,v} \times sgn\Delta y_t \qquad t = 1, \dots, M-1 \qquad v = 1, \dots, V$$

Finally, under the $B_{0,1}$ configuration the *textual window* is given by:

$$\boldsymbol{W}_t^{(B_{0,1})} = \begin{bmatrix} \widetilde{f}_{t-1,1} & \dots & \widetilde{f}_{t-1,V} \\ \widetilde{f}_{t,1} & \dots & \widetilde{f}_{t,V} \end{bmatrix} \qquad t = 2, \dots, M \tag{3.17}$$

while the corresponding *polarised textual window* is given by:

$$\begin{aligned} \boldsymbol{sgnW}_t^{(B_{0,1})} &= \boldsymbol{W}_t^{(B_{0,1})} \boldsymbol{sgn\Delta y} \\ &= \begin{bmatrix} sgnW_{t-1,1} & \dots & sgnW_{t-1,V} \\ sgnW_{t,1} & \dots & sgnW_{t,V} \end{bmatrix} \end{aligned} \tag{3.18}$$

where

$$sgnW_{t,v} = \widetilde{f}_{t,v} \times sgn\Delta y_t \qquad t = 2, \dots, M \qquad v = 1, \dots, V$$

For the sake of clarity, Figure 3.1 shows a timeline representation of the textual windows $B_{0,0}$, $B_{1,1}$ and $B_{0,1}$. A similar representation is possible also for the other five textual windows configurations.

Example Tables 3.8 to 3.10 better clarify the procedure.

FIGURE 3.1: Textual windows timeline representation

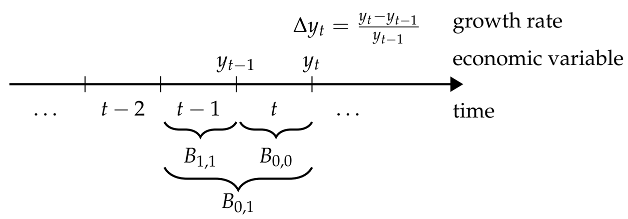


TABLE 3.8: $B_{0,0}$ Polarised Textual Windows Example

		v_1	v_2	v_3	v_4	v_5
w_1	t_1	0	0	0	-0.6	0
w_2	t_2	0	0	0	0	0
w_3	t_3	0	0.15	0	0	0.6
w_4	t_4	0	-0.15	0	0	0

TABLE 3.9: $B_{1,1}$ Polarised Textual Windows Example

		v_1	v_2	v_3	v_4	v_5
w_1	t_1	0	0	0	-0.6	0
w_2	t_2	0	0	0	0	0
w_3	t_3	0	-0.15	0	0	-0.6

TABLE 3.10: $B_{0,1}$ Polarised Textual Windows Example

		v_1	v_2	v_3	v_4	v_5
w_2	t_1	0	0	0	-0.6	0
	t_2	0	0	0	0	0
w_3	t_2	0	0	0	0	0
	t_3	0	0.15	0	0	0.6
w_4	t_3	0	-0.15	0	0	-0.6
	t_4	0	-0.15	0	0	0

2. The unique polarity for each word v is obtained by summing the frequencies across the rows of the polarised textual windows matrices $\boldsymbol{sgnW}_t$ as in Equation 3.19:

$$\underset{V\times 1}{\boldsymbol{p}^{(y)}} = \begin{bmatrix} p_1^{(y)} \\ \vdots \\ p_V^{(y)} \end{bmatrix} \qquad p_v^{(y)} = \sum_t sgnW_{t,v} \qquad v = 1, \dots, V \tag{3.19}$$

Table 3.11 reports the results for the given examples under the different textual windows $B_{0,0}$, $B_{1,1}$, $B_{0,1}$. Notably, v_2 receives polarity equal to zero under the

$B_{0,0}$ configuration while its polarity is negative under the other configurations. Also, v_5 receives a polarity equal to 0.6, -0.6 or 0 under the $B_{0,0}$, $B_{1,1}$ and $B_{0,1}$ configurations, respectively.

TABLE 3.11: $B_{0,0}$, $B_{1,1}$ and $B_{0,1}$ Unique Word Polarity Example

	v_1	v_2	v_3	v_4	v_5
$B_{0,0}$	0	0	0	-0,6	0,6
$B_{1,1}$	0	-0,15	0	-0,6	-0,6
$B_{0,1}$	0	-0,15	0	-0,6	0

3. The obtained polarities, which are continuous in their original range due to the word frequencies and weights, are then rescaled in the continuous interval $[-1, 1]$ through to the min-max normalisation:

$$\underset{V\times 1}{\widetilde{\boldsymbol{p}}^{(y)}} = \begin{bmatrix} \widetilde{p}_1^{(y)} \\ \vdots \\ \widetilde{p}_V^{(y)} \end{bmatrix} \qquad \widetilde{p}_v^{(y)} = 2\frac{p_v^{(y)} - \min \boldsymbol{p}^{(y)}}{\max \boldsymbol{p}^{(y)} - \min \boldsymbol{p}^{(y)}} - 1 \tag{3.20}$$

$$v = 1, \dots, V$$

Following the example so far, the normalised polarities under the $B_{0,0}$, $B_{1,1}$ and $B_{0,1}$ textual windows configurations are reported in Table 3.12.

TABLE 3.12: $B_{0,0}$, $B_{1,1}$ and $B_{0,1}$ Normalised Word Polarity Example

	v_1	v_2	v_3	v_4	v_5
$B_{0,0}$	0	0	0	-1	1
$B_{1,1}$	1	-0.5	1	-1	-1
$B_{0,1}$	1	-0.5	1	-1	1

3.4.4 Polarity indices time-series

Total Textual Political Polarity Index

After determining each word polarity on the first $M = 43$ *training sample* years, time-series of global polarity indices are constructed on the remaining 30 *out-of-sample* years $t = M + 1, \dots, T$.

Under this framework there might be some words that do not receive polarity at all - i.e. they receive polarity equal to zero - only because they were not used by orators (or did not exist at all) in training sample years. In the analysis at hand, such words amount to around 1,700 over a total of about 71,000 words - i.e. less than 2.5% of the entire corpus vocabulary. For this reason, the issue is disregarded as its impact on final results is assumed to be negligible.

The steps for constructing the polarity indices time-series are detailed here:

1. A matrix of normalised polarities (Equation 3.21) with the same dimensions of the TF-IDF weighted frequency matrix in Equation 3.4 is built. Each row is equal to the vector of normalised polarities in Equation 3.20.

$$\underset{(T-M)\times V}{\widetilde{\boldsymbol{P}}^{(y)}} = \begin{bmatrix} \widetilde{\boldsymbol{p}}^{(y)^\top} \\ \vdots \\ \widetilde{\boldsymbol{p}}^{(y)^\top} \end{bmatrix} \tag{3.21}$$

2. The Hadamard product between the matrices in Equations 3.4 and 3.21 is performed to get a new matrix whose elements are the "*polarised*" weighted frequencies - i.e. frequencies reflecting the sign and intensity of each word polarity as previously calculated. Such (test sample) polarised matrix is expressed in Equation 3.22.

$$\begin{aligned} \underset{(T-M)\times V}{\boldsymbol{G}^{(y)}} &= \underset{(T-M)\times V}{\widetilde{\boldsymbol{F}}} \circ \underset{(T-M)\times V}{\widetilde{\boldsymbol{P}}^{(y)}} \\ &= \begin{bmatrix} g^{(y)}_{M+1,1} & \cdots & g^{(y)}_{M+1,V} \\ \vdots & g^{(y)}_{t,v} & \vdots \\ g^{(y)}_{T,1} & \cdots & g^{(y)}_{T,V} \end{bmatrix} \end{aligned} \tag{3.22}$$

where

$$g^{(y)}_{t,v} = \widetilde{f}_{t,v} \times \widetilde{p}^{(y)}_v \qquad t = M+1,\dots,T \qquad v = 1,\dots,V$$

3. Time-series of positive, negative and total polarity are obtained by making the row sums of the elements in matrix $\boldsymbol{G}$.

   Their expressions are given in Equations 3.23, 3.24 and 3.25, respectively:

$$\underset{(T-M)\times 1}{\boldsymbol{pos}^{(y)}} = \begin{bmatrix} pos_{(M+1)} \\ \vdots \\ pos_T \end{bmatrix} \qquad pos^{(y)}_t = \sum_{v=1}^{V} g^{(y)}_{t,v} \mid g^{(y)}_{t,v} > 0 \tag{3.23}$$

$$\underset{(T-M)\times 1}{\boldsymbol{neg}^{(y)}} = \begin{bmatrix} neg_{(M+1)} \\ \vdots \\ neg_T \end{bmatrix} \qquad neg^{(y)}_t = \sum_{v=1}^{V} g^{(y)}_{t,v} \mid g^{(y)}_{t,v} < 0 \tag{3.24}$$

$$\underset{(T-M)\times 1}{\boldsymbol{tot}^{(y)}} = \begin{bmatrix} tot_{(M+1)} \\ \vdots \\ tot_T \end{bmatrix} \qquad tot^{(y)}_t = \sum_{v=1}^{V} |g^{(y)}_{t,v}| \tag{3.25}$$

where $t = M+1,\dots,T$.

4. Finally, the "*Total Textual Political Polarity Index*" (*TPPI-T*) at each time is calculated as in Equation 3.27:

$$\underset{(T-M)\times 1}{\boldsymbol{TPPI\text{-}T}^{(y)}} = \begin{bmatrix} TPPI\text{-}T^{(y)}_{(M+1)} \\ \vdots \\ TPPI\text{-}T^{(y)}_{T} \end{bmatrix} \tag{3.26}$$

where

$$TPPI\text{-}T^{(y)}_t = \frac{pos^{(y)}_t + neg^{(y)}_t}{tot^{(y)}_t} \qquad t = M+1, \ldots, T \tag{3.27}$$

The index in Equation 3.27 is referred to as "*Total*" as it is constructed on the entire set of texts without conditioning on the political group each orator belongs to at the time they pronounce her speech.

Group Specific Textual Political Polarity Indices

Following the same procedure detailed in Section 3.4.4, it is possible to construct analogous polarity indices by conditioning on the orators' political group belonging: *Government*, *Majority*, *Opposition* [15].

Therefore, also the following *four* "*Group Specific Textual Political Polarity Indices*", *TPPI-GS*, are built - one for each political group:

$$TPPI\text{-}GS^{(y)}_{i,t} \qquad i = \{Gov, Maj, Opp\} \qquad t = \{M+1, \ldots, T\} \tag{3.28}$$

3.4.5 Polarity divergence indices

Similarly to any oral speech, textual data convey information about the *tone* or *sentiment* of the discussion. This is the feature which the *TPPI-T* and the *TPPI-GS* described so far aim at modelling - i.e. what in this work has been called the *polarity* of the political debate during Assembly sessions over time.

Nonetheless, especially in the political debate contest, another relevant feature to analyse is the *degree of agreement or disagreement* between one political group and the others.e

Therefore this work also proposes text-based indices that measure such degree of agreement or disagreement between political groups. Specifically, such new indices are derived directly from the *TPPI-GS* and are referred to as "*Group Polarity Divergence*" indices, *TPPI-D*. Their main innovation is measuring the degree of agreement between political groups based on the different tones they show during Assembly debates.

[15] To obtain these new indices, texts are aggregated based on group belonging and the frequency matrices are built accordingly - i.e. a matrix for each group.

The derivation of these *TPPI-D* from the *TPPI-GS* requires the computation of the *squared differences* [16] between each of the four *TPPI-GS* at each time. The resulting time-series are given by the 3 possible combinations of the 3 political groups previously considered: *Government-Majority*, *Government-Opposition*, *Majority-Opposition*.

In formulae, let $\boldsymbol{TPPI\text{-}GS_i}$ be the $(T-M) \times 1$ vector containing the values of the i-th Group Specific TPPI with $i = \{Gov, Maj, Opp\}$, the (ij-th) *Group Divergence Polarity Index* vector, $\boldsymbol{TPPI\text{-}D_{ij}}$, is defined as follows:

$$\underset{(T-M)\times 1}{\boldsymbol{TPPI\text{-}D_{ij}}^{(y)}} = \begin{bmatrix} TPPI\text{-}D^{(y)}_{ij,(M+1)} \\ \vdots \\ TPPI\text{-}D^{(y)}_{ij,T} \end{bmatrix} \tag{3.29}$$

where

$$TPPI\text{-}D^{(y)}_{ij,t} = (TPPI\text{-}GS^{(y)}_{i,t} - TPPI\text{-}GS^{(y)}_{j,t})^2 \qquad i \neq j \qquad t = M+1, \ldots, T$$

In addition, in order to obtain a unique polarity divergence index, the *Average Group Divergence Index*, $TPPI\text{-}D_{Avg}$, is constructed by averaging the values of the 3 polarity divergence series at each time.

In formulae, the $TPPI\text{-}D_{Avg}$ is given by:

$$\underset{(T-M)\times 1}{\boldsymbol{TPPI\text{-}D_{Avg}}^{(y)}} = \begin{bmatrix} TPPI\text{-}D^{(y)}_{Avg,(M+1)} \\ \vdots \\ TPPI\text{-}D^{(y)}_{Avg,T} \end{bmatrix} \tag{3.30}$$

where

$$\begin{aligned} TPPI\text{-}D^{(y)}_{Avg,t} &= \frac{1}{C} \sum_{ij} TPPI\text{-}D^{(y)}_{ij,t} \\ C &= \textit{number of possible couples of i and j} \\ i,j &= \{Gov, Maj, Opp\} \qquad i \neq j \\ t &= M+1, \ldots, T \end{aligned}$$

In conclusion, a total of 4 "*Group Polarity Divergence*" indices, *TPPI-D*, are obtained [17]:

A synoptic scheme of the 8 *TPPI* is reported in Table 3.13.

[16] Actually any other suitable *loss function* may be used.

[17] To keep the notation simple, when possible and not misleading, divergence indices are abbreviated into the two groups names only - e.g. *Maj-Opp* instead of $TPPI\text{-}D_{MajOpp}$.

TABLE 3.13: Textual Political Polarity Indices Overview

Textual Political Polarity Indices	
Total	*Group-Specific*
TPPI-T	$TPPI\text{-}GS_{Gov}$ $TPPI\text{-}GS_{Maj}$ $TPPI\text{-}GS_{Opp}$
Divergence Textual Political Polarity Indices	
Total	*Group-Specific*
TPPI-D-Avg	$TPPI\text{-}D_{GovMaj}$ $TPPI\text{-}D_{GovOpp}$ $TPPI\text{-}D_{MajOpp}$

3.4.6 Selecting the textual window

As stated in Section 3.4.3, 8 textual windows configurations (from a $B_{0,0}$ to $B_{1,4}$) are used to construct the polarity indices. Therefore, each index - either "*Total*" or "*Group-Specific*" - is available in 8 different versions depending on the window employed for texts aggregation. In total there are 32 different indices: 8 *TPPI-T* and 24 *TPPI-GS* [18].

To tune the textual window length, W, and therefore choose the best configurations, the Model Confidence Set, MCS, approach of Hansen, Lunde, and Nason (2011) is used.

The technique is based on a sequence of tests "*which selects the best performing specifications over the considered time span*" (Amendola et al., 2020) with a given confidence level. "*The MCS procedure does not assume that a particular model is the true model; in fact, the MCS procedure can be used to compare more general objects, beyond the comparison of models*"(Hansen, Lunde, and Nason, 2011).

In this work the MCS is used to compare the *out-of-sample* values of the polarity indices time-series - both, *Total* and *Group Specific* - against the *out-of-sample* values of the chosen macroeconomic time-series.

The MCS procedure requires a "*a properly chosen evaluation criterion*" (Amendola and Storti, 2015) in order to compare different objects - i.e. a *loss function*. Here, the *Squared Error* function, *SE*, is adopted.

In formulae:

$$SE_t^{(w)} = (y_t - TPPI_t^{(w)})^2 \tag{3.31}$$

[18] For notation clarity, when necessary, indices built under different texts aggregation configurations are referred to indicating the textual window in the apex - e.g. $TPPI\text{-}T^{(B_{0,0})}$, $TPPI\text{-}GS_{Gov}^{(B_{0,1})}$, $TPPI\text{-}D_{GovMaj}^{(B_{0,0})}$.

where

$$t = M + 1, \dots, T$$
$$w = \{B_{0,0}, \dots, B_{0,3}, B_{1,1}, \dots, B_{1,4}\}$$
$$TPPI^{(w)} = \textit{the TPPI-T or TPPI-GS under textual window } w$$

After constructing the loss functions, 4 MCS are performed: one for each of the indices *TPPI-T*, *TPPI-GS*$_{Gov}$, *TPPI-GS*$_{Maj}$, *TPPI-GS*$_{Opp}$. These are "*primary*" indices as they are not derived from other indices as in the case of the *TPPI-D*.

The results of the MCS - all with a 0.99 confidence level - are reported in Tables 3.14. Out of $N = 32$ indices versions tested, $N^* = 9$ are selected as the best - i.e. entering the MCS:

- for *TPPI-T*: $B_{0,0}$, $B_{0,2}$;
- for *TPPI-GS*$_{Gov}$: $B_{0,1}$, $B_{0,0}$, $B_{1,1}$;
- for *TPPI-GS*$_{Maj}$: $B_{0,0}$, $B_{0,2}$;
- for *TPPI-GS*$_{Opp}$: $B_{0,0}$, $B_{1,1}$.

In light of these results, as far it concerns the *Total* and *Group Specific TPPI*, only these N^* indices selected via the MCS are considered in the following analysis.

Things are slightly different for the *TPPI-D*. In fact these *seven* indices are "*secondary*" indices in the sense they are derived from the *TPPI-GS*. In total there are 32 *TPPI-D* as each index comes in 8 different textual windows configurations. For such indices, the choice for the appropriate textual windows relies on the selection of the *TPPI-GS* they derive from.

Therefore, for each selected *TPPI-GS* version the corresponding *TPPI-D* with the same textual window are selected, whether or not the textual window for that *TPPI-GS* is the same as the textual window of the second *TPPI-GS* involved in the *TPPI-D* construction.

For instance, the *TPPI-D*$_{GovMaj}$ is considered not only in the configuration $B_{0,0}$ – which is the textual window common to both the *TPPI-GS*$_{Gov}$ and *TPPI-GS*$_{Maj}$ – but also in the configurations $B_{0,1}$, $B_{1,1}$ – which are the other best textual windows for the *TPPI-GS*$_{Gov}$ – and $B_{0,1}$ – which is the other best choice for the *TPPI-GS*$_{Maj}$.

Analogously, also the other *TPPI-D* are selected.

Therefore, out of a total of 32 versions, the following 12 are selected:

- for *TPPI-D*$_{GovMaj}$: $B_{0,0}$, $B_{1,1}$, $B_{0,1}$, $B_{0,2}$;
- for *TPPI-D*$_{GovOpp}$: $B_{0,0}$, $B_{1,1}$, $B_{0,1}$;
- for *TPPI-D*$_{MajOpp}$: $B_{0,0}$, $B_{1,1}$, $B_{0,2}$;
- for *TPPI-D*$_{Avg}$: $B_{0,0}$, $B_{1,1}$ (i.e. the only textual windows common to the *TPPI-D* selected in previous points).

TABLE 3.14: Test-sample TPPI-T and TPPI-GS: Textual Windows MCS and Mean Squared Errors

Indices							
TPPI-T		*TPPI-GS$_{Gov}$*		*TPPI-GS$_{Maj}$*		*TPPI-GS$_{Opp}$*	
Textual Window	*MSE*	*Textual Window*	*MSE*	*Textual Window*	*MSE*	*Textual Window*	*MSE*
$B_{0,0}$ [a]	2.128	$B_{0,1}$ [a]	2.137	$B_{0,0}$ [a]	2.209	$B_{0,0}$ [a]	2.299
$B_{0,2}$ [a]	2.342	$B_{0,0}$ [a]	2.138	$B_{0,2}$ [a]	2.373	$B_{1,1}$ [a]	2.377
$B_{1,1}$	2.376	$B_{1,1}$ [a]	2.139	$B_{0,1}$	2.406	$B_{0,2}$	2.421
$B_{1,2}$	2.382	$B_{0,2}$	2.237	$B_{1,1}$	2.408	$B_{1,2}$	2.458
$B_{0,1}$	2.389	$B_{1,2}$	2.286	$B_{1,2}$	2.426	$B_{0,1}$	2.459
$B_{0,3}$	2.520	$B_{0,3}$	2.297	$B_{0,3}$	2.542	$B_{0,3}$	2.536
$B_{1,3}$	2.571	$B_{1,4}$	2.307	$B_{1,3}$	2.569	$B_{1,3}$	2.576
$B_{1,4}$	2.634	$B_{1,3}$	2.346	$B_{1,4}$	2.677	$B_{1,4}$	2.650

[a] *Entering the superior set of models with 0.99 confidence level*

3.4.7 Testing causal effects

In this work the main goal of constructing textual indices is to quantify the effect of parliamentary discussion on Country economy. To this purpose the underlying necessary assumption is that such effect do exists. Moreover, as usual in the economic field, there might exists even a reverse causality. In other words, in the contest at hand, it way well happen that are economic dynamics determining the parliamentary discussion agenda and not the opposite. For this reasons Granger causality tests (Granger, 1969) are performed. As recalled by Scaramozzino, Cerchiello, and Aste, 2021, this type of test looks "*at causality in terms of the amount of extra information that the observation of a variable provides about another variable*" which "*in its original formulation, this corresponds to an additional term in a linear regression*".

Table 3.15 reports the results of Granger tests showing significant values at common significance levels for both hypothesis: causal effects of parliamentary discussion on the Italian GDP growth rates and the reverse relation.

What emerges is that, at least at lag 1, there is essentially no evidence in favour of the reverse hypothesis – i.e. GDP causing political debate. Interestingly, all indices but one showing significant causal effects on the GDP are of type TPPI-D, which suggests that the difference in tones between political groups plays a major role in predicting economic growth dynamics compared to the tone itself.

TABLE 3.15: Granger causality test between the textual indices and the Italian GDP growth rates

Tested Hypothesis	$GDP \leftarrow TPPI$			$TPPI \leftarrow GDP$		
Unrestricted Model	$y_t = \varphi y_{t-1} + \beta x_{t-1} + \varepsilon_t$			$x_t = \beta x_{t-1} + \varphi y_{t-1} + \varepsilon_t$		
Restricted Model	$y_t = \varphi y_{t-1} + \varepsilon_t$			$x_t = x_{t-1} + \varepsilon_t$		
			TPPI-GS			
$x_t; x_{t-1}$	*F Statistic*	*p-value*		*F Statistic*	*p-value*	
$B_{0,0}Gov$	4.261	0.049	*	0.792	0.382	
			TPPI-D			
$x_t; x_{t-1}$	*F Statistic*	*p-value*		*F Statistic*	*p-value*	
$B_{0,0}MajOpp$	6.503	0.017	*	0.015	0.905	
$B_{0,1}MajOpp$	4.568	0.042	*	0.142	0.710	
$B_{0,2}GovMaj$	6.686	0.016	*	0.648	0.428	
$B_{0,2}MajOpp$	4.407	0.046	*	1.870	0.183	
$B_{0,3}MajOpp$	3.235	0.084	.	1.730	0.200	
$B_{1,1}GovMaj$	8.298	0.008	**	0.012	0.915	
$B_{1,1}MajOpp$	5.169	0.031	*	1.129	0.298	
$B_{1,2}GovMaj$	4.527	0.043	*	0.260	0.614	
$B_{1,2}MajOpp$	4.426	0.045	*	3.114	0.089	.
$B_{1,3}MajOpp$	3.746	0.0639	.	1.368	0.253	
	*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$					

3.5 Evaluating annual TPPI predictive ability

3.5.1 Model estimation

After constructing the polarity indices, their effectiveness in improving the predictability of the economic variable they are constructed for is tested via time-series regressions.

Specifically, basing on Larsen and Thorsrud, 2019, *Auto-regressive with Exogenous Variable* (*AR-X*) models (Equation 3.32) are compared against the simpler *Auto-regressive* (*AR*) model of order one (Equation 3.33).

$$y_t = \beta x_{t-1} + \varphi y_{t-1} + \varepsilon_t \tag{3.32}$$

$$y_t = \varphi y_{t-1} + \varepsilon_t \tag{3.33}$$

In Equations 3.32 and 3.33, y is the Italian GDP annual growth rate described in Section 3.4.2, while the exogenous term x is the polarity index whose predictive ability is to evaluate.

As they are on different scales, x and y are standardised.

Following Larsen and Thorsrud, 2019, the predictive power of the polarity indices is tested at temporal lag 1 (x_{t-1}), as shown in Equation 3.32.

Predictions are performed on the *out-of-sample* years $t = M+1, \dots, T$ (i.e. from 1991 to 2020). This strategy ensures that no economic or textual information already exploited for the polarity indices construction may be also used for predictions and indices evaluation.

In detail, benchmark *AR* model aside, 21 regressions are estimated:

- 2 *AR-X* regressions for the *TPPI-T*. The indices evaluated in such regressions are the following:
  - $TPPI\text{-}T^{(B_{0,0})}$;
  - $TPPI\text{-}T^{(B_{0,2})}$;
- 7 *AR-X* regressions for the *TPPI-GS*. The indices evaluated in such regressions are the following:
  - $TPPI\text{-}GS_{Gov}^{(B_{0,0})}$, $TPPI\text{-}GS_{Maj}^{(B_{0,0})}$, $TPPI\text{-}GS_{Opp}^{(B_{0,0})}$;
  - $TPPI\text{-}GS_{Gov}^{(B_{0,1})}$;
  - $TPPI\text{-}GS_{Maj}^{(B_{0,2})}$;
  - $TPPI\text{-}GS_{Gov}^{(B_{1,1})}$, $TPPI\text{-}GS_{Opp}^{(B_{1,1})}$;
- 12 *AR-X* regressions for the *TPPI-D*. The indices evaluated in such regressions are the following:

- $TPPI\text{-}D^{(B_{0,0})}_{GovMaj}$, $TPPI\text{-}D^{(B_{0,0})}_{GovOpp}$, $TPPI\text{-}D^{(B_{0,0})}_{MajOpp}$, $TPPI\text{-}D^{(B_{0,0})}_{Avg}$;
- $TPPI\text{-}D^{(B_{0,1})}_{GovMaj}$, $TPPI\text{-}D^{(B_{0,1})}_{GovOpp}$;
- $TPPI\text{-}D^{(B_{0,2})}_{GovMaj}$, $TPPI\text{-}D^{(B_{0,2})}_{MajOpp}$;
- $TPPI\text{-}D^{(B_{1,1})}_{GovMaj}$, $TPPI\text{-}D^{(B_{1,1})}_{GovOpp}$, $TPPI\text{-}D^{(B_{1,1})}_{MajOpp}$, $TPPI\text{-}D^{(B_{1,1})}_{Avg}$.

The results of these *AR-X* regressions, together with the benchmark *AR* regression, are reported in Table 3.16. Specifically, only the regressions with estimated parameters whose significance level is below 10% are reported.

Results point out that in 9 regressions out of 21 the inclusion of a *TPPI* as an exogenous variable does improve the annual GDP growth rate predictions.

More in detail, regressions results show that the polarity divergence between political groups is the most informative feature. The substantial increment in the R-squared for the $TPPI\text{-}D^{(B_{0,0})}_{GovMaj}$ is coherent with the existence of a "*relationship of confidence*" between the Government and its Majority. The negative coefficient for this index confirms that as tone divergence between these two groups increases, the GDP growth rate decreases.

3.5.2 Model evaluation

After estimation, models performances are compared through an MCS. To this aim, at each time $t = M + 1, \dots, T$, the squared difference (or squared error) between observed *out-of-sample* macroeconomic variable and fitted values is calculated for each one of the seven *AR-X* with significant exogenous term as well as for the reference *AR* model.

The results of a MCS with a 0.99 confidence level are reported in Table 3.17 and point that two models outperform the others: the *AR-X* with $TPPI\text{-}D^{(B_{0,2})}_{GovMaj}$ series and the *AR-X* with $TPPI\text{-}GS^{(B_{0,0})}_{Maj}$, both at lag one. These two models also show the lowest *Mean Squared Error*, MSE, between observed and predicted values, while the benchmark *AR* model is the one showing the worst performance overall.

The temporal dynamic of observed and predicted values (only for the *AR* and the two best *AR-X* models) is presented in Figure 3.2 which confirms the superior performance of the regressive models including the textual indices. Notably, as shown in the same graph, the use of the proposed textual indices improves model performances also in peculiar years as the 2008 and 2020 - i.e. during the financial crisis and Covid-19 pandemic, respectively.

3.6 Rolling window polarity indices

The indices in Section 3.4 are constructed by splitting the original sample years $t = 1, \dots, T = 73$ (1948–2020) in two parts: a training sample (S_1) of 43 years – i.e. $t =$

TABLE 3.16: Italian yearly GDP, 2000 constant prices, annual growth rates. *AR* and *AR-X* models, 30 years estimation period: 1991–2020. Fixed window scheme textual indices.

AR Model $y_t = \varphi y_{t-1} + \varepsilon_t$						
x_{t-1}	φ		β		*Regr. SE*	R^2
	0.915 (0.125)	***			0.667	-0.013
AR-X Models $y_t = \beta x_{t-1} + \varphi y_{t-1} + \varepsilon_t$						
TPPI-GS						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{0,0}Gov_{t-1}$	0.924 (0.13)	***	0.156 (0.069)	*	0.66	0.044
TPPI-D						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{0,0}GovMaj_{t-1}$	0.948 (0.114)	***	-0.324 (0.156)	*	0.591	0.233
$B_{0,2}GovMaj_{t-1}$	0.887 (0.104)	***	-0.334 (0.146)	*	0.585	0.248
$B_{1,1}GovMaj_{t-1}$	0.904 (0.126)	***	-0.182 (0.081)	*	0.652	0.065

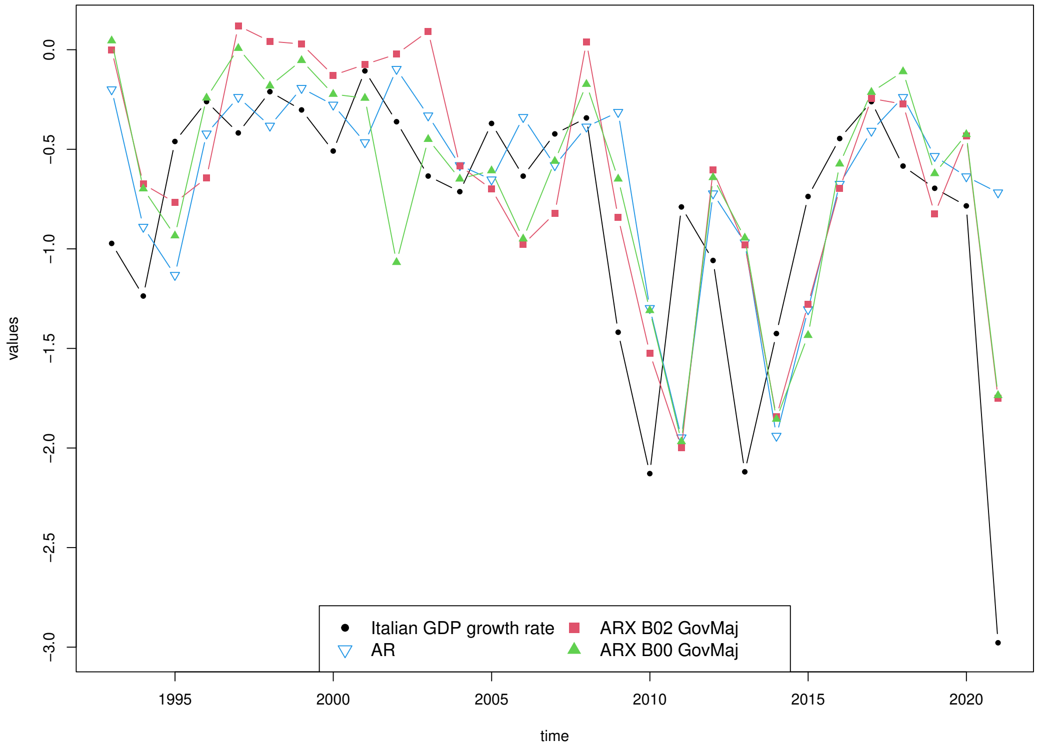
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$

TABLE 3.17: Annual growth rate of the Italian yearly GDP at 2000 constant prices: best models via Model Confidence Set.

Model	*Exogenous Variable*	*MSE*
AR-X	$TPPI\text{-}D^{(B_{0,2})}_{GovMaj}$	0.319[a]
	$TPPI\text{-}D^{(B_{0,0})}_{GovMaj}$	0.325[a]
	$TPPI\text{-}D^{(B_{1,1})}_{GovMaj}$	0.396
	$TPPI\text{-}GS^{(B_{0,0})}_{Gov}$	0.405
AR	-	0.429

[a] *Entering the MCS at 0.99 confidence level*

FIGURE 3.2: Annual growth rate of the Italian yearly GDP at 2000 constant prices: best *AR-X* models via MCS vs *AR* model



$1, \ldots, M = 43$ (1948–1990) – to compute words polarities, and a test sample (S_2) of 30 years – i.e. $t = M + 1, \ldots, T = 73$ (1991–2020) – to build and evaluate the (out-of-sample) time-series of polarity indices. In time-series literature this is called a "*fixed scheme*" approach. As noticed in Section 3.4.4, the vocabulary is fixed and composed of all words that appeared at least once in the training sample years. Therefore, as a drawback, about 1,700 words appearing in the test sample are excluded from the analysis as they receive no polarity during training, either because not used by orators or not existing yet in the first M years. As explained before, the impact of such an issue on final results is assumed to be negligible.

However, even if this is the case, the fixed approach poses a more delicate question: how to consider in the analysis the circumstance that, although slightly and smoothly, languages too do change over time. Such an issue is particularly relevant in the present analysis which spans 73 years, a very long time.

Therefore, to consider temporal dynamics in the construction of the proposed TPPI, following similar literature in sentiment analysis (Ahelegbey, Cerchiello, and Scaramozzino, 2022; Aprigliano et al., 2022; Caporin and Poli, 2017; Li et al., 2014; Shapiro, Sudhof, and Wilson, 2020), a "*rolling window*" approach is adopted.

Specifically, as the choice of different window sizes may affect the estimates, the robustness of results is evaluated under the 5 rolling window schemes in Table 3.18.

TABLE 3.18: Rolling window schemes details

$RW_{m,r}$	$T-m$	z	$S_{1,z}^{(m,r)}$	$S_{2,z}^{(m,r)}$
		1	1948–1970	1971–1975
$RW_{23,5}$	50	2	1953–1975	1976–1980
		...	...	...
		10	1993–2015	2016–2020
		1	1948–1975	1976–1980
$RW_{28,5}$	45	2	1953–1980	1981–1985
		...	...	...
		9	1988–2015	2016–2020
		1	1948–1980	1981–1985
$RW_{33,5}$	40	2	1953–1985	1986–1990
		...	...	...
		8	1983–2015	2016–2020
		1	1948–1985	1986–1990
$RW_{38,5}$	35	2	1953–1990	1991–1995
		...	...	...
		7	1978–2015	2016–2020
		1	1948–1990	1991–1995
$RW_{43,5}$	30	2	1953–1995	1996–2000
		...	...	...
		6	1973–2015	2016–2020

Let T be the size of the total sample where a number of Z rolling windows are constructed. Let m and r be respectively the size of the training, $S_{1,z}^{(m,r)}$, and test samples, $S_{2,z}^{(m,r)}$, within each rolling window $z = 1, \ldots, Z$. By construction, r is also the increment between successive rolling windows.

Within each rolling scheme, $RW_{m,r}$, the number of rolling windows is $Z = \frac{T-m}{r}$, while their total size is given by $s = m + r$.

Hence, the training and test samples for each z-th rolling window, under the rolling scheme $RW_{m,r}$, are given by:

$$S_{1,z}^{(m,r)} : \; t = 1 + (z-1)r, \ldots, M_z \tag{3.34}$$

$$S_{2,z}^{(m,r)} : \; t = M_z + 1, \ldots, T_z \tag{3.35}$$

where $M_z = m + (z-1)r$, $T_z = s + (z-1)r$, with $z = 1, \ldots, Z$. Under the generic rolling scheme $RW_{m,r}$, the length of the final out-of-sample TPPI time series is given by $T - m$. In fact, under the rolling scheme $RW_{m,r}$, the final out-of-sample series are obtained by chaining the Z (sub)series constructed on each $S_{2,z}^{(m,r)}$.

The decision to set the increments between consecutive rolling windows (r) equal to 5 years is motivated by two reasons. On one hand, the language during the first years after each training sample is expected not to change significantly in comparison. On the other hand, five years is also the duration of a Legislature prescribed by the Constitution. At the end of each Legislature a new Parliament must be elected, therefore a new Government formed. Hence, with a new Parliament (and Government), also the language during debates may be assumed to change.

After construction with the same procedure detailed in Sections 3.4.3 to 3.4.5, under each of the 5 rolling schemes, the obtained TPPI go through the same textual windows selection process in Section 3.4.6 by mean of Model Confidence Sets.

Finally, predictive ability is evaluated through *AR-X* models as in Section 3.5. Tables 3.19 to 3.23 report the results for the significant regressions. They confirm the major impact of TPPI-D compared to TPPI-GS (15 vs 8). As in the fixed scheme, also in the rolling schemes considered all the TPPI-GS are those of the Government. The fact that in all regressions such index receives a positive sign indicates the direct connection between the GDP growth rate and the tone expressed by Governments, which seems coherent with common sense and reality. The positive relation between GDP rates and divergence between Government-Majority, Government-Opposition and Majority-Opposition (3 cases in 5, 3 cases in 4 and 5 cases in 6, respectively) might be due to the fact that very long periods of low or no growth like the recent ones are probably associated with greater disagreement among political groups and uncertainty. As shown in Figure 3.3 rolling window indices seem to perform better than fixed scheme one in relatively less turbulent periods. For instance, in the graph the very low peak of 2020 due to the Covid-19 pandemic is better captured by the fixed scheme index. This might be due to the consideration that, although changing over time, languages evolve really slowly which translates in fixed schemes to retain longer memory of positive and negative events compensating the lack of relatively new information in updating the tone.

TABLE 3.19: Italian yearly GDP, 2000 constant prices, annual growth rates. AR and AR-X models, 50 years estimation period: 1971–2020. Rolling window scheme textual indices – $RW_{23,5}$

AR-X Models $y_t = \beta x_{t-1} + \varphi y_{t-1} + \varepsilon_t$						
TPPI-GS						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{1,1}Gov_{t-1}$	0.649 (0.184)	***	0.22 (0.099)	*	0.743	0.154
TPPI-D						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{0,1}MajOpp_{t-1}$	0.649 (0.181)	***	-0.194 (0.094)	*	0.748	0.143
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$						

TABLE 3.20: Italian yearly GDP, 2000 constant prices, annual growth rates. AR and AR-X models, 45 years estimation period: 1976–2020. Rolling window scheme textual indices – $RW_{28,5}$

AR-X Models $y_t = \beta x_{t-1} + \varphi y_{t-1} + \varepsilon_t$						
TPPI-D						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{0,1}GovMaj_{t-1}$	0.823 (0.11)	***	0.233 (0.084)	**	0.58	0.328
$B_{0,1}GovOpp_{t-1}$	0.839 (0.111)	***	0.219 (0.064)	**	0.586	0.315
$B_{0,2}GovOpp_{t-1}$	0.837 (0.113)	***	0.146 (0.066)	*	0.605	0.27
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$						

TABLE 3.21: Italian yearly GDP, 2000 constant prices, annual growth rates. AR and AR-X models, 40 years estimation period: 1981–2020. Rolling window scheme textual indices – $RW_{33,5}$

AR-X Models $y_t = \beta x_{t-1} + \varphi y_{t-1} + \varepsilon_t$						
TPPI-GS						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{1,2}Gov_{t-1}$	0.885 (0.118)	***	0.16 (0.076)	*	0.592	0.298
TPPI-D						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{1,1}GovMaj_{t-1}$	0.936 (0.104)	***	-0.221 (0.09)	*	0.575	0.337
$B_{1,1}GovOpp_{t-1}$	0.891 (0.111)	***	-0.201 (0.091)	*	0.583	0.32
$B_{0,0}MajOpp_{t-1}$	0.911 (0.114)	***	0.15 (0.078)	.	0.595	0.291
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$						

TABLE 3.22: Italian yearly GDP, 2000 constant prices, annual growth rates. AR and AR-X models, 35 years estimation period: 1986–2020. Rolling window scheme textual indices – $RW_{38,5}$

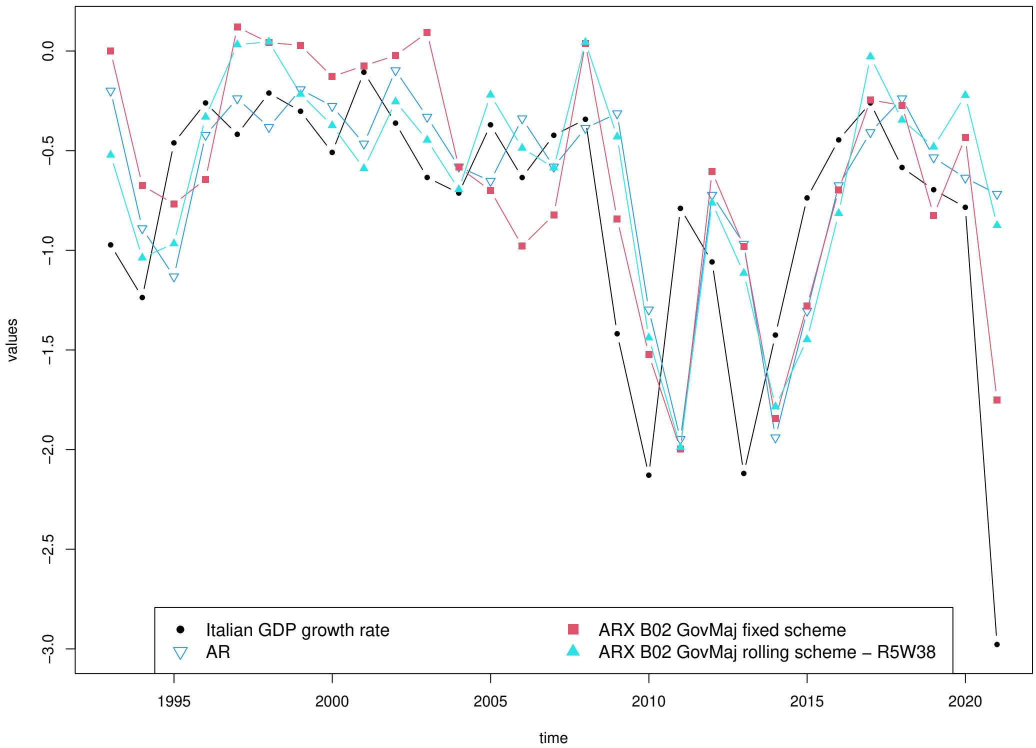
AR-X Models $y_t = \beta x_{t-1} + \varphi y_{t-1} + \varepsilon_t$						
TPPI-GS						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{0,2}Gov_{t-1}$	0.896 (0.125)	***	0.169 (0.093)	.	0.629	0.252
$B_{0,3}Gov_{t-1}$	0.893 (0.122)	***	0.201 (0.098)	*	0.621	0.271
$B_{1,2}Gov_{t-1}$	0.886 (0.124)	***	0.172 (0.087)	.	0.628	0.255
$B_{1,3}Gov_{t-1}$	0.89 (0.123)	***	0.192 (0.091)	*	0.623	0.266
TPPI-D						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{0,2}GovMaj_{t-1}$	0.897 (0.13)	***	0.24 (0.072)	**	0.608	0.303
$B_{0,3}GovMaj_{t-1}$	0.883 (0.126)	***	0.19 (0.074)	*	0.623	0.267
$B_{1,1}GovMaj_{t-1}$	0.951 (0.112)	***	-0.273 (0.099)	**	0.597	0.326
$B_{0,3}MajOpp_{t-1}$	0.914 (0.113)	***	0.238 (0.099)	*	0.613	0.291
$B_{1,3}MajOpp_{t-1}$	0.935 (0.122)	***	0.221 (0.1)	*	0.618	0.278
$B_{1,4}MajOpp_{t-1}$	0.934 (0.123)	***	0.236 (0.111)	*	0.613	0.29

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$

TABLE 3.23: Italian yearly GDP, 2000 constant prices, annual growth rates. AR and AR-X models, 30 years estimation period: 1991–2020. Rolling window scheme textual indices – $RW_{43,5}$

AR-X Models $y_t = \beta x_{t-1} + \varphi y_{t-1} + \varepsilon_t$						
TPPI-GS						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{0,1}Gov_{t-1}$	0.909 (0.135)	***	0.161 (0.075)	*	0.664	0.032
$B_{1,3}Gov_{t-1}$	0.907 (0.129)	***	0.135 (0.066)	.	0.668	0.019
TPPI-D						
x_{t-1}	φ		β		*Regr. SE*	R^2
$B_{1,4}GovOpp_{t-1}$	0.914 (0.121)	***	0.194 (0.092)	*	0.657	0.052
$B_{0,0}MajOpp_{t-1}$	0.972 (0.137)	***	0.199 (0.096)	*	0.655	0.058
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$						

FIGURE 3.3: Annual growth rate of the Italian yearly GDP at 2000 constant prices: best *AR*-*X* models via MCS vs *AR* model



Chapter 4

Conclusions

As stated in the Introduction, this thesis comprises two distinct papers sharing similar underlying research questions and data but applying different techniques for textual analysis within a statistical learning framework. The principal research purpose of measuring the relationship between political discourse and economic dynamics is achieved differently in the two previous chapters.

Specifically, Chapter 2 focuses on the relation between the content of the verbatim transcripts of the Italian Senate and eight quarterly national account statistics for the period 1996–2020. Transcripts content is analysed utilising a Correlated Topic Model (CTM) (Blei and Lafferty, 2007) which allows the estimation of the proportions of each discovered theme in the corpus. Such topics proportions are then combined to form *Textual Political Debate Indices* (TPDI) based on the evidence that their evolution in time is correlated (at least up to lag 2) with the growth rates time series of each economic variable considered. Hence, the main finding is that combining topic proportions makes it possible to obtain time series that can mimic the movements of a target economic indicator time series. This result is significant as it confirms that what elective representatives discuss in legislative assemblies does have a measurable impact on the economy. Therefore such qualitative information can improve the accuracy of traditional forecasting models. However, several aspects need further investigation.

First, remaining in the topic modelling area, CTM is not the only model in the literature. Hence, as a robustness check, several other models might be estimated. Such an option would need careful consideration as it may be computationally expensive. Actually, in the related literature, such comparisons are attempted when a new topic model is proposed rather than directly using such models in economics or political science applications. Nonetheless, it might be interesting to estimate topics through at least two other models in future research. Precisely, the first model might be the Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003), which is the first topic model proposed in the literature, hence the simplest to implement, and it may serve as a benchmark; The second model worth estimating is the Structural Topic Model (STM) (Roberts et al., 2014) whose more sophisticated architecture allows metadata in topic estimation. For example, among other features, such a model might reveal which theme is associated with a specific political party, providing greater detail on the parliamentary discussion.

Another delicate point in topic modelling is the dictionary size and composition. In the analysis conducted in Chapter 2, the dictionary is composed of only uni-grams (i.e. single words) to keep the vocabulary size relatively small (more than 50,000 uni-grams). However, such an approach has the limit of not considering the context around words. In future work, this issue might be managed by utilising bi-grams, tri-grams or n-grams – i.e. groups of two, three or *n* words (not necessarily successive words, which would be technically referred to as *skip-grams*). Even better than n-grams might be the recognition of the so called *collocations* or *phrasemes* – i.e. idiomatic phrases or multi-word expressions. On the other side, while expanding the vocabulary to allow for words context, also stop-word removal should be fine-tuned. In fact, in this work, not much care has been taken in considering context-specific stop-words. Recall that stop-words are words considered "useless" as they do not add meaningful information as being highly frequent words such as articles, conjunctions or pronouns. However, in addition to "usual stop-words" and depending on the specific domain and context, also common words may be considered as stop-words. For example, in the parliamentary context words such as *senator* or *president* or *government* are extremely frequent as they are used by orators when addressing each other during the debate. Removing stop-words and rare terms, is highly beneficial as it prevents model over-fitting and results in more interpretable topics. For handling the vocabulary, the use of deep learning methods such as word embedding might be beneficial as this would help reshape the word space in such a way as to preserve the semantic relations between terms. In fact, word embedding techniques assign each word a numeric vector in such a way that similar terms would be associated with similar vectors close to each other (e.g. *king* and *queen*). Deep learning models are attractive as they are available in pre-trained versions on enormous corpora. The only issue here would be the language and the context. Many available resources are built and tested on high-resource languages such as English. Translation, although possible, might affect results to some extent. Moreover, pre-trained models should be adapted to the specific domain and context as usually they are trained on "generalist" sources to capture the most common use of each word.

Moving on, concerning the indices construction and usage, a third delicate issue is the split of the corpus in training and test samples. Differently from Chapter 3 on sentiment analysis, in Chapter 2 the study is conducted on the entire data sample, motivating such choice with the need of having enough data for the topic model estimation. But, if they are going to be utilised for forecasting, indices should be constructed with the topic proportion of a test sample rather than those of a (unique) training sample. Hence, sample splitting is fundamental to checking indices' informative power and robustness. Such an issue may be overcome by extending the analysis to the entire corpus of data – i.e. back to 1948 – as in Chapter 3. However, it should be considered that the topics estimated on the training sample may not be relevant for use in the test sample as, meanwhile, the parliamentary discussion may have shifted to other more actual topics. What is reasonable to expect is that, on the test sample, very general and recurrent topics would be classified correctly, while highly specific, recent ones will not. Also the implementation of some rolling schemes may be challenging as, being generative probabilistic models, even the estimation of the same topic model on the same set of texts produces a slightly different output every time the model is run, let alone changing the training set composition. In this respect, further research is still necessary.

Finally, in Chapter 2, the indices are constructed by utilising linear models. This choice was motivated for consistency reasons with previous works in literature (for instance, Larsen and Thorsrud, 2019) and by the simple structure of the models adopted. Nonetheless, also non-linear relations could be handled and investigated. Moreover, after topic estimation, there might be some selection of the most relevant topics for a specific target variable. For this purpose, techniques such as LASSO or Ridge regression might be highly beneficial as they would help reduce the number of variables (topics).

Moving to Chapter 3, measuring the relation between parliamentary discussion and economy relies on a sentiment analysis technique and substantiates in the proposal of two categories of indices. The first category comprises the *Total Textual Political Polarity Indices* (TPPI-T) and *Group Specific Textual Political Polarity Indices* (TPPI-GS), which focus directly on the tone of the parliamentary discussion. The second category of indices, the *Divergence Textual Political Polarity Indices* (TPPI-D), investigates the tone divergence in the assumption that this feature can proxy the level of agreement or disagreement between three political groups – namely, Government, Majority and Opposition. The study analyses the entire corpus of verbatim reports of the Italian Senate from 1948 to 2020 and its relation to the annual GDP growth rate, and it presents at least two key contributions. First, it proposes a new method to estimate words sentiment or polarity. Actually, throughout the chapter, the term *polarity* is preferred over *sentiment* to remark the peculiarity that, differently from usual sentiment analysis techniques, the proposed approach aims to measure the positive or negative association between each word and the target economic variable, independently of the positive/negative sentiment or emotion commonly associated with each word. Instead, polarity computation based on the presence/absence of each word at each time, the frequency it has been used and the sign of the economic variable growth rate. The underlying idea is that a word has a positive (negative) polarity if, most of the times, it is associated with positive (negative) growth rates of the target economic variable. The motivation of such an approach is due to the assumption that parliamentary discussion and economic dynamic may be seen as a realisation of a unique probabilistic mechanism generating both simultaneously. Such a view, which is innovative and probably unusual, has at least a precedent in literature (Ke, Kelly, and Xiu, 2021). Nevertheless, the risk of such an approach lies in the fact that the information from growth rates, firstly, is used to assign word polarities which then are employed to construct indices which should explain the growth rates themselves. To prevent this circuit the original sample is split into a training and a test sample. In this way, word polarities are calculated on the training sample but indices construction and evaluation in terms of predictive power are performed on the test sample. This leads to the second major contribution of the study which is that word polarity computed in the proposed way does have an informative power. In other words, the circumstance that for many years a word has been associated with positive or negative dynamics of the target variable does help explain it as the same positive/negative words will tend to be associated with positive/negative growth rates of the target variable.

Another interesting finding is that, of the two categories, divergence indices seem to have more predictive power on the GDP dynamic. This might happen as a high disagreement between political groups is associated with negative GDP growth rates. Such assumption seems confirmed by a Granger causality test as described in Section 3.4.7. Moreover, the results highlight a significant negative linear relationship between the dynamic of the Italian GDP growth rate and the tone divergence in the discussion between Government and Majority members which, from a political perspective, seems coherent with the fundamental bond between a Government and its Majority.

Possible issues of applying the proposed procedure concern the sample splitting and the vocabulary used. Of course, the polarity can be estimated only for words present in the training sample while disregarding new words (or even words that already existed but never used during training sample years). In addition, also the way people speak does change over time. For this reason, the indices were built also under different rolling window schemes (see Section 3.6). However, the tested rolling window polarity indices do not seem to perform well as the fixed scheme ones. This can be due to two circumstances. First, languages tend to change very slowly (think that, depending on the language, native speakers can usually easily understand almost everything written in their language even some centuries ago), hence a rolling scheme may disregard some relevant past information in favour of some more updated but less relevant one. Employing an expanding window scheme may mitigate such an issue. Second, rolling window designs can determine a bias in the polarity computation when assigning the target variable growth rate signs to each word. Specifically, if growth rate signs remain positive (negative) for a long time then word polarities may be positively (negatively) biased and this will reflect in the indices. A way to handle this issue may be to set a threshold for the absolute value of the growth rates in the target variable so that positive (negative) signs are applied to words only if the growth rate is above (below) the threshold. Another way might be estimating the average bias in the vocabulary to correct word polarities accordingly. Anyway, also some deep learning techniques to discover subtle links between words and economic variables could be used.

Concerning the vocabulary size and composition, also the method in Chapter 3 presents limitations in considering the context of each word. This is relevant for the same reasons already discussed for Chapter 2. Moreover, sentiment analysis results are influenced by the presence of so-called *valence shifters* – i.e. words like negations which “invert” word polarity – or amplifiers (like some adverbs and adjectives) which may impact the overall sentiment of compound expressions or phrases. The analysis conducted may be extended to better allow for such linguistic peculiarities.

Moving to the indices construction and use in regression models, also in Chapter 3 only simple linear relations were tested. However, probably, the relationship between political discourse and the value of the target economic variable is not linear. Hence the analysis may be further extended also in this direction. Moreover, even national account statistics other than the GDP may be used. In this respect also the relation with current measures of sentiment or trust of economic operators may be investigated. Regarding the Italian case, the national institute of statistics (ISTAT), since 2018, publishes the daily time series of the *Social Mood on Economy Index* (SMEI) built with a lexicon-based technique on Twitter posts [1]. It would be interesting to study eventual relations between the parliamentary debate and the social media discussion, as well as their conjunct impact on the economy.

Finally, the analysis may be extended comparing the proposed procedure with the traditional sentiment analysis techniques reviewed in Section 3.2, namely, lexicon-based and deep learning-based methods to have benchmarks and make more robustness checks. Moreover, state-of-the-art methods could be combined with the proposed approach to achieve better results.

To conclude, both the analyses performed in this thesis confirmed the research hypothesis of a measurable relation between political discourse in the form of parliamentary debates and economic dynamics. Despite the issues detailed in this chapter, the first pieces of evidence seem promising. It would be interesting investigate if similar dynamics also occur in other countries to gain a better comprehension of the complex relationship between political discourse and economy.

[1] https://www.istat.it/it/archivio/219585

Appendix A

Italian stop-words

The list is an *ad hoc* stop-word list very short compared with other similar lists available online or in *NLP* software packages. It only comprehends conjunctions, articles, adjectives and pronouns of different kinds, simple and compound prepositions, interjections, the auxiliary verbs *have* (*avere*) and *be* (*essere*). It excludes all those grammatical connectives indicating negations (including negative adjectives or pronouns - e.g. *no*, *no one*, *nothing*, ...), alternative (e.g. *or*, *either*, *neither*, ...), contrast and concession (e.g. *but*, *however*, *although*, ...), temporal relation (e.g. *ever*, *never*, ...), causation (e.g. *why*, *because*, ...) as they may change the actual meaning of a sentence.

Here the detail:

- *aggettivi indefiniti - molteplicità indeterminata*: altra, altre, altri, altro, certa, certe, certi, certo, ciascuna, ciascuno, ogni, qualche, tale, tali, talune, taluni;
- *aggettivi indefiniti - qualità indeterminata*: qualsiasi, qualsivoglia, qualunque;
- *aggettivo dimostrativo*: codesta, codeste, codesti, codesto, quegli, quei, quel, quell, quella, quelle, quelli, quello, questa, queste, questi, questo, medesima, medesime, medesimi, medesimo, stessa, stesse, stessi, stesso;
- *aggettivo interrogativo o esclamativo*: che, quale, quali, quanta, quante, quanti, quanto;
- *aggettivo possessivo*: altrui, loro, mia, mie, miei, mio, nostra, nostre, nostri, nostro, propria, proprie, propri, proprio, sua, sue, suo, suoi, tua, tue, tuo, tuoi, vostra, vostre, vostri, vostro;
- *articolo determinativo*: gli, i, il, l, la, le, lo;
- *articolo indeterminativo*: un, una, uno;
- *articolo partitivo*: de, degli, dei, del, dell, della, delle, dello;
- *congiunzioni*: ed, od;
- *interiezioni proprie*: ah, ahi, ahimè, ahinoi, ahò, bah, beh, eh, ehi, ehilà, ehm, ih, mah, oh, ohé, ohi, ohibò, ohimè, olà, puah, puh, uff, uffa, uh, uhi, urrà;

- *pronome indefinito*: chiunque;
- *pronome interrogativo o esclamativo*: chi;
- *pronome personale complemento - forma debole*: ci, li, mi, ne, si, ti, vi;
- *pronome personale complemento - forma forte*: me, sé, te;
- *pronome personale soggetto*: egli, ella, essa, esse, essi, esso, io, lei, loro, lui, noi, tu, voi;
- *pronome relativo*: cui;
- *preposizioni proprie articolate*: agli, ai, al, all, alla, alle, alli, allo, dagli, dai, dal, dall, dalla, dalle, dalli, dallo, negli, nei, nel, nell, nella, nelle, nelli, nello, sugli, sui, sul, sull, sulla, sulle, sulli, sullo;
- *preposizioni proprie semplici*: a, ad, con, da, di, fra, in, per, su, tra;
- *verbo avere*: abbi, abbia, abbiamo, abbiano, abbiate, avemmo, avendo, avere, avesse, avessero, avessi, avessi, avessimo, aveste, avesti, avete, aveva, avevamo, avevano, avevate, avevi, avevo, avrà, avrai, avranno, avrebbe, avrebbero, avrei, avremmo, avremo, avreste, avresti, avrete, avrò, avuto, ebbe, ebbero, ebbi, ha, hai, hanno, ho;
- *verbo essere*: è, era, erano, eravamo, eravate, eri, ero, essendo, essere, fosse, fossero, fossi, fossimo, foste, foste, fosti, fu, fui, fummo, furono, sarà, sarai, saranno, sarebbe, sarebbero, sarei, saremmo, saremo, sareste, saresti, sarete, sarò, sei, sia, siamo, siano, siate, sii, sono.

Appendix B

Most frequent Italian names

The list of the most frequent female and male names since year 1999 is retrieved from *ISTAT* [1].

Here the names:

- *Female names*: Adele, Alessandra, Alessia, Alice, Angelica, Anna, Arianna, Aurora, Beatrice, Bianca, Camilla, Chiara, Elena, Elisa, Emma, Federica, Francesca, Gaia, Giada, Ginevra, Giorgia, Giulia, Greta, Ilaria, Ludovica, Maria, Marta, Martina, Matilde, Melissa, Mia, Nicole, Noemi, Rebecca, Sara, Sofia, Valentina, Viola, Vittoria;
- *Male names*: Alessandro, Alessio, Andrea, Antonio, Christian, Cristian, Daniele, Davide, Diego, Edoardo, Elia, Emanuele, Federico, Filippo, Francesco, Gabriel, Gabriele, Giacomo, Giovanni, Giulio, Giuseppe, Jacopo, Leonardo, Lorenzo, Luca, Marco, Matteo, Mattia, Michele, Nicolò, Pietro, Riccardo, Salvatore, Samuele, Simone, Tommaso.

[1] https://www.istat.it/it/dati-analisi-e-prodotti/contenuti-interattivi/contanomi.

Bibliography

Abercrombie, Gavin and Riza Batista-Navarro (Apr. 2020). "Sentiment and position-taking analysis of parliamentary debates: a systematic literature review". In: *Journal of Computational Social Science* 3.1, pp. 245–270. ISSN: 2432-2725. DOI: 10.1007/s42001-019-00060-w (cit. on pp. 1, 4, 53).

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. DOI: 10.1093/rof/rfaa007 (cit. on pp. 7, 13, 14, 16).

Ahelegbey, Daniel Felix, Paola Cerchiello, and Roberta Scaramozzino (2022). "Network based evidence of the financial impact of Covid-19 pandemic". In: *International Review of Financial Analysis* 81, p. 102101. ISSN: 1057-5219. DOI: https://doi.org/10.1016/j.irfa.2022.102101. URL: https://www.sciencedirect.com/science/article/pii/S1057521922000710 (cit. on pp. 2, 84).

Airoldi, Edoardo M. and Jonathan M. Bischof (2016). "Improving and Evaluating Topic Models and Other Models of Text". In: *Journal of the American Statistical Association* 111.516, pp. 1381–1403. DOI: 10.1080/01621459.2015.1051182 (cit. on p. 7).

Algaba, Andres et al. (2020). "Econometrics meets sentiment: an overview of methodologies and applications". In: *Journal of Economic Surveys* 34.3, pp. 512–547 (cit. on pp. 3, 52–54).

Amendola, Alessandra and Giuseppe Storti (2015). "Model Uncertainty and Forecast Combination in High-Dimensional Multivariate Volatility Prediction". In: *Journal of Forecasting* 34.2, pp. 83–91. DOI: https://doi.org/10.1002/for.2322 (cit. on p. 77).

Amendola, Alessandra et al. (2020). "A Model Confidence Set approach to the combination of multivariate volatility forecasts". In: *International Journal of Forecasting* 36.3, pp. 873–891. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.10.001 (cit. on p. 77).

Angelico, Cristina et al. (2022). "Can we measure inflation expectations using Twitter?" In: *Journal of Econometrics* 228.2, pp. 259–277. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2021.12.008. URL: https://www.sciencedirect.com/science/article/pii/S0304407622000227 (cit. on pp. 2, 4, 8).

Aprigliano, Valentina et al. (2022). "The power of text-based indicators in forecasting Italian economic activity". In: *International Journal of Forecasting*. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2022.02.006. URL: https://www.sciencedirect.com/science/article/pii/S0169207022000358 (cit. on pp. 2, 4, 52, 84).

Ardia, David, Keven Bluteau, and Kris Boudt (2019). "Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values". In: *International Journal of Forecasting* 35.4, pp. 1370–1386. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2018.10.010 (cit. on p. 53).

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani (May 2010). "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf (cit. on p. 54).

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. DOI: 10.1017/psrm.2018.31 (cit. on p. 8).

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. DOI: 10.1007/s10869-017-9528-3 (cit. on pp. 11, 16).

Benoit, Kenneth et al. (2018). "quanteda: An R package for the quantitative analysis of textual data". In: *Journal of Open Source Software* 3.30, p. 774. DOI: 10.21105/joss.00774. URL: https://quanteda.io (cit. on pp. 11, 14).

Bholat, David et al. (2015). *Text mining for central banks*. Ed. by Andrew Blake. Handbooks 33. Centre for Central Banking Studies, Bank of England (cit. on p. 2).

Birjali, Marouane, Mohammed Kasri, and Abderrahim Beni-Hssane (2021). "A comprehensive survey on sentiment analysis: Approaches, challenges and trends". In: *Knowledge-Based Systems* 226, p. 107134. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2021.107134. URL: https://www.sciencedirect.com/science/article/pii/S095070512100397X (cit. on pp. 3, 53–55).

Blei, David M. and John D. Lafferty (June 2007). "A correlated topic model of Science". In: *Ann. Appl. Stat.* 1.1, pp. 17–35. DOI: 10.1214/07-AOAS114 (cit. on pp. 7, 14, 15, 92).

— (2009). "Topic models". In: *Text mining: Classification, Clustering, and Applications*. Ed. by A. Sricastava and M. Sahami. Boca Raton, FL, USA: CRC Press. Chap. 4, pp. 71–93. ISBN: 978-1-4200-5940-3 (cit. on pp. 6, 12, 13).

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (Mar. 2003). "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3.null, 993–1022. ISSN: 1532-4435 (cit. on pp. 7, 14, 92).

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011). "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1, pp. 1–8. ISSN: 1877-7503. DOI: https://doi.org/10.1016/j.jocs.2010.12.007 (cit. on p. 52).

Bowden, James, Andrzej Kwiatkowski, and Dooruj Rambaccussing (2019). "Economy through a lens: Distortions of policy coverage in UK national newspapers". In: *Journal of Comparative Economics* 47.4, pp. 881–906. ISSN: 0147-5967. DOI: https://doi.org/10.1016/j.jce.2019.07.002 (cit. on p. 53).

Box, George EP and Gwilym M Jenkins (1970). *Time series analysis: forecasting and control*. 1st ed. Oakland: Holden-Day (cit. on p. 25).

Box, George EP et al. (2015). *Time series analysis: forecasting and control*. 5th ed. Wiley Series in Probability and Statistics. (See Box and Jenkins, 1970 for the original text). John Wiley & Sons. ISBN: 978-1-118-67502-1 (cit. on p. 25).

Brandt, Michael W. and Lin Gao (2019). "Macro fundamentals or geopolitical events? A textual analysis of news events for crude oil". In: *Journal of Empirical Finance* 51, pp. 64–94. ISSN: 0927-5398. DOI: https://doi.org/10.1016/j.jempfin.2019.01.007 (cit. on p. 52).

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. DOI: https://doi.org/10.1016/j.jeconom.2005.08.004 (cit. on p. 25).

Caporin, Massimiliano and Francesco Poli (2017). "Building News Measures from Textual Data and an Application to Volatility Forecasting". In: *Econometrics* 5.3. ISSN: 2225-1146. DOI: 10.3390/econometrics5030035 (cit. on pp. 52, 84).

Carboni, Marika, Vincenzo Farina, and Daniele A. Previati (2020). "ECB and FED Governors' Speeches: A Topic Modeling Analysis (2007–2019)". In: *Banking and Beyond: The Evolution of Financing along Traditional and Alternative Avenues*. Ed. by Caterina Cruciani, Gloria Gardenal, and Elisa Cavezzali. Cham: Springer International Publishing, pp. 9–25. ISBN: 978-3-030-45752-5. DOI: 10.1007/978-3-030-45752-5_2 (cit. on p. 8).

Casas, Isabel and Ruben Fernandez-Casal (2019). *tvReg: Time-Varying Coefficients Linear Regression for Single and Multi-Equations in R*. Tech. rep. R package version 0.5.3. SSRN. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3363526 (cit. on p. 25).

Catelli, Rosario, Serena Pelosi, and Massimo Esposito (2022). "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian". In: *Electronics* 11.3. ISSN: 2079-9292. DOI: 10.3390/electronics11030374. URL: https://www.mdpi.com/2079-9292/11/3/374 (cit. on pp. 4, 5).

Cerchiello, Paola and Paolo Giudici (July 2016). "How to measure the quality of financial tweets". In: *Quality & Quantity* 50.4, pp. 1695–1713. ISSN: 1573-7845. DOI: 10.1007/s11135-015-0229-6. URL: https://doi.org/10.1007/s11135-015-0229-6 (cit. on p. 4).

Cerchiello, Paola, Paolo Giudici, and Giancarlo Nicola (2017). "Twitter data models for bank risk contagion". In: *Neurocomputing* 264. Machine learning in finance, pp. 50 –56. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2016.10.101 (cit. on p. 52).

Cerchiello, Paola and Giancarlo Nicola (2018). "Assessing News Contagion in Finance". In: *Econometrics* 6.1. ISSN: 2225-1146. DOI: 10.3390/econometrics6010005 (cit. on p. 7).

Cerchiello, Paola and Roberta Scaramozzino (2020). "On the Improvement of Default Forecast Through Textual Analysis". In: *Frontiers in Artificial Intelligence* 3. ISSN: 2624-8212. DOI: 10.3389/frai.2020.00016. URL: https://www.frontiersin.org/article/10.3389/frai.2020.00016 (cit. on p. 2).

Chang, Jonathan et al. (2009). "Reading tea leaves: How humans interpret topic models". In: *Neural information processing systems*. Vol. 22. Citeseer, pp. 288–296 (cit. on p. 16).

Deerwester, Scott et al. (1990). "Indexing by latent semantic analysis". In: *Journal of the American Society for Information Science* 41.6, pp. 391–407. DOI: https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9 (cit. on p. 6).

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. DOI: https://doi.org/10.1017/pan.2017.44 (cit. on pp. 11, 61).

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (cit. on p. 3).

Dybowski, T.P. and P. Adämmer (2018). "The economic effects of U.S. presidential tax communication: Evidence from a correlated topic model". In: *European Journal of Political Economy* 55, pp. 511 –525. ISSN: 0176-2680. DOI: https://doi.org/10.1016/j.ejpoleco.2018.05.001 (cit. on p. 8).

Ehrmann, Michael and Jonathan Talmi (2020). "Starting from a blank page? Semantic similarity in central bank communication and market volatility". In: *Journal of Monetary Economics* 111, pp. 48–62. ISSN: 0304-3932. DOI: https://doi.org/10.1016/j.jmoneco.2019.01.028 (cit. on p. 52).

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as Data". In: *Journal of Economic Literature* 57.3, pp. 535–74. DOI: 10.1257/jel.20181020 (cit. on pp. 2, 3).

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3, pp. 424–438. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912791 (visited on 06/14/2022) (cit. on p. 79).

Grimmer, Justin (2010). "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases". In: *Political Analysis* 18.1, 1–35. DOI: 10.1093/pan/mpp034 (cit. on p. 12).

Grimmer, Justin and Brandon M. Stewart (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". In: *Political Analysis* 21.3, 267–297. DOI: 10.1093/pan/mps028 (cit. on pp. 1, 16).

Guan-Ru Wu, George, Tony Chieh-Tse Hou, and Jin-Lung Lin (2019). "Can economic news predict Taiwan stock market returns?" In: *Asia Pacific Management Review* 24.1, pp. 54–59. ISSN: 1029-3132. DOI: https://doi.org/10.1016/j.apmrv.2018.01.003 (cit. on p. 52).

Hannan, E. J. (1976). "The Identification and Parameterization of Armax and State Space Forms". In: *Econometrica* 44.4, pp. 713–723. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1913438 (cit. on p. 25).

Hansen, Peter R., Asger Lunde, and James M. Nason (2011). "The Model Confidence Set". In: *Econometrica* 79.2, pp. 453–497. DOI: https://doi.org/10.3982/ECTA5771 (cit. on pp. 37, 77).

Hansen, Stephen and Michael McMahon (2016). "Shocking language: Understanding the macroeconomic effects of central bank communication". In: *Journal of International Economics* 99. 38th Annual NBER International Seminar on Macroeconomics, S114 –S133. ISSN: 0022-1996. DOI: https://doi.org/10.1016/j.jinteco.2015.12.008 (cit. on pp. 2, 7).

Hansen, Stephen, Michael McMahon, and Andrea Prat (Oct. 2017). "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach". In: *The Quarterly Journal of Economics* 133.2, pp. 801–870. ISSN: 0033-5533. DOI: 10.1093/qje/qjx045 (cit. on pp. 2, 7).

Haselmayer, Martin and Marcelo Jenny (Nov. 2017). "Sentiment analysis of political communication: combining a dictionary approach with crowdcoding". In: *Quality & Quantity* 51.6, pp. 2623–2646. ISSN: 1573-7845. DOI: 10.1007/s11135-016-0412-4 (cit. on p. 53).

Hastie, Trevor and Robert Tibshirani (1993). "Varying-Coefficient Models". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 55.4, pp. 757–796. ISSN: 00359246 (cit. on p. 25).

Heston, Steven L. and Nitish Ranjan Sinha (2017). "News vs. Sentiment: Predicting Stock Returns from News Stories". In: *Financial Analysts Journal* 73.3, pp. 67–83. DOI: 10.2469/faj.v73.n3.3 (cit. on p. 52).

Hofmann, Thomas (Jan. 2001). "Unsupervised Learning by Probabilistic Latent Semantic Analysis". In: *Machine Learning* 42.1, pp. 177–196. ISSN: 1573-0565. DOI: 10.1023/A:1007617005950 (cit. on p. 6).

Istituto Nazionale di Statistica (n.d.). *I.Stat corporate data warehouse*. URL: http://dati.istat.it/ (cit. on p. 9).

Jo, Taeho (2019). "Introduction". In: *Text Mining: Concepts, Implementation, and Big Data Challenge*. Cham: Springer International Publishing, pp. 3–17. ISBN: 978-3-319-91815-0. DOI: 10.1007/978-3-319-91815-0_1 (cit. on p. 2).

Jurafsky, Daniel and James Martin (Dec. 2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. (Third Edition Draft) (cit. on p. 12).

Kao, Anne and Stephen R. Poteet (2007). "Overview". In: *Natural Language Processing and Text Mining*. Ed. by Anne Kao and Stephen R. Poteet. London: Springer London, pp. 1–7. ISBN: 978-1-84628-754-1. DOI: 10.1007/978-1-84628-754-1_1 (cit. on p. 2).

Ke, Zheng, Bryan T. Kelly, and Dacheng Xiu (Aug. 2021). "Predicting Returns with Text Data". University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2019-69, Yale ICF Working Paper No. 2019-10, Chicago Booth Research Paper No. 20-37 (cit. on p. 94).

Krippendorff, Klaus (2018). *Content analysis: An introduction to its methodology*. Sage publications (cit. on p. 2).

Larsen, Vegard H. and Leif A. Thorsrud (2019). "The value of news for economic developments". In: *Journal of Econometrics* 210.1. Annals Issue in Honor of John Geweke "Complexity and Big Data in Economics and Finance: Recent Developments from a Bayesian Perspective", pp. 203 –218. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2018.11.013 (cit. on pp. 4, 8, 9, 13, 16, 24, 25, 34, 36, 81, 94).

Larsen, Vegard H., Leif Anders Thorsrud, and Julia Zhulanova (2020). "News-driven inflation expectations and information rigidities". In: *Journal of Monetary Economics*. ISSN: 0304-3932. DOI: https://doi.org/10.1016/j.jmoneco.2020.03.004 (cit. on p. 8).

Larsen, Vegard Høghaug and Leif Anders Thorsrud (Sept. 2017). *Asset returns, news topics, and media effects*. Working Papers. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3057950 and https://ideas.repec.org/p/bny/wpaper/0054.html. Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School. DOI: https://dx.doi.org/10.2139/ssrn.3057950 (cit. on p. 7).

Leitch, Darren and Mohamed Sherif (2017). "Twitter mood, CEO succession announcements and stock returns". In: *Journal of Computational Science* 21, pp. 1–10. ISSN: 1877-7503. DOI: https://doi.org/10.1016/j.jocs.2017.04.002 (cit. on p. 52).

Li, Xiao (2020). "When financial literacy meets textual analysis: A conceptual review". In: *Journal of Behavioral and Experimental Finance* 28, p. 100402. ISSN: 2214-6350. DOI: https://doi.org/10.1016/j.jbef.2020.100402 (cit. on p. 7).

Li, Xiaodong et al. (2014). "News impact on stock price return via sentiment analysis". In: *Knowledge-Based Systems* 69, pp. 14–23. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2014.04.022 (cit. on pp. 52, 84).

Liu, Lin et al. (Sept. 2016). "An overview of topic modeling and its current applications in bioinformatics". In: *SpringerPlus* 5.1, p. 1608. ISSN: 2193-1801. DOI: 10.1186/s40064-016-3252-8 (cit. on pp. 6, 7).

Liu, Ling et al. (2015). "A social-media-based approach to predicting stock comovement". In: *Expert Systems with Applications* 42.8, pp. 3893–3901. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2014.12.049 (cit. on p. 52).

Loughran, Tim and Bill McDonald (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks". In: *The Journal of Finance* 66.1, pp. 35–65. DOI: https://doi.org/10.1111/j.1540-6261.2010.01625.x (cit. on pp. 52, 54).

— (2016). "Textual Analysis in Accounting and Finance: A Survey". In: *Journal of Accounting Research* 54.4, pp. 1187–1230. DOI: https://doi.org/10.1111/1475-679X.12123 (cit. on p. 2).

Loughran, Tim, Bill McDonald, and Ioannis Pragidis (2019). "Assimilation of oil news into prices". In: *International Review of Financial Analysis* 63, pp. 105–118. ISSN: 1057-5219. DOI: https://doi.org/10.1016/j.irfa.2019.03.008 (cit. on p. 52).

Lüdering, Jochen and Peter Tillmann (2020). "Monetary policy on twitter and asset prices: Evidence from computational text analysis". In: *The North American Journal of Economics and Finance* 51, p. 100875. ISSN: 1062-9408. DOI: https://doi.org/10.1016/j.najef.2018.11.004 (cit. on p. 7).

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. DOI: doi:10.1515/jbnst-2015-1026 (cit. on p. 7).

Manela, Asaf and Alan Moreira (2017). "News implied volatility and disaster concerns". In: *Journal of Financial Economics* 123.1, pp. 137 –162. ISSN: 0304-405X. DOI: https://doi.org/10.1016/j.jfineco.2016.01.032 (cit. on p. 2).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press. DOI: 10.1017/CBO9780511809071 (cit. on p. 12).

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Red Hook, NY, USA: Curran Associates Inc., 3111–3119 (cit. on p. 3).

Mimno, David et al. (July 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 (cit. on pp. 16, 17).

Mogaji, Emmanuel, Janarthanan Balakrishnan, and Tai Anh Kieu (2021). "Examining consumer behaviour in the UK Energy sector through the sentimental and thematic analysis of tweets". In: *Journal of Consumer Behaviour* 20.2, pp. 218–230. DOI: https://doi.org/10.1002/cb.1820 (cit. on p. 53).

Naderi Semiromi, Hamed, Stefan Lessmann, and Wiebke Peters (2020). "News will tell: Forecasting foreign exchange rates based on news story events in the economy calendar". In: *The North American Journal of Economics and Finance* 52, p. 101181. ISSN: 1062-9408. DOI: https://doi.org/10.1016/j.najef.2020.101181 (cit. on p. 53).

Pang, Bo and Lillian Lee (Jan. 2008). "Opinion Mining and Sentiment Analysis". In: *Found. Trends Inf. Retr.* 2.1–2, 1–135. ISSN: 1554-0669. DOI: 10.1561/1500000011 (cit. on p. 3).

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162 (cit. on p. 3).

Picault, Matthieu and Thomas Renault (2017). "Words are not all created equal: A new measure of ECB communication". In: *Journal of International Money and Finance* 79, pp. 136–156. ISSN: 0261-5606. DOI: https://doi.org/10.1016/j.jimonfin.2017.09.005 (cit. on p. 54).

Porter, M. F. (2001). "Snowball: A language for stemming algorithms". In: URL: http://snowball.tartarus.org/ (cit. on p. 12).

Pota, Marco et al. (2021a). "An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian". In: *Sensors* 21.1. ISSN: 1424-8220. DOI: 10.3390/s21010133. URL: https://www.mdpi.com/1424-8220/21/1/133 (cit. on p. 4).

Pota, Marco et al. (2021b). "Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets". In: *Expert Systems with Applications* 181, p. 115119. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.115119. URL: https://www.sciencedirect.com/science/article/pii/S0957417421005601 (cit. on p. 4).

Proksch, Sven-Oliver et al. (2019). "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches". In: *Legislative Studies Quarterly* 44.1, pp. 97–131. DOI: https://doi.org/10.1111/lsq.12218 (cit. on p. 53).

Quinn, Kevin M. et al. (2010). "How to Analyze Political Attention with Minimal Assumptions and Costs". In: *American Journal of Political Science* 54.1, pp. 209–228. ISSN: 00925853, 15405907 (cit. on p. 12).

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (cit. on pp. 14, 16, 25).

Rauh, Christian (2018). "Validating a sentiment dictionary for German political language—a workbench note". In: *Journal of Information Technology & Politics* 15.4, pp. 319–343. DOI: 10.1080/19331681.2018.1485608 (cit. on p. 53).

Renault, Thomas (2017). "Intraday online investor sentiment and return patterns in the U.S. stock market". In: *Journal of Banking & Finance* 84, pp. 25–40. ISSN: 0378-4266. DOI: https://doi.org/10.1016/j.jbankfin.2017.07.002 (cit. on p. 52).

Rheault, Ludovic et al. (Dec. 2016). "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis". In: *PLOS ONE* 11.12, pp. 1–18. DOI: 10.1371/journal.pone.0168843 (cit. on p. 53).

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley (2019). “stm: An R Package for Structural Topic Models”. In: *Journal of Statistical Software, Articles* 91.2, pp. 1–40. ISSN: 1548-7660. DOI: 10.18637/jss.v091.i02 (cit. on p. 16).

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. DOI: https://doi.org/10.1111/ajps.12103 (cit. on pp. 7, 16, 17, 92).

Robinson, Peter M. (1989). “Nonparametric Estimation of Time-Varying Parameters”. In: *Statistical Analysis and Forecasting of Economic Structural Change*. Ed. by Peter Hackl. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 253–264. ISBN: 978-3-662-02571-0. DOI: 10.1007/978-3-662-02571-0_15 (cit. on p. 25).

— (1991). “Time-Varying Nonlinear Regression”. In: *Economic Structural Change*. Ed. by Peter Hackl and Anders Holger Westlund. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 179–190. ISBN: 978-3-662-06824-3 (cit. on p. 25).

Scaramozzino, Roberta, Paola Cerchiello, and Tomaso Aste (2021). “Information Theoretic Causality Detection between Financial and Sentiment Data”. In: *Entropy* 23.5. ISSN: 1099-4300. DOI: 10.3390/e23050621. URL: https://www.mdpi.com/1099-4300/23/5/621 (cit. on p. 79).

Sebastiani, Fabrizio (Mar. 2002). “Machine Learning in Automated Text Categorization”. In: *ACM Comput. Surv.* 34.1, 1–47. ISSN: 0360-0300. DOI: 10.1145/505282.505283 (cit. on pp. 2, 66).

Senato della Repubblica (n.d.). URL: http://www.senato.it/3801 (cit. on p. 8).

Shapiro, Adam Hale, Moritz Sudhof, and Daniel J. Wilson (2020). “Measuring news sentiment”. In: *Journal of Econometrics*. ISSN: 0304-4076 (cit. on pp. 2, 53, 84).

Snowball (n.d.). URL: https://snowballstem.org/ (cit. on p. 12).

Stieglitz, Stefan et al. (2014). “Social Media Analytics - An Interdisciplinary Approach and Its Implications for Information Systems”. In: *Business & Information Systems Engineering* 6 (2), pp. 89–96. DOI: 10.1007/s11576-014-0407-5 (cit. on p. 2).

Stone, Philip J. and Earl B. Hunt (1963). “A Computer Approach to Content Analysis: Studies Using the General Inquirer System”. In: *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*. AFIPS ’63 (Spring). New York, NY, USA: Association for Computing Machinery, 241–256. ISBN: 9781450378802. DOI: 10.1145/1461551.1461583. URL: https://doi.org/10.1145/1461551.1461583 (cit. on p. 54).

Taddy, Matt (2012). “On Estimation and Selection for Topic Models”. In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, pp. 1184–1193 (cit. on pp. 16, 17).

Ter Ellen, Saskia, Vegard H. Larsen, and Leif Anders Thorsrud (2022). “Narrative Monetary Policy Surprises and the Media”. In: *Journal of Money, Credit and Banking* 0.0. DOI: https://doi.org/10.1111/jmcb.12868 (cit. on p. 52).

Tetlock, Paul C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market". In: *The Journal of Finance* 62.3, pp. 1139–1168. DOI: https://doi.org/10.1111/j.1540-6261.2007.01232.x (cit. on pp. 2, 52).

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy (2008). "More Than Words: Quantifying Language to Measure Firms' Fundamentals". In: *The Journal of Finance* 63.3, pp. 1437–1467. DOI: https://doi.org/10.1111/j.1540-6261.2008.01362.x (cit. on p. 52).

Thorsrud, Leif Anders (2020). "Words are the New Numbers: A Newsy Coincident Index of the Business Cycle". In: *Journal of Business & Economic Statistics* 38.2, pp. 393–409. DOI: 10.1080/07350015.2018.1506344 (cit. on pp. 1, 2, 8, 51).

Tillmann, Peter and Andreas Walter (2019). "The effect of diverging communication: The case of the ECB and the Bundesbank". In: *Economics Letters* 176, pp. 68–74. ISSN: 0165-1765. DOI: https://doi.org/10.1016/j.econlet.2018.12.035 (cit. on p. 52).

Vayansky, Ike and Sathish A.P. Kumar (2020). "A review of topic modeling methods". In: *Information Systems* 94, p. 101582. ISSN: 0306-4379. DOI: https://doi.org/10.1016/j.is.2020.101582 (cit. on pp. 3, 6, 7, 14).

Wallach, Hanna M. et al. (2009). "Evaluation Methods for Topic Models". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. New York, NY, USA: Association for Computing Machinery, 1105–1112. ISBN: 9781605585161 (cit. on pp. 16, 17).

Wehrheim, Lino (Jan. 2019). "Economic history goes digital: topic modeling the Journal of Economic History". In: *Cliometrica* 13.1, pp. 83–125. ISSN: 1863-2513. DOI: 10.1007/s11698-018-0171-7 (cit. on p. 7).

Welbers, Kasper, Wouter Van Atteveldt, and Kenneth Benoit (2017). "Text Analysis in R". In: *Communication Methods and Measures* 11.4, pp. 245–265. DOI: 10.1080/19312458.2017.1387238 (cit. on p. 11).

Whittle, Peter (1951). *Hypothesis testing in time series analysis*. Vol. 4. Almqvist & Wiksells boktr. (cit. on p. 25).

Yadav, Ashima and Dinesh Kumar Vishwakarma (Aug. 2020). "Sentiment analysis using deep learning architectures: a review". In: *Artificial Intelligence Review* 53.6, pp. 4335–4385. ISSN: 1573-7462. DOI: 10.1007/s10462-019-09794-5. URL: https://doi.org/10.1007/s10462-019-09794-5 (cit. on p. 55).

Yadollahi, Ali, Ameneh Gholipour Shahraki, and Osmar R. Zaiane (May 2017). "Current State of Text Sentiment Analysis from Opinion to Emotion Mining". In: *ACM Comput. Surv.* 50.2. ISSN: 0360-0300. DOI: 10.1145/3057270 (cit. on pp. 3, 4).

Yang, Shanxiang, Zhechen Liu, and Xinjie Wang (2020). "News sentiment, credit spreads, and information asymmetry". In: *The North American Journal of Economics and Finance* 52, p. 101179. ISSN: 1062-9408. DOI: https://doi.org/10.1016/j.najef.2020.101179 (cit. on p. 52).

Young, Lori and Stuart Soroka (2012). "Affective News: The Automated Coding of Sentiment in Political Texts". In: *Political Communication* 29.2, pp. 205–231. DOI: 10.1080/10584609.2012.671234 (cit. on pp. 53, 54).

Zeileis, Achim (2019). *dynlm: Dynamic Linear Regression*. R package version 0.3-6. URL: https://CRAN.R-project.org/package=dynlm (cit. on p. 25).