Tracking ECB's communication: Perspectives and Implications for Financial Markets

FORTES, Roberta and Le Guenedal, Theo

December 2020

Tracking ECB's Communication: Perspectives and implications for financial markets

Roberta Fortes*
Ph.D. Candidate
Paris 1 Panthéon-Sorbonne, Paris

Théo Le Guenedal*
Quantitative Analyst
Amundi Asset Management, Paris

December 2020

Abstract

This article assesses the communication of the European Central Bank (ECB) using Natural Language Processing (NLP) techniques. We show the evolution of discourse over time and capture the main themes of interest for the central bank that go beyond its traditional mandate of maintaining price stability, enlightening main concerns and themes of discussion among board members. We also built sentiment signals compatible with any form of language, both formal and informal, an important step as the ECB aims to enhance communication with non-expert audiences. In a second step, we measure the impact of the ECB's communication on the EUR/USD exchange rate. We found that our quantitative series, both topics and sentiment, improve financial-linked models consistently in all periods analyzed (2.5% on average). Meaningful signals comprise a broad range of subjects and vary in time. This suggests that overall ECB's talk matters for asset prices, including themes not directly related to monetary policy. This result is particularly important in a context in which the ECB, as well as other major central banks, are moving towards integrating issues closer to the society into their scope of action, implying that subjects, which were considered peripheral, may become central. This emphasizes the importance for markets to effectively track central banks' communication to improve investment processes.

Keywords: Quantitative trading, Central Bank, Fixed Income, Foreign Exchange, Text mining.

JEL classification: C38, C63, E44, F31, G12.

*Contact information: roberta.alves-de-paiva-fortes@etu.univ-paris1.fr and theo.leguenedal@amundi.com. The authors are very grateful to Bastien Drut, Philippe Ithurbide, Edmond Lezmi, Takaya Sekine, Lauren Stagnol and Jiali Xu for helpful comments.

Table of Contents

1 Introduction

"Since I've become a central banker, I've learned to mumble with great coherence. If I seem unduly clear to you, you must have misunderstood what I said."
Alan Greenspan, 1987.

"Today, central bank communication is at the heart of monetary policy. It is actually a monetary policy tool in itself."
Mario Draghi, 2014.

Central bank communication has evolved significantly over time. From Greenspan (1987)[1] to Draghi (2014a), we have moved from a *secrecy* period in which central banks systematically limited their communication to an environment in which discourse is not only used extensively, but has also assumed an effective role for monetary policy.

According to the literature, two main factors have created the terrain that brought about major changes observed in the way central banks voice themselves from the mid-90s. First, an overall understanding that clear communication ensured the effectiveness of monetary policy. Second, the greater independence granted to central banks. This implied the need for an increase in democratic accountability, which required central banks to explain how their decisions helped them to achieve their objectives (Blinder *et al.*, 2008; Draghi, 2014a; Yellen, 2012).

As such, the quest for transparency became primordial and major central banks have started to (gradually) disclose information. For instance, the Federal Reserve (Fed) began publishing public monetary policy statements in 1994 and the votes of individual members in 2002. From 2003, the statements also contained explicit information, or forward guidance, on the likely course of inflation. From 2005, the minutes of the Federal Open Market Committee (FOMC) meetings started to be released. The European Central Bank (ECB), which was formulated when these changes were taking shape, has adopted transparency from its inception[2]. The ECB was the first central bank to offer monthly press conferences (Draghi, 2014a) and its communication strategy generally relied on the use of *code words* such as *"strong vigilance"*, *"heightened alertness"* as a way to signal to the markets forthcoming changes in the conduct of monetary policy[3] (De Haan, 2008).

However, the 2007-2008 Global Financial crisis (GFC) and the advent of the zero lower bound (ZLB) environment in major economies have requested greater central bank trans-

[1]Speech to the Subcommittee of the US Congress, November–December 1987 (See Ratcliffe (2017)).

[2]According to Trichet (2008a), a new and bold level of transparency was important to i) ensure credibility of the new currency, the euro, ii) render uniform the communication in the euro area, avoiding various interpretations that could rise from different cultures and languages in the region and iii) establish and consolidate credibility of the institution.

[3]Markets were very attentive to the use of these words. See, for example, https://ftalphaville.ft.com/2011/02/03/478326/the-ecbs-code-words/

parency, prompting a new communication revolution. Under such environment, agents may find it difficult to foresee monetary policy decisions as signals are less clear to extract. Indeed, in addition to a highly uncertain economic environment, skepticism may arise regarding the effectiveness of the new and numerous instruments used. Figure 1 highlights the diversity of the non-standard monetary policy measures taken by the ECB since the GFC inception. Different instruments have been used over time: from innovative liquidity facilities to asset purchase programmes and negative deposit rates. In this context, to steer expectations about future policies, central banks had to provide clearer guidance to markets and be more explicit about how they understood their new environment (Coenen *et al.*, 2017; Draghi, 2014b). As a matter of fact, central banks' talks have then become one of the main instruments of monetary policy[4].

More recently, we have also noticed both a great effort and commitment of major central banks to diversify their audiences as well as a great willingness to listen to the public[5]. The combination of unconventional monetary policies and extraordinary fiscal stimulus (specially provoked by the COVID-19 pandemic crisis), has challenged public perception of central bank independence (Macklem, 2020). Strengthening trust between the public and central banks has become both desired and necessary. As such, communication has proved to be an important tool not only in influencing investor's expectations, but also in handling crisis, ensuring markets[6] and getting closer to the society.

Reflecting this growing importance, central banks have engaged in spreading their message. Regarding the ECB, there has been an increasing number of media channels over time (e.g. speeches, interviews, twitters, podcasts) as well as a larger amount of interventions given by different board members. From a purely analytical point of view, this shows that communication has become more widespread, diverse and less focused on the figure of the president of the institution. From a quantitative perspective, this constitutes an increase in the volume and frequency of information available - and which remains broadly under explored.

The objective of this paper is twofold. First, we aim at capturing and assessing the evolution of ECB's communication over time using text mining techniques. Second, from the text mining based signals extracted, we evaluate the impact of the ECB's discourse on asset prices. Even though there are several studies on the topic[7], our paper brings some novelties to the literature on central bank communication. First, research on this subject

[4] Jens Weidmann, president of the Bundesbank, recently emphasized the central role of communication to monetary policy. According to him *"ECB watchers scrutinize every world. Changes in communication are very keenly noticed and analyzed. And it is not only what is said but also what is unsaid that can send a message"* and *"the language in central bank statements about the likely future course of monetary policy (known as forward guidance) became more important than any concrete decision taken (or not taken) during the monetary policy meetings" Weidmann (2018).*

[5] See e.g. Ehrmann and Wabitsch (2020). In addition, note that the ECB held its first *ECB listens* in October 2020, an event that brought together a range of European-level civil society organizations to hear their views on the impact of the ECB's monetary policy and communication and on the global challenges ahead.

[6] See e.g. B. S. Bernanke (2013) and Trichet (2008b, 2008c). Coenen *et al.* (2017), in particular, found that announcements of asset purchase programs have lowered market's uncertainty effectively. Forward guidance too, specially when providing information for long horizons.

[7] See Blinder *et al.* (2008) for a review.

is considered relatively young mostly due to the difficulties in measuring qualitative information in a direct, transparent, objective and reproducible way (Lucca & Trebbi, 2009). We use Natural Language Processing (NLP) techniques in our study hence contributing to the growing literature that applies computational linguistics tools to analyze central bank communication. Specifically, we apply the Structural Topic Model (STM) introduced by M. E. Roberts *et al.* (2019).

STM allows estimating general semantic themes within a collection of documents and to simultaneously analyze the prevalence of each topic conditional on a number of factors. Therefore, we estimated the main topics that characterize our corpus of texts as well as the variation of their prevalence over time. This process allowed us to track the main subjects of central bank communication over time and by speaker, highlighting the main research and interests of the institution which goes beyond its core mandate of ensuring price stability. By doing so, we built and analyzed the ECB's history through quantitative (and reproducible) signals. This approach adds to the recent literature that uses NLP to identify diverse messages contained in central banks' discourse. Moreover, to our knowledge, our article is the first to trace the evolution of the ECB's communication through the speaker's perspective including extra-financial topics.

From our database, we also constructed *sentiment* signals to capture the overall emotion displayed by speakers during their interventions. To do so, we used measures based on three different common language lexicons (AFINN, NCR and BING) and a specific financial dictionary (Loughran-McDonald). Most articles normally use only one of these measures to capture the emotions possibly conveyed in a discourse. We believe that including these set of broad-based lexicons, general and financial, is important because it allows to capture the different language usage which varies according to the communication tools employed. In this way, our quantitative process generates signals compatible with any means of communication the central bank may use, whether formal (e.g. a speech) or informal (e.g. twitter)[8].

We then use these text mining based time-series to measure the effects of ECB's communication on the evolution of asset prices, notably foreign exchange, which is little explored in the literature despite the undeniable influence that central banks' discourse has on the FX market. Indeed, research has mostly focused on the impact of communication on the yield curve. More precisely, we aim at measuring the impact of the ECB's discourse on the EUR/USD exchange rate. We employ the whole set of information (topics and sentiments) extracted from its different interventions in our analysis. Our results show that the central bank's discourse exerts a systematic influence on the currency's return over time and that a variety of topics, which goes beyond monetary policy, plays an important role to explain the EUR/USD performance.

[8] We implemented Principal Component Analysis (PCA) in order to understand the correlations and variances within the different signals based on formal (financial) lexicon and informal (general) lexicons.

Figure 1: ECB's non-standard policy measures

ECB's Non-Standard Policy Measures (ECB-NSM)
Responses to Different Phases of the crisis

First Phase:

→ *ECB as a Lender of Last resort* (2008 –2009)

Objective: Provision of liquidity to banks and keep financial market functionning.

Instruments:

- LTRO (6 months): 28/03/2008
- Fixed-rate full allotment: 15/10/2008
- LTRO (1Y)/CBPP1: 07/05/2009

Second Phase:

→ *Addressing the Sovereing Debt Crisis*

Objective: Addressing market's malfunctionning and to reduce differences in financial conditions faced by business and households in different euro area countries.

Instruments:

- SMP1: 10/05/2010
- SMP2: 08/08/2011
- LTRO (3Y): 08/12/2011
- CBPP2: 06/10/2011
- OMTs: 06/09/2012

Third Phase:

→*Dealing with credit crunch and risk of deflation*

Objective: as short-term interest rates were close to the ZLB, ECB policy measures aimed at influencing the broad set of interest rates relevant for financing condition in the euro area.

Instruments:

- Forward Guidance: 04/07/2013
- ABSPP and TLTRO1: 05/06/2014
- Negative Interest Rates: 12/06/2014
- CBPP3: 04/09/2014
- Extended APP (PSPP incl.): 22/01/2015
- APP1 (60 EUR bn): 09/03/2015 extended in 03/12/2015
- APP2 (80 EUR bn) and CSPP: 10/03/2016
- TLTRO2: 10/03/2016
- APP3 (60 EUR bn): 08/12/2016
- APP4 (30 EUR bn): 26/10/2017
- APP5 (15 EUR bn): 14/06/2018
- TLTRO3: 07/03/2019
- APP6 (20 EUR bn): 12/09/2019
- Additional APP funding (120 EUR bn): 12/03/2020 (coronavirus)
- PEPP (750 EUR bn): 18/03/2020 (coronavirus)
- Additional PEPP funding (600 EUR bn): 04/06/2020 (coronavirus)

Source: ECB and authors' compilation.

2 Relation to Literature

"*Words speak at least as loud as actions*". In early 2000s, Gürkaynak *et al.* (2004) aiming at accurately assessing the effects of US monetary policy decisions on asset prices, pointed out that two major elements were necessary to tackle the task: i) the current federal funds rate and ii) the future path of policy rates, which was closely related with the Federal Open Market Committee (FOMC) announcements. As a conclusion, the authors claimed that "*words speak at least as loud as actions*". From that point, the literature on central bank communication expanded and now it comprises studies for a large number of central banks and a variety of asset classes. Our work associates to this research. While findings are mixed, studies show that central bank talk can influence asset prices in a significant manner (B. Bernanke *et al.*, 2004; Blinder *et al.*, 2008; Connolly, Kohler, *et al.*, 2004; Schmeling & Wagner, 2019).

Our research is closely related to studies specifically linked to the ECB's communication. Overall, empirical evidence suggests that the central bank communication has a significant impact on euro area interest rates for all maturities (Brand *et al.*, 2010; Ehrmann & Fratzscher, 2003). In particular, Rosa and Verga (2005) constructed an indicator based on counting *code words*[9] in the ECB introductory statement (during the press conferences) and found that communication affects short term dynamics of interest rates. In the same fashion, Ehrmann and Fratzscher (2007) found that press conferences add substantial information often exerting an even larger effect on financial markets than the release of the monetary policy decisions itself. For the authors, it was clear that 3-months Euribor futures returns, as well as trading activity, responded to the ECB's communication. According to De Haan (2008), there is substantial evidence that ECB communications increase the predictability of interest rates decisions by the Governing Council.

Stepping back from the literature which (until then) mostly relied on subjective indicators[10] to quantify central bank communication, Brand *et al.* (2010) constructed a set of multidimensional indicators of monetary policy news[11] and studied their impact on the euro area yield curve. Their results suggested that the yield curve was highly sensitive to

[9] *Code words* refer to the use of some key words such as "strong vigilance" and "extremely alert" by the ECB, in an attempt to signal markets futures changes in the conduct of the monetary policy. They were often used during Jean-Claude Trichet's mandate.

[10] Among subjective indicators, we can find indicators based on counting code words in the ECB introductory statement and the use of dummy variables to classify, for example, ECB's assessments on the 'inflation outlook' as tilted towards greater accommodation or tightening of monetary policy. Very often, in the case of the former, authors attributed a score of -1 whereas that in the later, +1. Such analysis were thus highly dependent on the individual's interpretation. Literature has been quite critical about the use of such approaches. Blinder *et al.* (2008) pointed out that as these indicators are constructed ex-post, they might mitigate the unexpected component in the statement and fail to capture how the financial market understood the message at the release time. For Brand *et al.* (2010) "*subjective indicators cannot possibly reflect all the information that is used by financial markets when forming expectations about monetary policy.*"

[11] Specifically, the authors used three different methodologies to extract monetary policy news from the money market yield curve: i) a segregation of the time windows into decision and communication windows, ii) rotated principal components and iii) recursive regressions. Note that the last two methods were inspired on early work of Gürkaynak (2005).

the ECB press conference, with more sizable and significant changes on the medium-to-long tenors. The front-end of the curve, in turn, was essentially driven by the immediate policy decision. Recent studies corroborate this result. Leombroni *et al.* (2019) decomposed ECB monetary policy surprises into *target* and *communication* shocks. The authors found that in days of monetary policy announcements, euro area bond yields are mostly driven by communication shocks with pronounced effects at medium maturities. Moreover, the authors pointed out that the ECB monetary policy was no longer homogeneous after the European debt crisis. In particular, communication shocks drove apart core and peripheral yields.

Central bank communication during unconventional monetary policy times also affects financial markets. Overall, empirical evidence suggests that the announcements of asset purchase programs (APP) and Securities Markets Programs (SMPs) impacted significantly government bonds yields, especially in the peripheral euro area economies (Altavilla *et al.*, 2019; Altavilla *et al.*, 2015). Coenen *et al.* (2017) therefore drew attention to the language used and the information disclosed during the announcements of new policies. According to the authors, statements that use more difficult language raise stock market volatility, a notable evidence for APP announcements. The authors also pointed out that it is very important to communicate the expected workings of the tools and to provide as many details about their implementation as possible (e.g. extent of the policy) otherwise uncertainty would increase. Clarity is hence key. This finding is in line with studies comprising other major central banks such as the Federal Reserve Bank (Jansen, 2011) and the Bank of Canada (Ehrmann & Talmi, 2020).

Forward guidance is also influential. Literature shows that its use in a particular economy affects local market interest rates (Hansen & McMahon, 2016). Specifically regarding the ECB, empirical research has showed that this policy has lowered the full term structure of private short term interest rates, with stronger and persistent results on longer maturities (Hubert, Labondance, *et al.*, 2018). Altavilla *et al.* (2019) found similar effects for longer-term sovereign yields. Coenen *et al.* (2017) noted that this tool seems to be particularly effective when the ECB gives guidance for long horizons and even stronger when jointly unveiled with asset purchase programmes. Taking into account a sample of eight major central banks and focusing on the formation of expectations, Sutherland (2020) provided evidence that forecasters revise their interest rate prediction in the intended direction by about five basis points (on average) in response to a change in forward guidance.

While transparency about policies has proven to be influential on markets, studies have also shown that other factors such as the language tone also matters. For instance, Schmeling and Wagner (2019) analyzed the ECB press conferences and found that a positive tone is associated with higher equity market returns, lower volatility risk premia, and lower credit spreads.

Our work is also closely related to the research that combines central bank communication and Natural Language Processing (NLP), in particular topic modeling. One approach is to use dictionary based techniques such as the one applied on Lucca and Trebbi (2009). The authors used the Google semantic orientation score to capture policy inclinations in the FOMC statements. Regarding the ECB, Picault and Renault (2017) constructed a field-specific dictionary based on the content of the ECB press conferences. Specifically, the authors proposed a term-weighting and contiguous sequence of words (n-grams) to better capture the subjectivity of central bank communication. They found that the lexicon helps to explain future ECB monetary policy decision when considering an augmented Taylor rule and that the European stock market is more volatile following a negative tone of the ECB on the economic outlook.

Although dictionary models are information based, they do not capture the context of the documents, a major limitation for text analysis. Other techniques, such as latent variables models overcome this issue. As a broad view, they assume that words are not independent but linked together by underlying, unobserved topics[12](Bholat *et al.*, 2015).

A widely known approach is the Latent Dirichlet Allocation (LDA, see Blei *et al.* (2003)). Essentially, this probabilistic topic model treats each document as a mixture of topics, and each topic as a mixture of words. Formally speaking, each topic is defined as a unique distribution of words which yields a word-topic matrix. This provides a conditional probability for every word given each hidden topic (Wesslen, 2018)[13]. Therefore, this method allows the same word to appear in multiple topics with different probabilities, whereas a standard mixture model would force each word to appear only in one topic. According to Hansen *et al.* (2018), this is one of the main advantages of the LDA. The authors used this technique to study the effects of increased central bank transparency, having the Federal Reserve as case-study and found evidence of increased conformity after the release of FOMC meeting transcripts.

Such flexibility made the LDA one of the most used methods in topic modeling with many applications. Research on this subject has evolved in the past years to allow models to also capture correlations between topics, a major weakness of the LDA. This is an important limitation as topic correlations are common in real-word data. Ignoring them limits LDA's ability to express the large scale data and to predict the new data (Cao *et al.*, 2009). A quite new, efficient and easy-to-implement approach to solve this issue is the Structural Topic Model (STM) (M. E. Roberts *et al.*, 2019), which we apply in this study[14].

In a nutshell, the STM adds richer structures to the traditional probabilistic topic models (notably the LDA) and other topic models that have extended these such as the Correlated Topic Model (CTM) (Blei, Lafferty, *et al.*, 2007). The model allows researchers to examine the relationship between topics and various covariates of interest (e.g. gender, political party, location), a key innovation. The STM has been used in a variety of

[12]See Bholat *et al.* (2015) for a review of the main text mining techniques applied for central banks.

[13]Wesslen (2018) pointed out that the probabilistic nature of the LDA works similar to the SVD used in earlier models such as the LSA. It acts as a dimensionality reduction process by reducing the information about each document from the large number of columns (words) to a much smaller number of columns (topics).

[14]A 'STM' package is available in R. The model applies a fast variational EM algorithm which contributes to fast and optimal computation. Note that while we are aware of new promising machine learning techniques for language modeling, notably Word2vec (Mikolov *et al.*, 2013) and the Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.*, 2018), we opted not to use these approaches as research has yet to understand their full capabilities. For instance, Ettinger (2020) found that while BERT can generally distinguish good from bad completions involving shared category or role reversal (albeit with less sensitivity than humans), it showed clear insensitivity to the contextual impacts of negation.

contexts, including political economy and central banking. For instance, Moschella *et al.* (2020) recently assessed the effects of negative public opinion of the ECB and found that it apparently leads the institution to expand its scope of communication beyond that designated in its main mandate. For the authors, this is an indication that the ECB listens to the public.

3 First insights on ECB's communication

3.1 Database structure

Central banks have several communication channels to interact with the public. Among them, we find a formal communication, in which policymakers make statements during regular pre-scheduled events such as monetary policy meetings and Parliament hearings. Normally, the language and communication adopted in such situations have a notion of group strategy with the objective of transmitting an institutional and unified message. For example, when the ECB President speaks at the Press Conference following the Governing Council monetary policy meeting, she (he) speaks of a decision taken on behalf of the group and not on a personal basis[15]. Yet, there are other more informal situations, such as speeches and interviews, being more irregular and in which the individual aspect can (often unintentionally) stand out. The spontaneous aspect of these events acts as an important strategic tool to emphasize some elements of a given policy or, rather, during a crisis situation to reassure markets. Said differently, during circumstances in which communicating more frequently is made absolutely necessary.

Our article covers formal and informal communication instruments used by the ECB. Precisely, our database contains three of the main tools used: i) *Press Conference*, ii) *Speech* and iii) *Interview* extracted from the ECB's website[16].

Our database covers the period of February 02, 1997 to July 31, 2020 and is organized as follows. Each communication event (speech, press conference or interview) is an observation for which the date, speaker(s), title, subtitle and content are provided. We count for 25 individual speakers and the full cleaned database allows us to access 2977 interventions of these types since the inception of the ECB. To avoid translation issues, we only considered

[15]The introductory statement, in particular, might reflect the position and the views of the Governing Council, agreed upon a word-by-word basis by its members (De Haan, 2008).

[16]Precisely, the ECB started publishing from 2019 a pre-compiled data set containing the content of all speeches delivered by its board members, which is updated every two months. See https://www.ecb.europa.eu/press/key/html/downloads.en.html. In addition, we used the database available on Kaggle, proposed by Roberto Lofaro, which contains the list links to the ECB's website for every press conference, interviews and speeches since they started to be released to nowadays. See https://www.kaggle.com/robertolofaro/ecb-speeches-1997-to-20191122-frequencies-dm/data. We checked every link in order to verify the validity of the information provided. Note that the use of this database require to scrap the html or pdf content of each link. The content was cleaned from hashtags and related topics proposed, biasing the models, and of the menu rubrics. For the *Press Conference*, a limitation is that the "Question & Answer" session still contains the questions affecting term frequencies. Ultimately, the two inputs were compared and provide similar results. Finally, the *Press Conferences* were assigned to the main speaker. In the vast majority of the cases, it was held by the President.

documents in English, which represents 94% of the sample. Hence, the remaining data contains 2815 lines. Table 1 provides a grouping of the content by speaker (individual and groups) and it classifies speakers according to their latest intervention (column 'Last')[17]. We reiterate that this table accounts for interventions off all categories.

Table 1 and Figure 2 show the main statistics. We observe that the number of interventions given by the board members has increased since the inception of the ECB, which puts in evidence the central bank efforts to broaden communication with the public. The Presidents, in particular, have a significant role in conveying the institution's message. Note, however, that the ECB communication's strategy seems to have changed by the end of Jean-Claude Trichet term in office. Apparently, until his mandate, the President was the lead communicator of the institution. Jean-Claude Trichet, in particular, stands as an outlier with 3.9 interventions alone per month on average which goes up to nearly 5 when adding the *Press Conferences* (Figure 2). These figures are far above the historical marks of the other Presidents (about 2.2 interventions alone and 2.9 when adding press conferences) and board members (around 2). The occurrence of two major crises during his mandate, the 2007-2008 Global Financial crisis followed by the 2010-2012 European Sovereign Debt crisis, could be a plausible explanation for his outstanding performance as communication plays (clearly) a major role during difficult times. It is worth noting that this fact was emphasized by Jean-Claude Trichet on different occasions in 2008[18].

Nevertheless, this explanation is certainly not sufficient as following Presidents have also faced major adversities. For instance, Mario Draghi's mandate started during the European Sovereign Debt crisis, followed by the introduction of unconventional monetary policy instruments wich led to growing skepticism about the effectiveness of monetary policy. Christine Lagarde's subsequent term has recently begun and she has had to deal with the Pandemic crisis and innumerable economic uncertainties. Just as before, these Presidents have engaged on talking and spreading the ECB's strategy. Yet, other members of the executive board have spoken more frequently during both, Mario Draghi and Christine Lagarde presidencies, which represents a change from the pattern observed in the early years of the ECB.

Specifically, we observe a general increase in the number of interventions on average – formal and informal – given by the board members. In particular, note the rise in the number of interviews granted, which was almost nonexistent in earlier 2000s. All in all, this may indicate a change in the European Central Bank's communication strategy to a model in which information is disseminated more widely, through different channels and with less focus on the President of the institution. Delivering the message of the ECB seems to have become a task of all[19].

[17] Note that due to space limitations, we opted to abbreviate the first name of the speakers.

[18] Indeed, at least three interventions given in 2008 had as main topic 'communication' emphasizing its changing role and importance in difficult times: "*Central Bank communication - if supported by an established strategy - might well be a quiet and uneventful activity. In difficult times, however, when the economic outlook darkens exceptionally and confidence falters, communication becomes even more important to explain how the central bank intends to gear its policy.*"(Trichet, 2008c). In such context, the need for more frequent communication became clear.

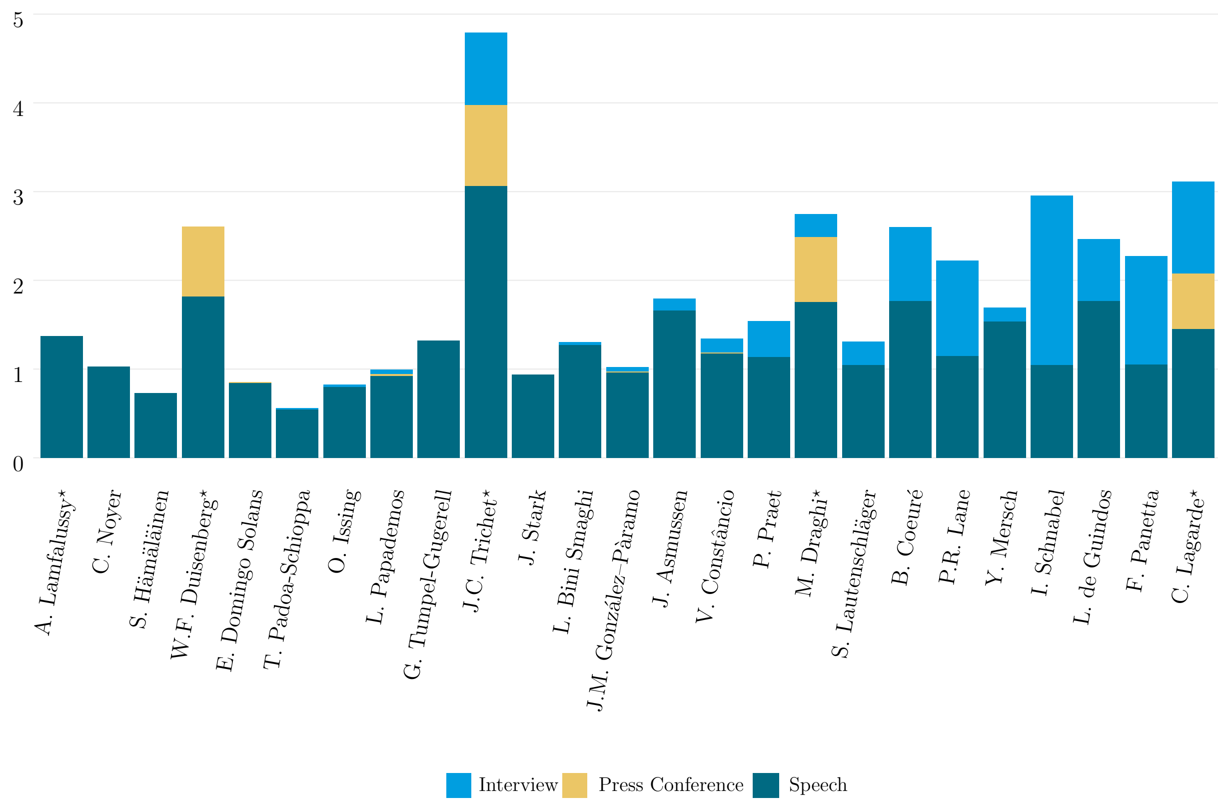
[19] Indeed, the ECB has recently introduced new communication instruments and apparently adopted a

Table 1: Speaker statistic table

Speakers	N	First	Last	Frequency per month (av)
C. Lagarde	24	04/11/2019	31/07/2020	2.5
F. Panetta	13	18/02/2020	27/07/2020	2.3
L. de Guindos	67	26/06/2018	26/07/2020	2.5
I. Schnabel	17	11/02/2020	21/07/2020	3.0
Y. Mersch	162	27/02/2013	02/07/2020	1.7
P. R. Lane	29	01/07/2019	01/07/2020	2.2
B. Coeuré	265	06/02/2012	18/12/2019	2.6
S. Lautenschläger	98	03/02/2014	30/10/2019	1.3
M. Draghi	226	18/11/2011	28/10/2019	2.2
P. Praet	159	16/06/2011	15/05/2019	1.5
V. Constâncio	137	09/07/2010	29/05/2018	1.3
J. Asmussen	40	27/03/2012	12/12/2013	1.8
J. M. González-Páramo	101	18/10/2004	18/05/2012	1.0
L. Bini Smaghi	110	27/06/2005	16/12/2011	1.3
J. Stark	66	17/07/2006	02/12/2011	0.9
J. C. Trichet	407	20/11/2003	30/10/2011	3.9
G. Tumpel-Gugerell	136	30/06/2003	19/05/2011	1.3
L. Papademos	94	07/03/2003	31/05/2010	1.0
O. Issing	85	02/07/1998	22/05/2006	0.8
T. Padoa-Schioppa	47	03/09/1998	17/02/2005	0.6
E. Domingo Solans	60	04/12/1998	23/04/2004	0.9
W. F. Duisenberg	172	30/06/1997	29/10/2003	2.1
S. Hämäläinen	43	26/10/1998	28/04/2003	0.7
C. Noyer	46	18/11/1998	23/04/2002	1.0
A. Lamfalussy	7	07/02/1997	30/06/1997	1.4
C. Lagarde, L. de Guindos	6	12/12/2019	16/07/2020	0.8
M. Draghi, L. de Guindos	11	14/06/2018	24/10/2019	0.6
V. Constâncio, M. Draghi	45	12/01/2012	26/04/2018	0.5
V. Constâncio, M. Draghi, E. Nowotny	1	02/06/2016	02/06/2016	-
V. Constâncio, D. Nouy	1	26/10/2014	26/10/2014	-
M. Draghi and I. Rimsevics	1	06/10/2013	06/10/2013	-
B. Coeuré and J. Nagel	1	18/09/2013	18/09/2013	-
B. Coeuré and J. Asmussen	1	29/07/2013	29/07/2013	-
M. Draghi, V. Constâncio, B. Coeuré	2	03/11/2011	08/12/2011	1.6
J. C. Trichet, V. Constâncio	17	10/06/2010	06/10/2011	1.0
L. Papademos, J. C. Trichet	74	06/11/2003	06/05/2010	0.9
G.Tumpel-Gugerell, L. Papademos, J. C. Trichet	1	08/03/2007	08/03/2007	-
W. F. Duisenberg, L. Papademos	14	06/06/2002	02/10/2003	0.8
W. F. Duisenberg, O. Issing, L. Papademos	1	08/05/2003	08/05/2003	-
W. F. Duisenberg, C. Noyer	26	30/03/2000	02/05/2002	1.0
W. F. Duisenberg, R. Rato, P. Solbes	1	03/01/2002	03/01/2002	-
W. F. Duisenberg, E. Domingo Solans	1	30/08/2001	30/08/2001	-

policy much in line with this view. Besides of being more active on social media (e.g. '# *ECB hashtag*' on Twitter), the central bank launched the '*ECB Podcast*' in late 2019. In March 2020, it started releasing the '*ECB Blog*' which offers insights on recent policy decisions and economic analysis. Note that while the *Blog* is authored by policymakers, in the *Podcast*, ECB's staff other than the Executive board members are often invited for discussions on major subjects affecting the euro area and external guests may also participate. Altogether, these steps certainly underline the efforts of the central bank to democratize the

Figure 2: Evolution of ECB's interventions by speaker - average monthly frequency



Source: Authors' calculations. Note that 'Press Conferences' were assigned to the main speaker, mostly the President of the institution. Speakers appear in increasing order, from the oldest to most recent, according to the latest intervention granted. Dates for both, first and last interventions by speaker, can be found in Table 1. * corresponds to the Presidents of the ECB. Note that Alexandre Lamfalussy was President of the European Monetary Institute (EMI), forerunner to the European Central Bank.

From this database, we applied text mining techniques to clean, process and analyze the unstructured data. We manually performed the preprocessing (e.g. removal of stop words and punctuation, creation of n-grams, term normalization) and extensive discussion on the different methodological steps taken is provided in the technical appendix[20].

In the next two sections, we perform both analytical and quantitative studies regarding the ECB's communication. Specifically, in section 3.2, we evaluate the transformed data from a bottom-up perspective (speaker to institution) through a standard statistical method. Then, in section 3.3, we apply the Structural Topic Model (STM) developed by M. E. Roberts *et al.* (2019) to identify the main subjects addressed by the ECB over time and extract quantitative *topic* signals.

access – but also the offer – of information to a wider public.

[20]See Appendix A.

3.2 ECB's narrative through the voices of the speakers

"Our main aim is to maintain price stability"... and beyond? The main objective of the ECB is to *maintain price stability*, a commitment that is consistently emphasized by executive board members during their interventions. But while monetary policy is and should be the center of interest of a central bank, peripheral issues also have their importance. After all, the ECB has also responsibilities in other domains such as banking supervision, systems, international relations. Each executive board member is responsible for specific areas. For instance, Benoît Coeuré, by the end of his mandate, was in charge of *Market Operations International and European Relations*. Philip Lane, is the current chief economist, responsible for *Monetary Policy and Economics*[21]. Hence, while board members normally share an unified broad view of the economy and monetary policy path, they also have some specific areas and themes of interest. This may highlight ongoing research and key topics at the center of discussion for the central bank.

Aiming at capturing these specificities, we computed the *Term frequency - Inverse Term frequency (TF-IDF)*. In a nutshell, this is a statistical measure used to evaluate the relevance of a world in a document given a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of corpus of texts[22]. It works by increasing proportionally to the number of times a word appears in a document, but is offset by the number of documents that contain the word. Hence, a word that appears many times across texts is penalized.

Figure 3 shows the TF-IDF statistics applied in all the documents in our database classified by speaker, who appear according to the last intervention given (most recent appears first). The terminologies and definitions of the acronyms can be found in the *Glossary* [23]. This figure reveals then, the key words for every speaker and, in a sense, it highlights the main topics and issues faced by the ECB throughout its existence, which goes beyond its central mission of maintaining price stability. We highlight some of them below:

European Construction Themes related to the European construction and its institutions were specially highlighted in the beginning of the ECB (e.g. Alexandre Lamfalussy, Christian Noyer, Willem Duisenberg). The so-called 'Stage Three' of the Economic and Monetary Union (or 'Stage Three EMU'), which started on January 1999, contemplated (among other factors) the introduction of the euro, the conduct of the single monetary policy by the European System of Central Banks (ESCB) besides of carrying into effect the intra-EU exchange rate mechanism (ERM II). The latter is an instrument ruling 'accession' of new members to the bloc, linking the non-euro area members states to the euro, with the objective of helping these countries to enhance policies to achieve stability and fostering

[21] A complete of view of the distribution of responsibilities among the board members and updates can be found in the ECB's website. See, e.g., https://www.ecb.europa.eu/ecb/pdf/orga/distributionofresp3_EB.pdf?4db2d62776f862d6302dc19ec0251e39

[22] Further details are provided in Appendix A.

[23] See Appendix C.

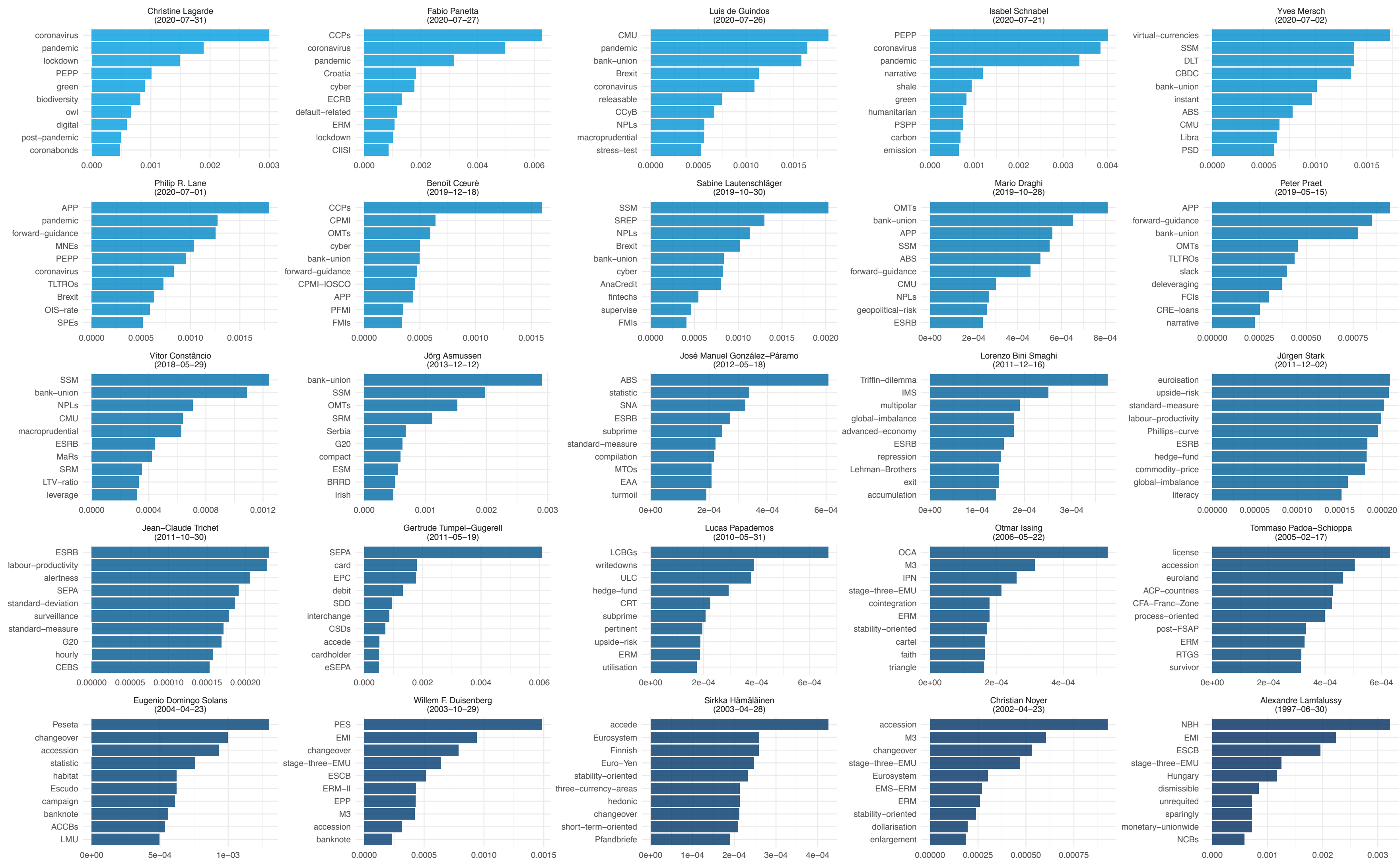
convergence. One step further toward the adoption of the European single currency.

Productivity Labor productivity also gained special attention in early 2000s and stands as a central theme for the former President Jean-Claude Trichet. In particular, he perceived the slowdown in productivity growth in the euro area as one of the most important economic challenges in the region. Differences regarding the trajectory observed in the United States were remarkable[24]. Further European integration, particularly in financial markets, reduction of the barriers to competition and continued structural reforms were perceived as key elements to help reversing the trend. Moreover, it is interesting to note the appearance of the word 'alertness' on Jean-Claude Trichet's statistics. This certainly highlights (as pointed out previously) his distinct style of communication on which *code words* (e.g.“*heightened alertness*”) were frequently used to send signals about future monetary policy path to the markets. By saying that the central bank was 'alert', he delivered the message that the ECB was ready to act promptly to the emergence of threats to the price stability.

Deepening Integration and CMU Themes related to *integration deepening* have been importantly highlighted. It includes several subjects from the enlargement of systems and payments (e.g. SEPA, eSEPA, cards) to the homogenization of statistics (e.g. AnaCredit) and unification of financial markets (See e.g. Gertrude Tumpel-Gugerell, Jean-Claude Trichet; Jörg Asmussen; Vitor Contâncio, Sabine Lautenschläger, Luis de Guindos). Regarding the latter, note that following the euro area decision to move towards the Bank Union on June 2012, subjects related to this matter such as the Single Supervisory Mechanism (SSM), which grants the ECB the supervisory authority over all banks within the EU member states, became key (e.g. Jörg Asmussen; Vitor Contâncio). More recently, the implementation of the Capital Markets Union (CMU) agenda, aiming at fostering deep and diversified capital markets in the European Union amplifying financing options in the region, has been an important topic for the European Commission and to the ECB. Notably, the advent of the Brexit implying the departure of the largest financial center in the EU, imposed challenges and reassessment to this agenda. The vice-president, Luis de Guindos, has given particular attention to these matters.

[24]In the 1980s, annual growth in hourly productivity was 2.5% in Europe and only 1.3% in the United States. But over the period 1996-2004, hourly productivity growth in the United States rose from 1.3% to 2.5%, whereas it fell from 2.5% to 1.3% in Europe. Facing these figures, Trichet stated “*I am perplexed as to the spectacular decline in productivity gains observed in Europe, despite the fact that we live in the same technological world and the same global environment.*” (“Banque de France International Symposium: Productivity, competitiveness and globalisation”, 2005). Jean-Claude Trichet identified structural rigidities within the euro area as the main constraint to the improvement in this figure, rather than euro area's failure to ensure widespread technological innovation (often used as an argument).

Figure 3: Term-frequency inverse document frequency – identifying speakers



Notes : The date under the speakers refers to the latest intervention they gave at the time of the study. See definitions for the acronyms on Appendix C.

Names of authors assigned in citations (e.g. studies, articles, events) and cross-references were removed manually in the preprocessing.

Crisis and Emergency Programs The 2007-2008 Global Financial Crisis and the subsequent European Sovereign Debt crisis were also (unsurprisingly) addressed by all board members. Issues on asset-backed securities and subprime were articulated by José Manuel Gozalez-Paramo while macroprudential policies and bank risk assessment (e.g.NPLs, LTV-ratio) were central to Vitor Constâncio. The Outright Monetary Transactions (OMTs), an emergency bond-buying program to protect the euro area from speculation that could have forced some countries out of the single currency during the European crisis, specially characterizes Mario Draghi's interventions. The OMTs were announced one week after the notorious *26 July, 2012 speech*, on which Mario Draghi stated that "*Within our mandate, the ECB is ready to do whatever it takes to preserve the euro. And believe me, it will be enough.*" This program can be considered as a concrete manifestation of Draghi's commitment to save the currency. Indeed, the former President is credited with rescuing the euro given his decisive posture in a time when the risk of the euro area breaking up was extremely high[25].

Unconventional Monetary Policy and Digital world More recently, asset purchase programmes and forward guidance are key subjects, especially for the chief economists, (former) Peter Praet and (current) Philippe Lane. Note that in addition to monetary policy, themes related to digital currencies and underlying technology have also been part of recent discussions within the ECB, lead by Yves Mersch.

Pandemic The Coronavirus pandemic has been (as expected) exhaustively highlighted, being a major topic for all board members. Note, however, that the Pandemic Emergency Purchase Programme (PEPP), launched in response to the risks imposed by the COVID-19, has been particular debated by Isabel Schnabel and Christine Lagarde.

Climate Change President Lagarde has brought a '*green touch*' to the ECB. She has given particular attention to environment and climate change matters, putting green policy in the top of the ECB's agenda. The topic has been central in her discourse since her confirmation as ECB president: "*Any institution has to actually have climate change risks and protection of the environment at the core their understanding of their mission*" and she elected the subject to be a "*mission-critical priority*" for the ECB[26].

3.3 Enhanced topic analysis

"*Central banking has never been a static business.*" The Structural Topic Model (STM) allows the estimation of main topics that characterizes our corpus of texts and their corresponding proportion in every document. We used their evolution over time to generate

[25] See e.g. https://www.ft.com/content/a62b221c-eb64-11e9-a240-3b065ef5fc55 and https://www.bloomberg.com/news/features/2018-11-27/3-words-and-3-trillion-the-inside-story-of-how-mario-draghi-saved-the-euro. Ultimately the OMTs were never activated.

[26] See https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/ECON/PR/2019/09-04/1187645EN.pdf.

topic analysis and assess whether some topics prevalence helps to explain asset prices fluctuations at a determined point in time[27].

Note that there is no 'right' answer to the number of topics that are appropriate for a given collection of texts (Grimmer & Stewart, 2013; M. E. Roberts *et al.*, 2019). This implies, hence, that this variable is set according to the author's discretion[28]. In our work, we opted to use the selection strategy based on the work of D. Mimno and Lee (2014) which has the advantage of automatically selecting the number of topics. As highlighted by M. E. Roberts *et al.* (2019), this procedure should not be seen as estimating the "true" number of topics, but it can be a useful place to start and has the computational advantage that it needs to be run only once. The methodological choice was to apply a spectral initialization as recommended by M. E. Roberts *et al.* (2019), as the authors found it produces the best results consistently[29]. From this procedure, we obtained a total of 87 topics[30].

Figure 4 shows the main topics found and their associated top ten words. In order to simplify visualization and assessment, we classified them in broader subjects, in a total of 8 clusters. They are:

1 '*Banks, Regulation and Financial Instability*';

2 '*Monetary Policy and Inflation*';

3 '*Trade and International Issues*';

4 '*Payments, Settlements and Statistics*';

5 '*EU construction, EMU setup and Enlargement*';

6 '*Growth, Employment and Structural reforms*';

7 '*Fiscal Policy, Debt and Sovereign crisis*';

8 '*Environment and Social Concerns*'.

[27] See Appendix A.4 for technical details.

[28] There are several tests available to help choosing the number of topics such as calculating the held out likelihood (Wallach *et al.*, 2009) and performing a residual analysis (Taddy, 2012).

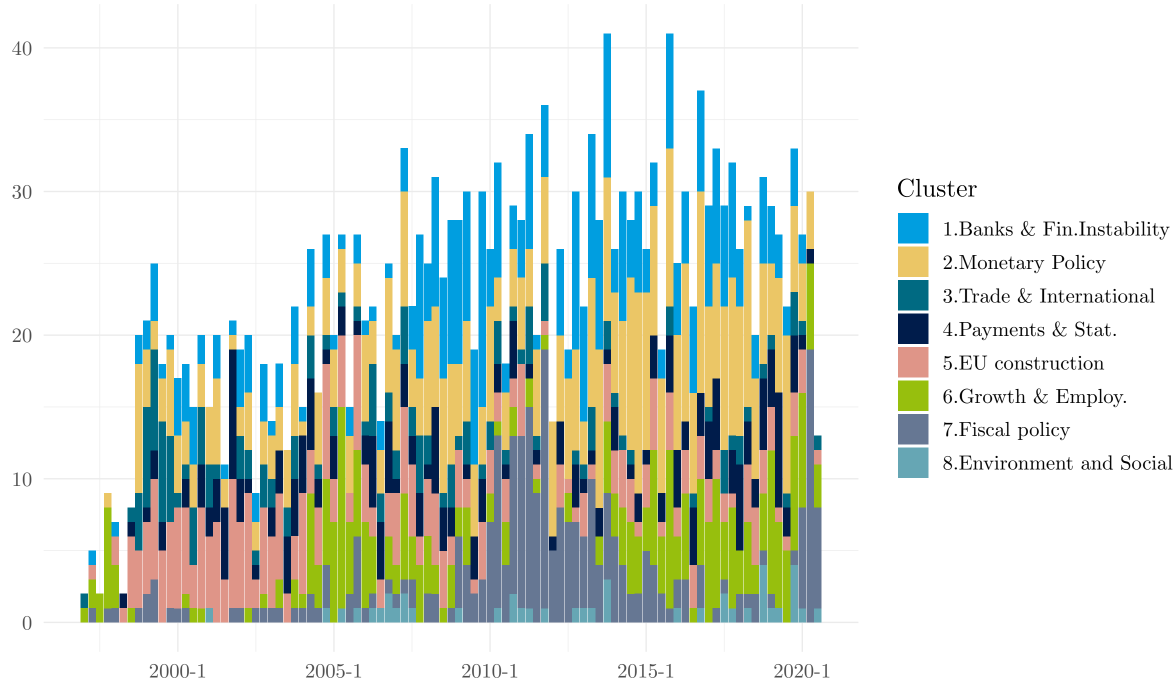
[29] Note that, from an operational point of view, using text mining in investment processes requires retraining the model and testing its predictive performance on "future", to improve its predictive capabilities, which is the cross-validation step. In this paper, however, we use the full sample and no machine learning or cross validation algorithm, mostly due to the lack of observations for this exercise to be truly meaningful. However, from an analytical perspective, our study to transform the ECB's communication into time-series, providing multiple interesting insights.

[30] As robustness check, we tested four other techniques to identify the optimal number of topics. They all converge towards a number between 70 and 100 topics. See Figure 22 on Appendix A.4.

Figure 4: Topics top betas tf-idf (characteristic words)

Structural Topic Model (STM), Spectral initiation (K=0), manual preprocessing

Figure 5: Shift in main topic composition



Source: Structural Topic Model (STM). Authors' calculation.

Figure 5 shows the evolution of these aggregate themes over time[31]. They are broadly in line with the findings and analysis showed in the TF-IDF statistics. We can see that *European construction* thematic was predominant in the beginning of the central bank and the euro area, an important phase of adjustments and continued implementation of processes aimed at deeper European integration (e.g. the Stage Three EMU). *Monetary policy* is, of course, a major topic for the ECB. A certain intensification of its relative importance has been noted more recently, possibly due to the introduction of non-conventional monetary measures such as the quantitative easing and the forward guidance in which communication is particularly necessary.

Issues related to *banking and financial risks* have also always been present, but with a sharp increase in its relative participation after the Global Financial crisis, in which the need for greater financial and banking regulation became self-evident. *Fiscal* matters gained strong prominence during the European Sovereign Debt crisis, in which the increase in indebtedness of some countries (notably Greece, Portugal) was so pronounced that investors called into question their ability and willingness to repay their debt.

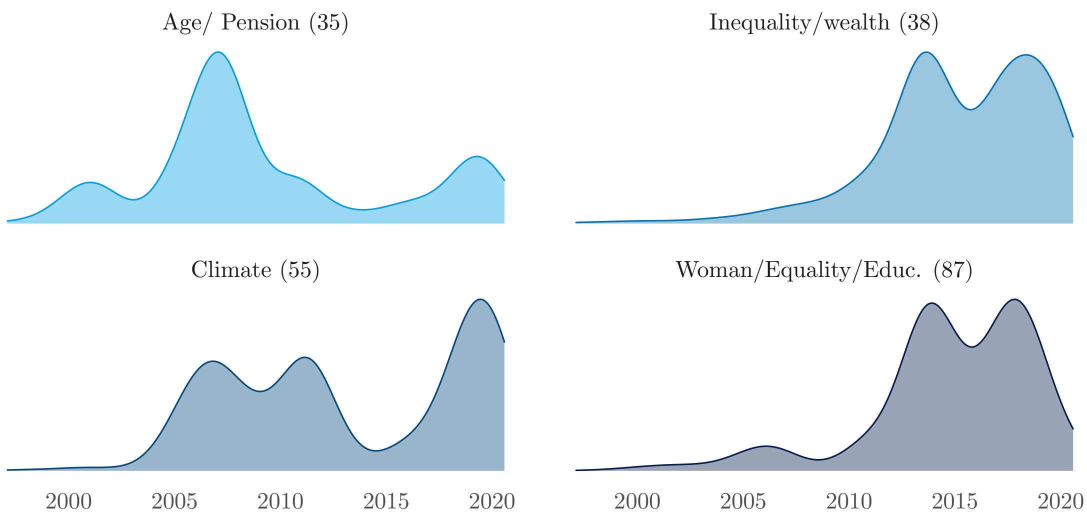
Environment and social issues started to be tackled more significantly in the past decade (Figure 6 for details) and we note that the occurrences for these subjects were higher than traditional themes such as fiscal policy for certain dates. We expect these subjects to increase in importance from now onward following President Lagarde's commitment to

[31]See Appendix A.4.3 for the underlying topics of each cluster.

these matters underpinned by the support of the European Parliament[32]. Indeed, as part of its strategy review, the ECB aims to encompass new challenges that people value and care about like climate challenge or inequality (Lagarde, 2020).

As Alexandre Lamfalussy, former President of the European Monetary Institute, forerunner of the ECB, once highlighted "*Central banking has never been a static business. Throughout its long history it has performed different tasks in different periods.*"[33]

Figure 6: Shift in '*Environment and Social Concerns*' discourse



Source: Structural Topic Model (STM). Authors' calculation.
The original sparse distributions of probability of presences of each topics are displayed on Figure 36 on page 76. The vertical axis gives the relative probability of presence of the topic, normalized per its own maximum prevalence score. The normalized value follows: $\gamma_{k,t}/\max_t(\gamma_k)$.

[32]Specifically, on February 12, 2020, the European Parliament approved (by a large majority) for a resolution recommending the ECB to look at ways in which central banks can tackle the climate change: "*As an EU institution, the ECB is bound by the Paris Agreement on climate change and that this should be reflected in its policies, while fully respecting its mandate and its independence; welcomes the emergence of a discussion about the role of central banks and supervisors in supporting the fight against climate change; calls on the ECB to implement the environmental, social and governance principles (ESG principles) into its policies, while fully respecting its mandate and its independence.*". See https://www.europarl.europa.eu/doceo/document/TA-9-2020-0034_EN.pdf.

[33]See Lamfalussy (1994).

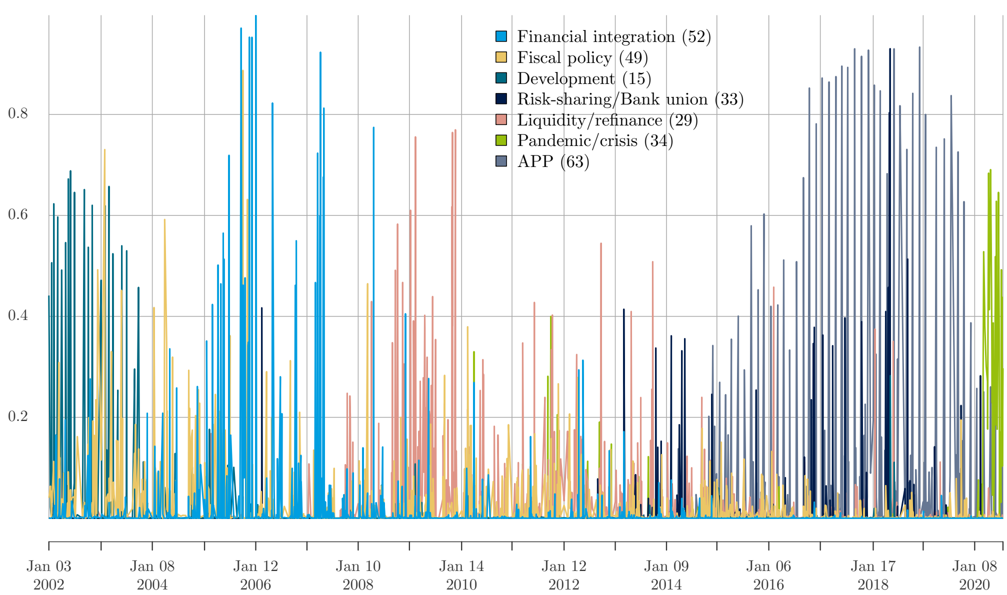
4 Financial Markets Modeling

4.1 Explanatory variables of interest

Using topic modeling and dictionary analysis, we extract two quantitative signals of interest to be considered in our EUR/USD model.

Topic signals The STM provides a probability of presence of each topic in each intervention. The metric representing this probability is called gamma[34]. There is a bijection between document (e.g. speech) and dates. Therefore, we used the gamma distributions by date as a time-series proxy for the presence of a determined topic in a discourse. An illustration of the topics dispersion is provided Figure 7. Topic 15 – development/ growth/ construction of the European alliance – is mostly concentrated in the beginning of the period. Topic 34 – Coronavirus / Pandemic –, in turn, is naturally present more recently. Note, however, that some topics are not distributed this way, being more persistent over time. An example could be Topics 49 (Fiscal policy) and 52 (Financial Integration) while Topic 29 (Liquidity) has been addressed since the Global Financial Crisis. Topic 63 (APP) highlights the ECB's quantitative easing program, started on 2015.

Figure 7: Example of topic probability over time (2002-01-03 to 2020-07-31)

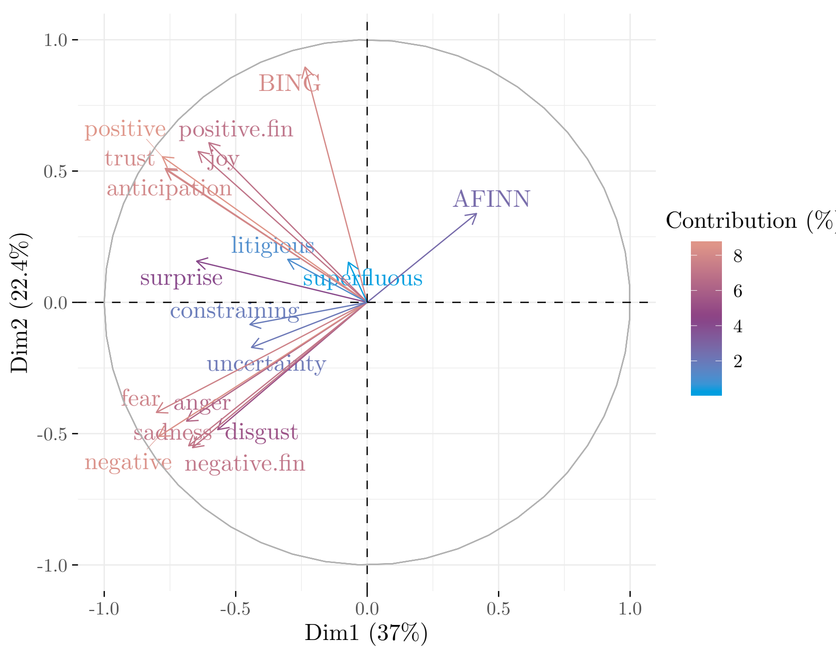


Source: Structural Topic Model (STM). Authors' calculation.

[34]See Appendix A.4 for more information on Topic Modeling.

Sentiment signals We compute a broad set of scores, based on both general lexicons (e.g. AFINN, BING, NRC[35]) and a specialized financial dictionary (Loughran-McDonald[36]), totalizing 18 series. Our purpose is to take into account the different style of discourse – formal or informal – used by central banks which may vary depending on the audience and type of communication media chosen. As we showed previously, they are becoming increasingly diverse.

Figure 8: Projection on the PC1 and PC2

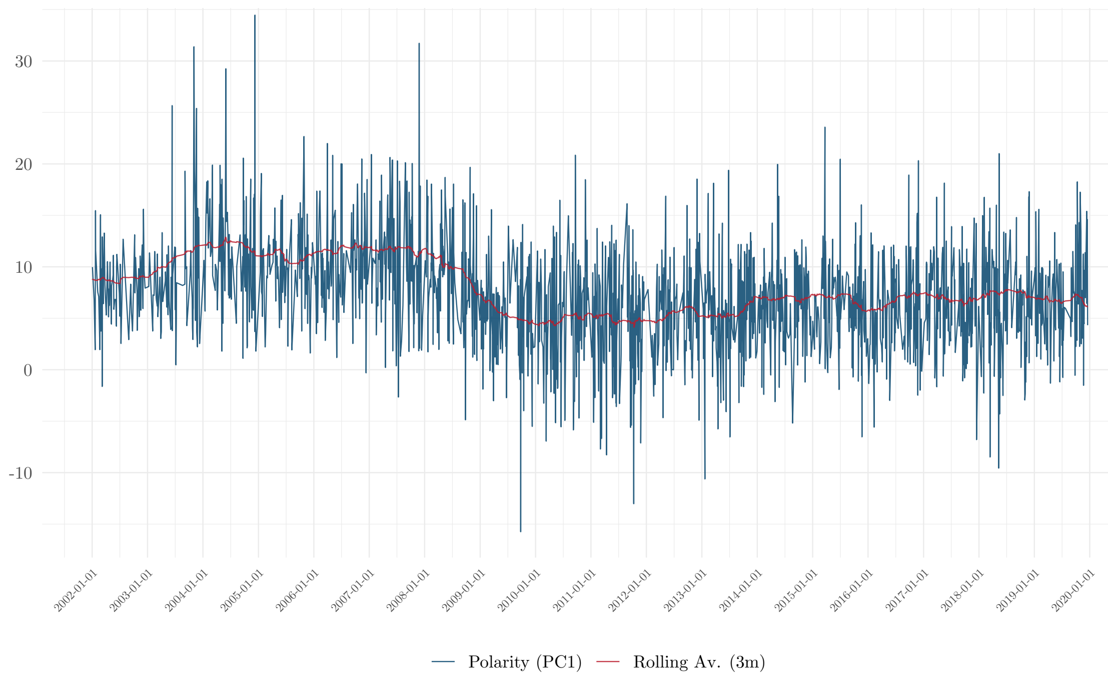


To reduce the dimensionality of these dictionary-based series, we perform a Principal Component Analysis (PCA). Figure 8 shows that the PC1 captures well the binary (positive and negative) polarity dimension[37] and Figure 9 evidences that the average tone employed by the ECB shifted from enthusiastic to more neutral since the Global Financial crisis[38].

However, we can also note that the first two components do not capture well the Loughran-McDonald dictionary which is known to have value added when applied in financial contexts. Therefore, we take all the sentiment variables into account in the quantitative analysis.

[35]The NRC Emotion Lexicon has eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive).
[36]The Loughran-McDonald dictionary has six categories: negative, positive, uncertainty, litigious, constraining and superfluous.
[37]See Appendix A.3 for details in building the scores and for details on the PCA methodology.
[38]Armelius *et al.* (2020) found a similar shape.

Figure 9: Polarity component - PC1 (2002-01-03 to 2020-01-01)



Source: Authors' calculation.

4.2 EUR/USD return as dependent variable

The economic literature comprises a wide range of models to assess the value of exchange rates on the middle and long run which are based on macroeconomic fundamentals (e.g. Behavioral Equilibrium Exchange Rate (BEER) model and Fundamental Equilibrium Exchange Rate model (FEER)[39]. In the short term, exchange rate variations are notoriously more difficult to explain and forecast[40].

Exchange rates fluctuate in the short run due to a wide range of factors, notably interest rates differential, economic surprises, equity market performance, but also by more 'qualitative' factors such as political uncertainties and central bank communication. Figure 10 highlights this point. It shows a 4 years time-window for the EUR/USD exchange rate (from mid-2014 to mid-2018), a period on which a set of 'qualitative' events played a major role to the currency's trajectory. Precisely, from June 2014 until the beginning of 2015, the euro was especially driven by the expectations, announcement and implementation of unconventional monetary policy measures by the ECB.

From the beginning of 2015 until the end of 2016, a new (lower) level for the euro was set and the currency traded around the 1.10-1.15 range – excluding the four times in which the

[39] In particular, the BEER model tries to find a long term relationship between the real effective exchange rate (REER) and different macroeconomic fundamentals, such as terms of trade or the productivity of an economy. The FEER model, takes into account the dynamics of the balance of payments into account. See e.g. Clark and MacDonald (1999) and Cline, Williamson, *et al.* (2008) for more details.

[40] See Rossi (2013) for a review.

1.05 floor was 'tested'. Note that the events driving the currency to such a low level were mostly related to news about quantitative easing measures (in the second episode, there were also fears of a Greek default, which was dismissed by the ECB). From November 2016, the currency traded lower, on the 1.05-1.10 range, a combination of expectations of better economic momentum in the US – as Donald Trump won the US presidential elections a massive fiscal stimulus was expected - and an extreme monetary policy easing in the euro area. Moreover, we should note that the region had a challenging political calendar in the beginning of 2017 with, for example, rising risks of populism in France (French presidential elections).

From the middle of 2017, ECB's communication pointing out 'confidence' regarding the macroeconomic scenario supported the euro above 1.15. Later, in January 2018, ECB's Minutes pointing out that the central bank would change its communication to a normalization earlier than expected drove the currency above 1.20. Finally, in late April 2018, the rise of populist government in Italy weighted on the euro and the currency depreciated against the dollar and breaking below 1,20.

While some might argue this is a quick extrapolation of causal-relationship-effects and that other factors might also have played a role during a specific event-date, the chart view highlights that some qualitative information does affect the euro. Therefore, from the topic analysis, we aim to quantify the ECB's communication and assess its effects on the euro on the days an intervention took place.

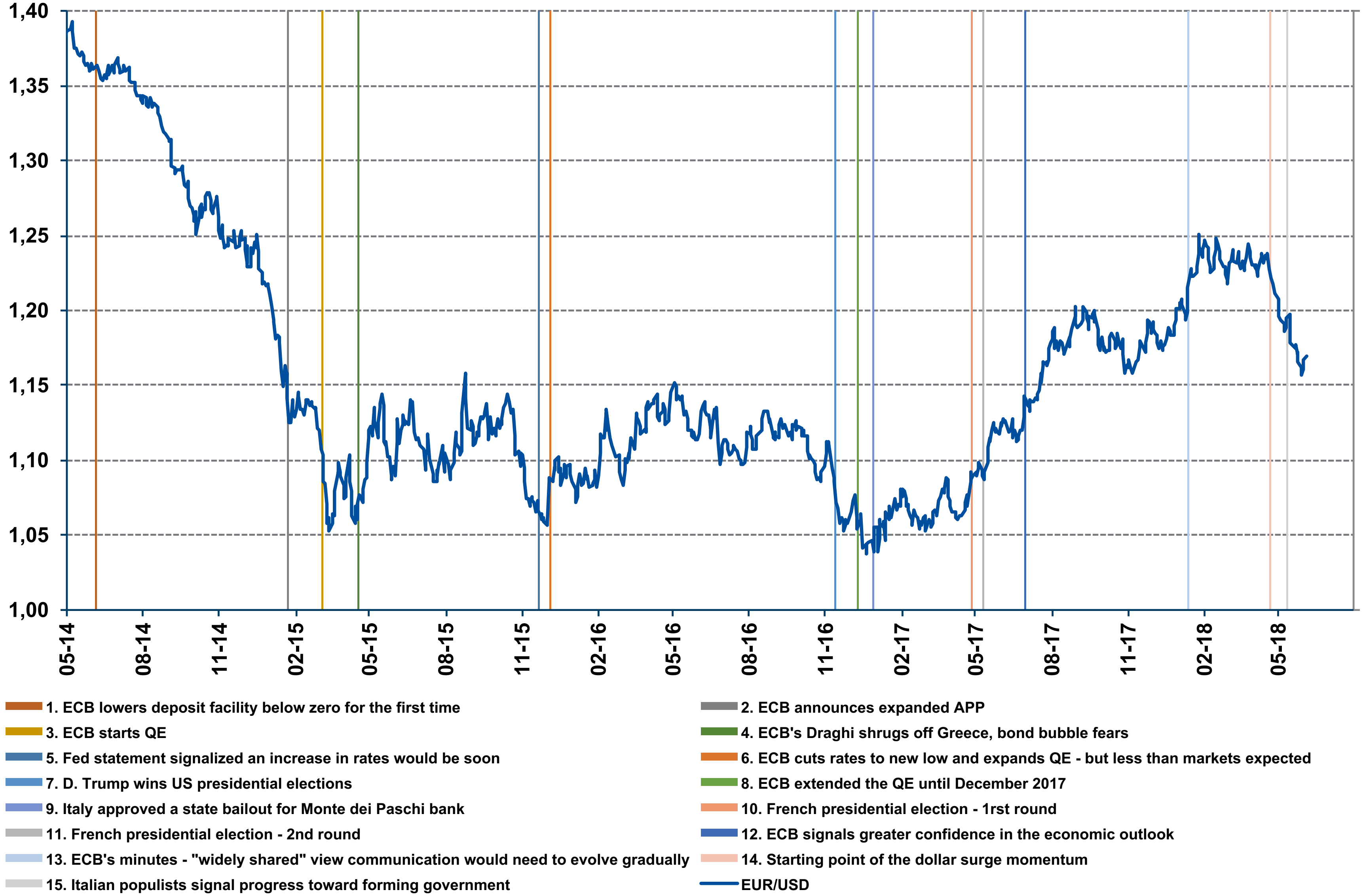
4.3 Control variables

To test the value added of our signals, we introduced a set of (mostly) market-based control variables, as they have the advantage of being measure in daily basis. Economic variables are, of course, very important to exchange rate but they are, in turn, often measured and/or available at lower frequency (monthly or quarterly) basis (e.g. industrial production, current account), which represents a major drawback for our exercise as we would like to explain the euro's return in the very short term, at some specific days.

That been said, we made full-use of the economic variables that could be found and/or approximated at higher frequency such as the Economic Surprise Index, which could be a good indicator for growth and drive short term deviations of the currency. We also took into account a set of interest rates for the Euro area (e.g. Eonia, MRO) including short and long-term bond yields (2Y and 10Y) for different economies within the region, from the core (France, Germany, Netherlands) and peripheral countries (Spain, Portugal, Italy) as well as the OIS curve, which is the best proxy for a risk-free yield curve in the euro area[41]. Stock markets return (STOXX 50) and risk aversion indicators (volatility of the German stock market - VDAX and VSTOXX, as the euro area's implied stock volatility) were also considered. The complete list of the control variables used in our model(s) is provided in Appendix D.

[41] See e.g. De Santis and Stein (2016) and Lane (2020 (accessed October 9, 2020)).

Figure 10: EUR/USD - Selected Window



Source : Datastream; Author's indications.

4.4 Quantifying the impacts

Modeling EUR/USD return We aim at assessing the value-added of the quantitative signals extracted from the ECB's communication on the EUR/USD time series. To control for general market effects and in line with the literature, we considered only days in which the ECB made an intervention (speech, interview or press conference). The full sample is daily, spans the time from 01 January 2002 to 31 July 2020, comprising a total of $N = 167$ variables[42]. The general structure of the model over the full database can be expressed as follows:

$$\Delta \text{EUR/USD}_t = \beta Z_t + \epsilon_t$$

where Z_t is the matrix gathering the 167 time series. We identified four major groups of explanatory variables: i) '*Bond yields and spreads*'; ii) '*Risk-free rates*'; iii) '*Macro-financial indicators*'and iv) '*ECB's communication*'. Thus, this equation becomes:

$$\Delta \text{EUR/USD}_t = \underbrace{\sum_{j \in \mathcal{A}} \beta_{j,T} \Delta y_{j,T,t}}_{\text{Bond yields and spreads}} + \underbrace{\sum_{T} \beta_T \Delta \text{OIS}_{T,t}}_{\text{Risk-free rates}} + \underbrace{\sum_{l \in \mathcal{M}} \beta_l \Delta \text{FC}_{l,t}}_{\text{Macro-financial indicators}} + \underbrace{\sum_{k \in \mathcal{T}} \beta_k \gamma_{k,t} + \sum_{s \in \mathcal{E}} \beta_s Sm_{s,t}}_{\text{ECB's communication}} + \epsilon_t \quad (1)$$

where Δ is the operator of time differentiation, $\mathcal{A}$ is the set of countries/areas, y_j is the bond yield of Country j[43], T is the maturity, OIS corresponds to the OIS rates in the euro area, $\mathcal{M}$ is a set of macro-financial control variables FC_l, $\mathcal{T}$ is the set of topic prevalence series γ_k , $\mathcal{E}$ is the set of sentiment series Sm_s.

To measure the impact of the communication series on the euro, along with other macro-financial variables, we perform a Lasso regression similarly to Bennani *et al.* (2018). As the database is large, we fix the degree of freedom to a maximum of 15 variables[44]. The penalized Lasso estimate is defined by:

$$\begin{aligned} \hat{\beta}_{lasso} &= \text{argmin}_\beta \sum_{i=1}^{N} \left(\Delta \text{EUR/USD} - \beta_0 - \sum_{n=1}^{N} \beta_n Z_n \right)^2 \\ \text{s.t} \quad & \text{Card}\big(\beta_n \neq 0\big) < 15 \end{aligned} \quad (2)$$

[42]We considered all the topics built up, with the exception of those in the '*4.Payments and statistics*' cluster, which highlights, from an operational point of view, the introduction of payment systems in the euro area. We opted not to take it into account to eliminate potential noise in the model since this cluster is not directly related to market movements. In addition, its connotation of *euro area integration* is already captured by other topics contained in the cluster '*5.EU construction, EMU setup and Enlargement*'.

[43]It can also represent the spread against the German bond yield for the same maturity.

[44]In this context, it is necessary to normalize the series as we working with signals in different units. For instance, macro-financial variables are daily returns in bps or percentage, *topics* indicate probability of presence in a corpus and *sentiment* represents the fraction of words belonging to each dictionary.

subject to $\sum_{n=1}^{N} |\beta_n| < \frac{\lambda}{\lambda_0}$ where $\lambda_0 = \sum_{n=1}^{N} |\beta_n^{\text{OLS}}|$ (β_n^{OLS} the least squares estimates) and λ ranges from 0 to t_0[45].

Topics generated are gathered in time (as it can be observed in Figure 7), thus, we decided to distinguish three time periods in line with what is used in the literature:

- **Pre-crisis**: January 01, 2002 to July 30, 2007;
- **Crisis** (Global Financial Crisis and European Sovereign Debt crisis): August 01, 2007 to September 30, 2012;
- **Post-crisis**: October 01, 2012 to July 31, 2020.

As an example, Figure 11 shows variables selection for the post-crisis period. From this process, we select the most significant variables to perform a ordinary least square regression. For each period, we compare the outcomes with and without the inclusion of time series signals based on central bank communication.

4.5 Results

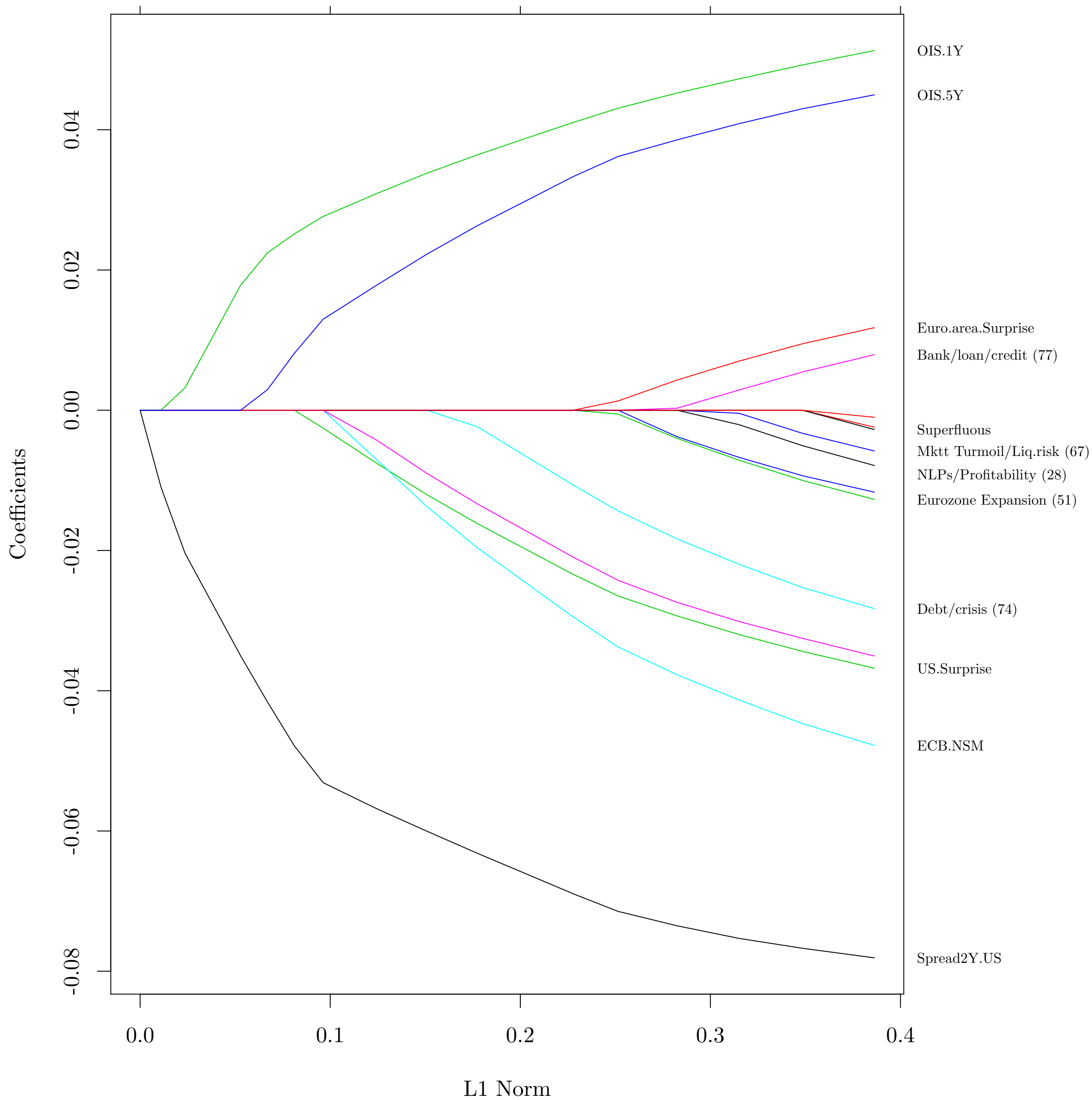
Table 2 shows the ordinary least square estimates for the EUR/USD based on this preliminary Lasso selection. Overall, while the variables that explain the euro's return change over time, the three sub-periods under study comprise communication-related variables. Indeed, the topic-based time series, constructed with the STM model, are systematically selected out of 167 variables, meaning that ECB's talk does matter for the euro. Indeed, their inclusion in the regressions improve a traditional financial model in an important manner in all periods (on average 2.5% to the adjusted R-squared). Note, in particular, that the last period exhibits larger gains as well as a larger number of topic signals selected. This may highlight the growing importance of the ECB's communication during the zero lower bound environment, the introdcution of asset purchase programmes and the adoption of forward guidance. Entering the model details, we present the main findings for each of the four blocs of variables as it follows:

Bond yields and spreads We observe that interest rate differentials between the United States and Germany (Spread2Y.US) impacts significantly and negatively the euro. This is in line with the carry trade strategies and the empirical literature which evidences that an increase in the interest rate in the home country, induces capital inflows and appreciates the domestic currency[46]. This finding holds only for the crisis and post-crisis periods. The latter was particularly marked by monetary policy divergences between the US and the euro area. For instance, while during a Congress hearing in May 22 2013 former Fed chairman, Ben Bernake, announced the central bank could slow the pace of its asset bond purchases,

[45] We use the glmnet R package to perform this operation. See e.g. Friedman *et al.* (2010) for further details. Note that equation 2 is given under the consideration of $\alpha = 1$, which corresponds to the Lasso penalty. Other possibilities for α values, e.g. the ridge-regression ($\alpha = 0$), were not explored.

[46] See e.g. Nadal-De Simone and Razzak (1999).

Figure 11: Δ EUR/USD – Post-crisis Lasso regression



Source: Authors' calculation.

Period: October 01, 2012 to July 31, 2020.

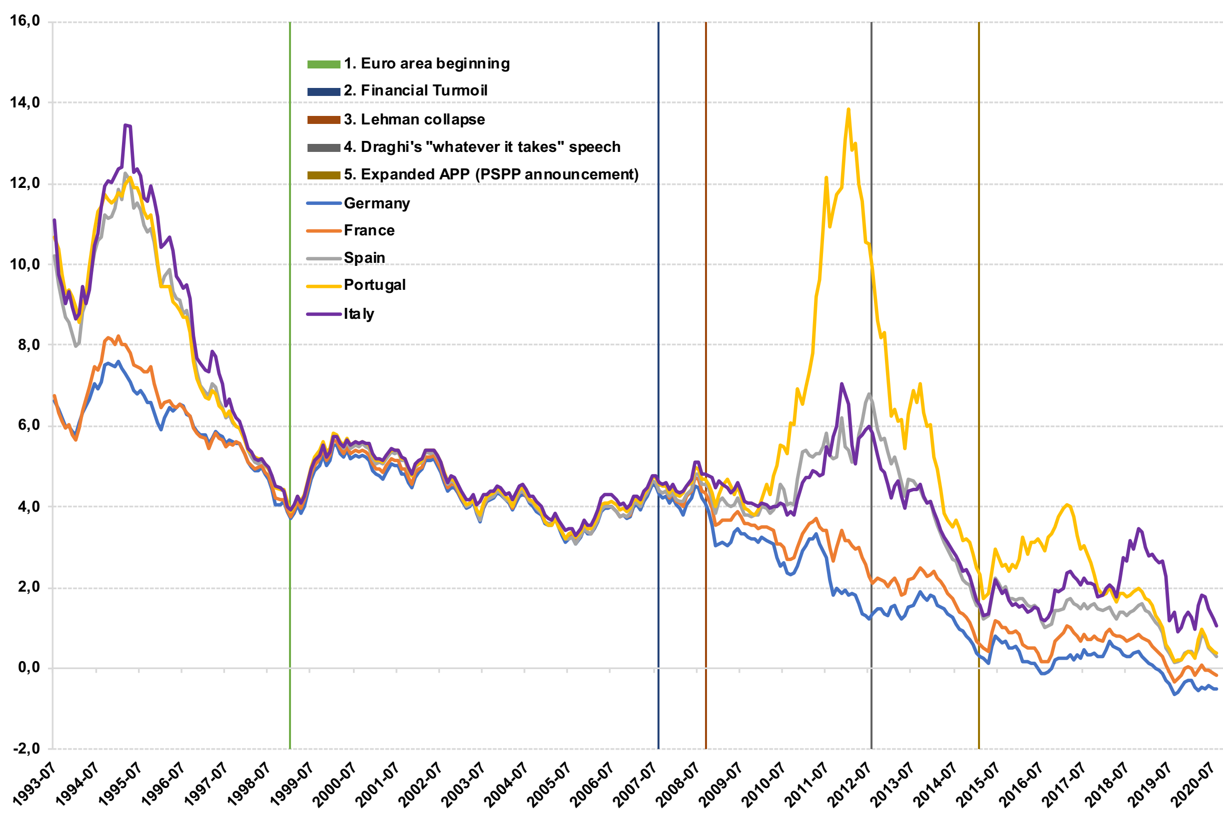
the ECB started signaling in early 2014 that a quantitative easing would be on the way to fight deflation and lift the economy.

Looking at the crisis period, we note that the spread between Peripheral countries and Germany long-term bond yields (Spread10Y.Peripheral) also stands out as an explanatory variable. This metric is often used as a proxy of financial distress in the euro area. Hence, higher spread tends to negatively affect the euro.

In the pre-crisis, peripheral long-term yields (Peripheral.10Y) played a major role for the currency. Special attention goes to the (*à priori* unexpected) negative linkage between the yields and the euro. More than a puzzle, this might be a consequence of the euro area construction. The introduction of the euro in January 1999, prompted an interest rate convergence of the bond yields of countries within the euro area (See Figure 12). Specifically, interest rates in the peripheral economies drop to levels closer to the ones observed in the core countries. This lower interest rate environment prompted a credit-finance boom in these economies and the region became particularly attractive for investments. Hence, the inverse relationship between yields and the euro could hold in this period.

Risk-free rates The OIS curve also seems to play an important role for the euro, specially middle-term tenors. Shorter maturities (OIS.6M), in particular, gained relatively importance during the crisis period. In times of low inflation, 'low-for-long' short term interest rates and the adoption of unconventional monetary policies in major economies, an increase in middle to longer-term yields stand out as a positive influence for the currency.

Figure 12: Euro area Sovereign Bond 10Y yields



Source: Datastream; Author's indication.

Macro-financial indicators this group is also influential. Starting with oil prices, we note the positive EUR/USD - WTI linkage during the first 2 periods, which might essentially reflect the known strong inverse relationship between oil prices and the US dollar. Historically, such relationship holds given two main factors. As crude oil prices are quoted in USD, a weaker USD implies oil prices were higher in US dollar terms. Moreover, the United States were during a long time a net oil importer implying that an increase in oil prices affected negatively its trade balance, putting the USD under pressure. However, the shale revolution has substantially changed this configuration. From 2011, the US became net oil exporter of refined petroleum products, implying that higher oil prices no longer contribute to a further deterioration of the US trade balance. It has actually helped the US to improve its trade balance deficits. Hence, the historically strong inverse relationship between oil prices and the US dollar has become more unstable. Accordingly, in the post-crisis period, the WTI no longer figures as a major variable selected to explain EUR/USD variations.

Higher equity returns (STOXX50) are normally associated with a negative returns for the euro, as a stronger currency means lower international competitiveness for European companies, which make up a big part of the index. Note, however, that a positive relationship was captured during the crisis period, which could be associated with the nature of the risk-off periods itself. Under such situations, safe havens assets (USD, US treasuries, Japanese Yen) normally over-perform. Hence, a decline of the financial stress lead, for instance, by improvement in markets sentiment and positive growth news, might prompt investors to undo conservative portfolio positions and look for diversification.

Macroeconomic surprises and policy also play a major role for the euro. While unanticipated negative news on the macro front in the US (US.Surprise) positively affects the euro, better than expected positive news regarding the macro indicators in the euro area (Euro.Area.Surpise) drives the currency up. ECB's non-standard monetary policy measures (ECB.NSM) negatively impacts the currency. All in all, the standard macro-financial model for the euro presents the expected signs.

ECB's communication signals Regarding our time-series based on the ECB's communication, we observe that topics and sentiments that exert an influence on the euro also evolve over time. During the pre-crisis period, the topic related to the development of the euro area (Development (15)) played a positive and substantial role for the euro. This provides evidence that at its inception, the new currency created (the euro came into circulation in 2002) depended first and foremost on the success of the creation of the euro area to exist. Topics related to monetary policy strategy (Monetary Policy (72)) and incrementation of financial markets and bank union (Europe/Fin.Mkt/CMU (56)) were also significant for the currency. The latter influenced the euro's return negatively, which should reflect the timid political action on this front during this phase, as concrete advances related to the banking union began to be observed only from 2012 onward.

During the crisis period, themes related to growth (Growth (27)), fiscal policy (Fiscal Policy (49)) and the need for low rates environment (Low/rates/inflation (8)) positively affected the currency. This emphasizes the particularity of the period, in which expansionary monetary and fiscal policies were extremely necessary to support economies.

Looking at the most recent period, we note that a greater number of topics were selected as explicative variables for the euro. This could reflect the ascension of communication as a key monetary policy instrument in the past decade, under the unconventional monetary policies environment. In particular, themes related to debt and crisis management (Debt/crisis (74)) among member states have a significant negative influence on the euro, highlighting worrying of fiscal situation in the region given to the high indebtedness of many countries. Lower bank profitability under the low/negative policy rate environment (NLPS/Profitability (28)) is also major source of concern, weighing on the currency. Subjects related to the euro area expansion (topic 51) have also exert pressure on the currency under this period, possibly reflecting markets perception on the need of a more stable (both economic and political) environment within region before taking further steps to enlargement. Improvements in financing conditions and credit (Bank/loan/credit (77)), in turn, positively affect the currency. It is interesting to note the appearance of the sentiment variable '*Superfluous*' which is negatively related to the euro. In times on which central bank communication has been at the "*heart of monetary policy*" central banks need to be attentive in delivering precise and clear policy messages.

4.6 Discussion

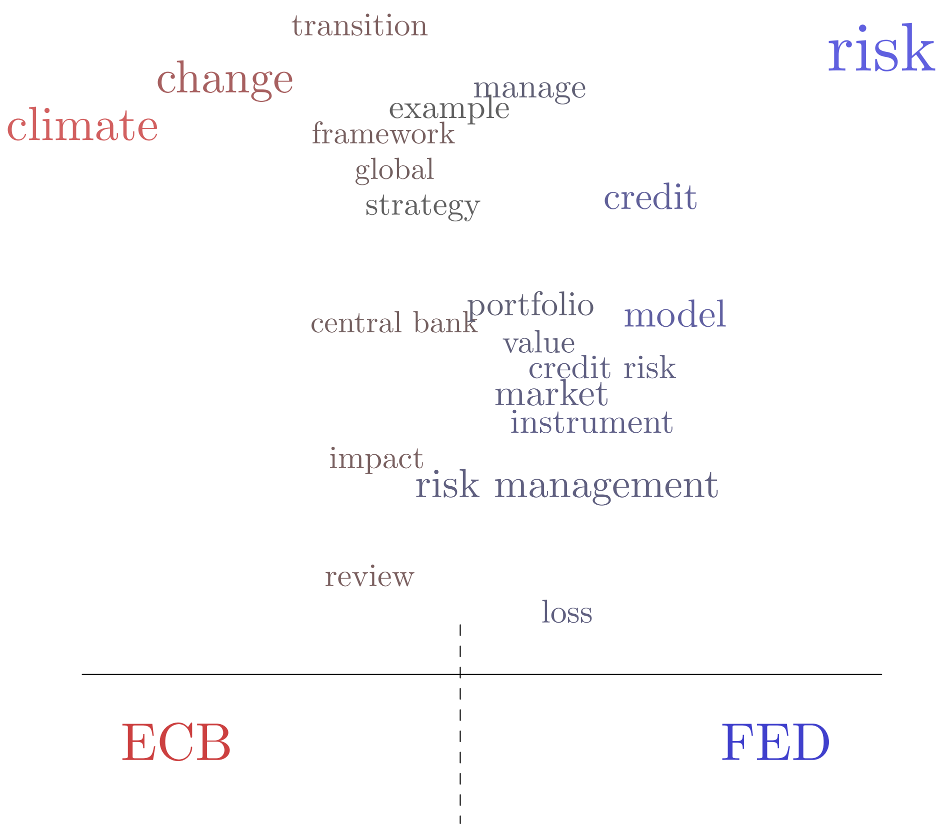
Results show the added-value of using Natural Language Processing (NLP) for assessing central bank communication and construct relevant signals. The unsupervised and parsimonious methods applied allowed us to highlight the diversity and temporal dispersion of the subjects addressed by the ECB – disentangling the multiple strands of its communication.

Through these set of signals, we measured the relative importance of ECB's talk on the euro performance. We note that the text mining based series increase the explanatory power of the euro/US dollar models by more than 2% on a market as complex as the Forex. Importantly, the topic signals prevalence is systematically selected over other financial variables by the Lasso method implemented, stressing the importance of the ECB's communication as driver for the euro's return. Potentially, a refinement of the model by the inclusion of additional information about the dataset (i.e. covariates), an option provided by the Structural Topic Model (STM), could generate even more promising results.

Indeed, this is a major innovation of the STM in comparison to other probabilistic topic models (see Appendix A.4.1). The generative process of the STM integrates contextual information in the process of topic construction. This allows to highlight perspectives of different agents within a group and/or of different entities on a particular subject. To illustrate the latter, we enlarged our dataset to also comprise Federal Reserve Bank[47] (Fed) communication. Figure 13 outlines the main 'wording' used by the ECB and the Fed related to the topic *Financial Stability Risk*. This Figure suggests that the Fed still focuses on a more traditional risk dimension, while the ECB's discourse has shifted to include climate change and transition as major risks for financial stability. This finding corraborates the idea of "*transatlantic divide*" introduced by Drei *et al.* (2019). However, we should note

[47]We used partial content scrapped dataset for public speeches given by Fed governors. The data is provided by Natan Mish and available on Kaggle at https://www.kaggle.com/natanm/federal-reserve-governors-speeches-1996-2020/version/2.

Figure 13: Financial Stability Risk



Source: Structural Topic Model (STM); Author's calculation. This Figure highlights how the vocabulary used to talk about a particular topic (in this case, *Financial Stability*) differs between the ECB and the Fed.

that the Fed is catching up on this matter and has recently put potential risks arising from climate change on the table[48]. This analytical tool has several applications. For example,

[48]During the Fed Press Conference on November 05, 2020, chairman Jerome Powell pointed out that the institution is taking important setps toward integrating climate risks into their analysis "*incorporating climate change into our thinking about financial regulation is relatively new, as you know. And we are very actively in the early stages of this, getting up to speed, working with our central bank colleagues and other colleagues around the world to try to think about how this can be part of our framework.*"Also note that

under the current take-off in ESG investing, analysts could use such process to encompass cultural disparities within discourse.

Overall, our findings suggest that topic modelling could be useful in ex-post expert analysis as it enlarges the camp of information available, having value for both central banks and market practitioners. The former, can have a better assessment of their discourse. For instance, if they detect a negative market response to a given topic, they might rethink their tone or approach. As for asset owners and investment managers, such a tool can rapidly process information of several central banks communication if data is properly integrated. In practice, this approach can be used to better understand risks and opportunities in their investment universe.

That been said, our methodology is subject to some limitations:

(i) Looking at daily variations on speech days only reduce significantly both number of observations and the possibility for the market to react over several days/weeks to some announcements,

(ii) Both sentiment and topic series were constructed in a simple manner (and could be improved – e.g. including covariates),

(iii) Our approach considers fixed content which can be made dynamic (as in Lima *et al.* (2020)) and compatible with forecasting,

(iv) Variances of financial variables are generally time-varying. Splitting the period helps limiting these effects, but modelers often use GARCH/ARCH, i.e. volatility clustering models to better capture the relationships. Some adaptions can be made to the Lasso to be compatible with such modeling (Medeiros & Mendes, 2017).

(v) Finally, we did not test for the value added of this database in a predictive process.

We justified our choices along the methodology and in the Appendix. We emphasize that our objective was to use text mining to provide a new reading of central bank's communication, and to develop a prove of concept tool, assisting analysts in their task when it comes to dissect the semantic content of central banks' discourse. Therefore, if the limitations above applies, they all refer to existing techniques than could be integrated straightforwardly.

in the November 2020 Financial Stability Report, the Fed adressed the implications of climate change for financial stability for the first time (System, 2020).

Table 2: Sub-periods models based on Lasso selection (Figures 31 and 32)

	Dependent variable:					
	EUR/USD					
	01/01/2002 to 30/07/2007		01/08/2007 to 30/09/2012		01/10/2012 to 31/07/2020	
Peripheral.10Y	-0.058^{***}	-0.058^{***}				
	(0.008)	(0.008)				
Spread10Y.Peripheral			-0.017^{***}	-0.017^{***}		
			(0.004)	(0.003)		
Spread2Y.US			-0.019^{***}	-0.021^{***}	-0.033^{***}	-0.034^{***}
			(0.004)	(0.004)	(0.006)	(0.006)
OIS.6M			0.020^{**}	0.019^{**}		
			(0.009)	(0.009)		
OIS.1Y					0.061^{***}	0.062^{***}
					(0.019)	(0.019)
OIS.5Y	0.042^{***}	0.038^{***}			0.025^{***}	0.026^{***}
	(0.011)	(0.011)			(0.008)	(0.008)
WTI	0.027^{**}	0.029^{***}	0.053^{***}	0.054^{***}		
	(0.011)	(0.011)	(0.011)	(0.011)		
STOXX50	-0.064^{***}	-0.066^{***}	0.101^{***}	0.100^{***}	-0.049^{***}	-0.049^{***}
	(0.019)	(0.019)	(0.019)	(0.018)	(0.014)	(0.013)
US.Surprise			-0.012^{***}	-0.012^{***}	-0.010^{***}	-0.011^{***}
			(0.004)	(0.004)	(0.003)	(0.003)
Euro.area.Surprise					0.006^{**}	0.007^{**}
					(0.003)	(0.003)
ECB.NSM					-0.696^{***}	-0.669^{***}
					(0.137)	(0.136)
'Mon.Pol./Rate/Inf./FG (72)'		-0.056^{**}				
		(0.023)				
'Development (15)'		0.402^{**}				
		(0.190)				
'Europe/Fin.Mkt/CMU (56)'		-0.047^{**}				
		(0.023)				
'Growth (27)'				0.082^{***}		
				(0.027)		
'Low/rates/inflation (8)'				0.065^{**}		
				(0.026)		
'Fiscal policy (49)'				0.062^{**}		
				(0.026)		
'Debt/crisis (74)'						-0.056^{***}
						(0.016)
'NLPs/Profitability (28)'						-0.039^{**}
						(0.016)
'Bank/loan/credit (77)'						0.039^{**}
						(0.016)
'Euro area Expansion (51)'						-0.038^{**}
						(0.016)
'Superfluous'						-0.037^{**}
						(0.016)
Constant	0.010	-0.002	0.007	0.007	0.028^{*}	0.028^{*}
	(0.023)	(0.024)	(0.027)	(0.026)	(0.016)	(0.016)
Observations	436	436	495	495	830	830
R^2	0.206	0.233	0.331	0.359	0.160	0.193
Adjusted R^2	0.199	0.221	0.323	0.347	0.152	0.181
Residual Std. Error	0.476 (df = 431)	0.470 (df = 428)	0.590 (df = 488)	0.580 (df = 485)	0.458 (df = 822)	0.450 (df = 817)
F Statistic	27.994^{***} (df = 4; 431)	18.616^{***} (df = 7; 428)	40.323^{***} (df = 6; 488)	30.197^{***} (df = 9; 485)	22.285^{***} (df = 7; 822)	16.233^{***} (df = 12; 817)
Note:						$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Source: Author's calculations.

5 Conclusion

This article narrates the history of the European Central Bank through its communication using quantitative tools. We used text mining techniques for this purpose, valuing the unstructured data available and extracting information hidden in corpus of texts, revealing patterns and trends. This procedure allowed us to identify the main *topics* of interest to the ECB. We showed that they go beyond the core objective of maintaining price stability, enlightening research focus and main themes of discussion among board members. Furthermore, we built *sentiment* signals compatible with any form of intervention, formal and informal. This is an important step, as emotions are also a distinctive element of discourse, being important in the transmission of a message.

We then used both of these signals, *topics* and *sentiment*, to measure the impact of central bank communication on asset prices, especially on the euro. We analyzed in an "agnostic" way the main signals of influence on the currency, whether or not they are related to monetary policy. Our results show that both signals are significant drivers for the euro's return. They are consistently chosen by the Lasso selection over time and improved the regressions in all sub-periods analyzed (2.5% on average). We also note that more communication variables are significant under the zero lower bound period, which emphasize the greater role of communication in this environment. Consistently with the literature, our sentiment variables suggest that tone matters and that delivering clear messages is key. We showed that meaningful topics change over time and encompass a broad range of subjects: from the development and enlargement of the euro area, to fiscal policies, liquidity risk and banks' profitability. As a whole, this implies that financial markets are sensitive to a large and diverse scope of the ECB's communication, well beyond the traditional thematic of monetary policy. Clearly, there is a central bank communication factor at play in the European financial markets, including an extra-monetary one.

This is particularly important in a context where the ECB is conducting a strategy review and aims at integrating more social-related issues into its scope of actions - certainly, a road of one-way. By trying to understand how the institution through its stewardship (of course within its mandate) can help the society to move forward on some issues such as climate change and/or inequality, the central bank could give market incentives, easing the movement in the intended direction. While this is hot-topic discussion that surpasses the objectives of this article (social-related themes should be addressed only by governments? Should central bank address market failures?), the important message to take is that the ECB has changed, has enhanced communication and intends to be fully understandable and closer to the society. This certainly means that themes that until then were observed as secondary, will become central. Being able to identify such themes may become essential from an active management point of view, for both alpha generation and stock selection. Taking into account that central banks of other major economies have also advanced in the same direction, markets may have to adjust, making this issue to become imperative.

References

Airoldi, E. M., & Bischof, J. M. (2012). *A Poisson convolution model for characterizing topical content with word frequency and exclusivity. arXiv preprint arXiv:1206.4631.*

Airoldi, E. M., & Bischof, J. M. (2016). *Improving and evaluating topic models and other models of text. Journal of the American Statistical Association*, *111*(516), 1381–1403.

Airoldi, E. M., Cohen, W. W., & Fienberg, S. E. (2005). *Bayesian models for frequent terms in text. Proceedings of the Classification Society of North America and INTERFACE Annual Meetings*, *990*, 991.

Aletras, N., & Stevenson, M. (2013). *Evaluating topic coherence using distributional semantics. Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, 13–22.

Altavilla, C., Brugnolini, L., Gürkaynak, R. S., Motto, R., & Ragusa, G. (2019). *Measuring euro area monetary policy. Journal of Monetary Economics*, *108*, 162–179.

Altavilla, C., Carboni, G., & Motto, R. (2015). *Asset purchase programmes and financial markets: lessons from the euro area.*

Armelius, H., Bertsch, C., Hull, I., & Zhang, X. (2020). *Spread the Word: International spillovers from central bank communication. Journal of International Money and Finance*, *103*, 102116.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). *On finding the natural number of topics with latent dirichlet allocation: Some observations. Pacific-Asia conference on knowledge discovery and data mining*, 391–402.

Banque de France International Symposium: Productivity, competitiveness and globalisation. (2005).

Bennani, L., Le Guenedal, T., Lepetit, F., Ly, L., Mortier, V., Roncalli, T., & Sekine, T. (2018). *How ESG Investing Has Impacted the Asset Pricing in the Equity Market.*

Bernanke, B., Reinhart, V., & Sack, B. (2004). *Monetary policy alternatives at the zero bound: An empirical assessment. Brookings papers on economic activity*, *2004*(2), 1–100.

Bernanke, B. S. (2013). *Communication and Monetary Policy. Federal reserve Bank.* https://www.federalreserve.gov/newsevents/speech/bernanke20131119a.htm

Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). *Text mining for central banks. Available at SSRN 2624811.*

Blei, D. M., Lafferty, J. D. et al. (2007). *A correlated topic model of science. The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent dirichlet allocation. Journal of machine Learning research*, *3*(Jan), 993–1022.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., & Jansen, D.-J. (2008). *Central bank communication and monetary policy: A survey of theory and evidence. Journal of Economic Literature*, *46*(4), 910–45.

Brand, C., Buncic, D., & Turunen, J. (2010). *The impact of ECB monetary policy decisions and communication on the yield curve. Journal of the European Economic Association*, *8*(6), 1266–1298.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). *A density-based method for adaptive LDA model selection. Neurocomputing*, *72*(7-9), 1775–1781.

Clark, P. B., & MacDonald, R. (1999). *Exchange rates and economic fundamentals: a methodological comparison of BEERs and FEERs. Equilibrium exchange rates* (pp. 285–322). Springer.

Cline, W. R., Williamson, J. et al. (2008). *New estimates of fundamental equilibrium exchange rates*. Peterson Institute for International Economics.

Coenen, G., Ehrmann, M., Gaballo, G., Hoffmann, P., Nakov, A., Nardelli, S., Persson, E., & Strasser, G. (2017). *Communication of monetary policy in unconventional times.*

Connolly, E., Kohler, M. et al. (2004). News And Interest Rate Expectations: A Study Of Six Central Banks. Citeseer.

De Haan, J. (2008). *The effect of ECB communication on interest rates: An assessment. The Review of International Organizations*, *3*(4), 375–398.

De Santis, R. A., & Stein, M. (2016). *Correlation changes between the risk-free rate and sovereign yields of euro area countries.*

Deveaud, R., SanJuan, E., & Bellot, P. (2014). *Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique*, *17*(1), 61–84.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.*

Draghi, M. (2014a). *Central Bank communication. European Central Bank.* https://www.ecb.europa.eu/press/inter/date/2014/html/sp140804.en.html

Draghi, M. (2014b). *Monetary Policy communication in turbulent times. European Central Bank.* https://www.ecb.europa.eu/press/key/date/2014/html/sp140424.en.html

Drei, A., Le Guenedal, T., Lepetit, F., Ly, L., Mortier, V., Roncalli, T., & Sekine, T. (2019). *ESG Investing in Recent Years: New Insights from Old Challenges.* Amundi Discussion Paper, 42, www. research-center. amundi. com.

Ehrmann, M., & Fratzscher, M. (2003). *Monetary policy announcements and money markets: A transatlantic perspective. International Finance*, *6*(3), 309–328.

Ehrmann, M., & Fratzscher, M. (2007). *Explaining monetary policy in press conferences.*

Ehrmann, M., & Talmi, J. (2020). *Starting from a blank page? Semantic similarity in central bank communication and market volatility. Journal of Monetary Economics*, *111*, 48–62.

Ehrmann, M., & Wabitsch, A. (2020). *Central Bank Communication with Non-Experts–A Road to Nowhere?*

Ettinger, A. (2020). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics*, *8*, 34–48.

Feldman, R. (2013). *Techniques and applications for sentiment analysis. Communications of the ACM*, *56*(4), 82–89.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). *Regularization paths for generalized linear models via coordinate descent. Journal of statistical software*, *33*(1), 1.

Griffiths, T. L., & Steyvers, M. (2004). *Finding scientific topics. Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

GRIMMER, J., & STEWART, B. M. (2013). *Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis*, *21*(3), 267–297.

GÜRKAYNAK, R. S. (2005). *Using federal funds futures contracts for monetary policy analysis.*

GÜRKAYNAK, R. S., SACK, B. P., & SWANSON, E. T. (2004). *Do actions speak louder than words? The response of asset prices to monetary policy actions and statements.*

HANSEN, S., & MCMAHON, M. (2016). *Shocking language: Understanding the macroeconomic effects of central bank communication. Journal of International Economics*, *99*, S114–S133.

HANSEN, S., MCMAHON, M., & PRAT, A. (2018). *Transparency and deliberation within the FOMC: a computational linguistics approach. The Quarterly Journal of Economics*, *133*(2), 801–870.

HU, M., & LIU, B. (2004). *Mining and Summarizing Customer Reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177. https://doi.org/10.1145/1014052.1014073

HUBERT, P., LABONDANCE, F. et al. (2018). *The effect of ECB forward guidance on the term structure of interest rates. International Journal of Central Banking*, *14*(5), 193–222.

JANSEN, D.-J. (2011). *Does the Clarity of Central Bank Communication Affect Volatility in Financial Markets? Evidence from Humphrey-Hawkins Testimonies. Contemporary Economic Policy*, *29*(4), 494–509.

LAGARDE, C. (2020, September 30). *The monetary policy strategy review: Some preliminary considerations*. Retrieved November 3, 2020, from https://www.ecb.europa.eu/press/key/date/2020/html/ecb.sp200930~169abb1202.en.html

LAMFALUSSY, A. (1994). *Central banking in transition. The Future of Central Banking: The Tercentenary Symposium of the Bank of England*, 330.

LANE, P. (2020 (accessed October 9, 2020)). *The ECB's monetary policy response to the pandemic: liquidity, stabilisation and supporting the recovery*. https://www.ecb.europa.eu/press/key/date/2020/html/ecb.sp200624~d102335222.en.html

LEOMBRONI, M., VEDOLIN, A., VENTER, G., & WHELAN, P. (2019). *Central bank communication and the yield curve. Available at SSRN 2873091.*

LIMA, L. R., GODEIRO, L. L., & MOHSIN, M. (2020). *Time-Varying Dictionary and the Predictive Power of FED Minutes*. *Computational Economics*, 1–33.

LIN, C., & HE, Y. (2009). *Joint sentiment/topic model for sentiment analysis*. *Proceedings of the 18th ACM conference on Information and knowledge management*, 375–384.

LIU, B. (2012). *Sentiment analysis and opinion mining*. *Synthesis lectures on human language technologies*, *5*(1), 1–167.

LIU, B., & ZHANG, L. (2012). *A survey of opinion mining and sentiment analysis*. *Mining text data* (pp. 415–463). Springer.

LOUGHRAN, T., & MCDONALD, B. (2011). *When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks*. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

LUCCA, D. O., & TREBBI, F. (2009). *Measuring central bank communication: an automated approach with application to FOMC statements*.

MACKLEM, T. (2020, August 27). *The imperative for public engagement*. Retrieved November 5, 2020, from https://www.bankofcanada.ca/2020/08/imperative-for-public-engagement/

MCAULIFFE, J. D., & BLEI, D. M. (2008). *Supervised topic models*, 121–128.

MEDEIROS, M. C., & MENDES, E. F. (2017). *Adaptive LASSO estimation for ARDL models with GARCH innovations*. *Econometric Reviews*, *36*(6-9), 622–637.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., & DEAN, J. (2013). *Distributed representations of words and phrases and their compositionality*. *Advances in neural information processing systems*, 3111–3119.

MIMNO, D., & LEE, M. (2014). *Low-dimensional embeddings for interpretable anchor-based topic inference*. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1319–1328.

MIMNO, D., WALLACH, H. M., TALLEY, E., LEENDERS, M., & MCCALLUM, A. (2011). *Optimizing semantic coherence in topic models*. *Proceedings of the conference on empirical methods in natural language processing*, 262–272.

MIMNO, D. M., & MCCALLUM, A. (2008). *Topic models conditioned on arbitrary features with Dirichlet-multinomial regression*. *UAI*, *24*, 411–418.

Mohammad, S. M., & Turney, P. D. (2013). *Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Moschella, M., Pinto, L., & Martocchia Diodati, N. (2020). *Let's speak more? How the ECB responds to public contestation. Journal of European Public Policy*, *27*(3), 400–418.

Nadal-De Simone, M. F., & Razzak, W. (1999). Nominal Exchange Rates And Nominal Interest Rate Differentials. International Monetary Fund.

Naldi, M. (2019). *A review of sentiment computation methods with R packages. arXiv preprint arXiv:1901.08319*.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). *Automatic evaluation of topic coherence. Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 100–108.

Nielsen, F. Å. (2011). *A new ANEW: evaluation of a word list for sentiment analysis in microblogs*. In M. Rowe, M. Stankovic, A.-S. Dadzie, & M. Hardey (Eds.), *Proceedings of the eswc2011 workshop on 'making sense of microposts': Big things come in small packages* (pp. 93–98). http://ceur-ws.org/Vol-718/paper_16.pdf

Nikita, M. (2016). *Package 'ldatuning'*.

Picault, M., & Renault, T. (2017). *Words are not all created equal: A new measure of ECB communication. Journal of International Money and Finance*, *79*, 136–156.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). *Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora*, 248–256.

Ratcliffe, S. (2017). Oxford Essential Quotations. Oxford University Press.

Roberts, M., Stewart, B., Tingley, D., Benoit, K., Stewart, M. B., Rcpp, L., Imports matrixStats, R., & KernSmooth, N. (2020). *Package 'stm'. Imports matrixStats, R. & KernSmooth*, (2017).

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). *A model of text for experimentation in the social sciences. Journal of the American Statistical Association*, *111*(515), 988–1003.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). *Stm: An R package for structural topic models. Journal of Statistical Software*, *91*(2).

Röder, M., Both, A., & Hinneburg, A. (2015). *Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Rosa, C., & Verga, G. (2005). *The Importance of the Wording of the ECB. CEP Discussion Paper*, (0694).

Rossi, B. (2013). *Exchange rate predictability. Journal of economic literature*, *51*(4), 1063–1119.

Schmeling, M., & Wagner, C. (2019). *Does central bank tone move asset prices? Available at SSRN 2629978.*

Sutherland, C. S. (2020). *Forward Guidance and Expectation Formation: A Narrative Approach.*

System, F. R. (2020). *Financial Stability Report*. Federal Reserve System.

Taddy, M. (2012). *On estimation and selection for topic models. Artificial Intelligence and Statistics*, 1184–1193.

Trichet, J.-C. (2008b). *Central banks and the public - the importance of communication. European Central Bank.* https://www.ecb.europa.eu/press/key/date/2008/html/sp081118_2.en.html

Trichet, J.-C. (2008c). *The changing role of communication. European Central Bank.* https://www.ecb.europa.eu/press/key/date/2008/html/sp081215_2.en.html

Trichet, J.-C. (2008a). *A few remarks on communication by central banks. European Central Bank.* https://www.ecb.europa.eu/press/key/date/2008/html/sp080116_1.en.html

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). *Evaluation methods for topic models. Proceedings of the 26th annual international conference on machine learning*, 1105–1112.

Weidmann, J. (2018). *Central Bank communication as an instrument of monetary policy. Bank of International Settlements.* https://www.bis.org/review/r180511a.htm

Wesslen, R. (2018). *Computer-assisted text analysis for social science: Topic models and beyond. arXiv preprint arXiv:1803.11045.*

Yellen, J. L. (2012). *Revolution and Evolution in Central Bank Communications. Federal Reserve Bank*. https://www.federalreserve.gov/newsevents/speech/yellen20121113a.pdf

Zhu, J., Ahmed, A., & Xing, E. P. (2012). *MedLDA: maximum margin supervised topic models. Journal of Machine Learning Research*, *13*(Aug), 2237–2278.

A Text mining methods

In this section, we introduce the main text mining principles used in our analysis. The objective is to provide a review of the parsimonious techniques and algorithms that can be used on any text database. Therefore, from a sample extracted in the speech database provided by the European Central Bank[49], we explain how to extract information about the words used, their frequency and relation with surrounding words. Then, we present some examples of sentiment analysis and we provide illustrative examples to calibrate contextual preprocessing, automatically or manually, which is recommended for a process of better quality.

In order to provide a comprehensive pedagogical review of the main principles of text mining, we restricted the corpus of texts to few Mario Draghi's speeches in this section. Our objetive is to introduce the first notions of text mining and data science tools '*naively*' i.e. without using any knowledge of economy or finance.

[49]See European Central Bank. (25 October 2019). Speeches database. Retrieved from: https://www.ecb.europa.eu/press/key/html/downloads.en.html

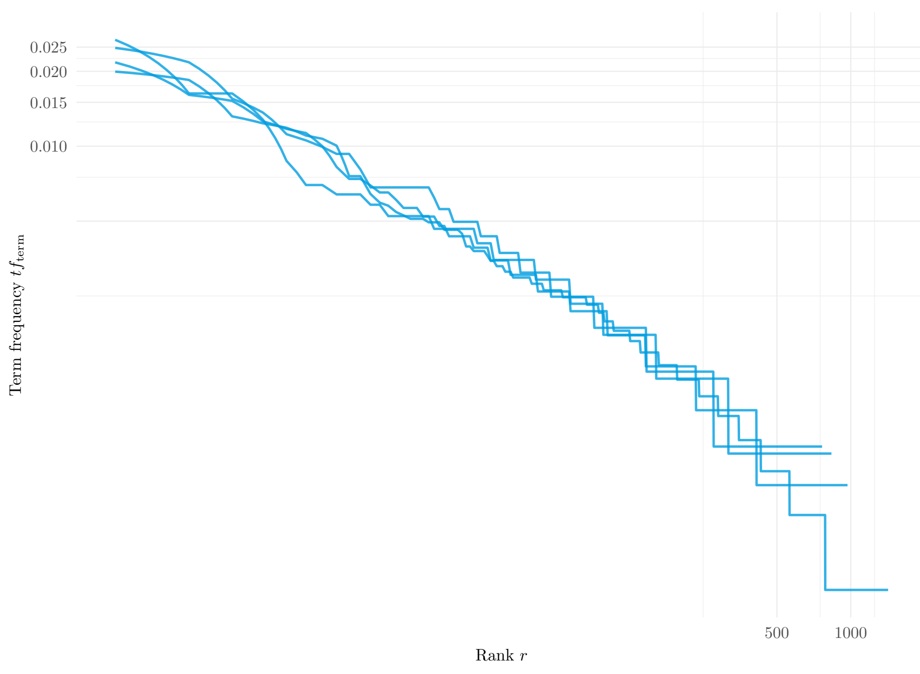
A.1 Basic statistics

Data analysis in text mining starts normally with a simple word count. The referring metric is called term-frequency:

$$tf(w,u) = \frac{\text{Number of times word } w \text{ appears in document } u}{\text{Total number of word in the document } u}$$

This statistic is nothing else than the number of times we observe a given word divided by the number of words (n) in the corpus of texts (e.g. speeches, interviews, etc.). The rank of the word r evolve between 1 and n from the most to less used word. For instance, the word '*euro*' in the IMFC Statement speeches appears 121 times on a total of 6071 words, its frequency is, hence, $tf_{\text{euro}} = 0.0199$ and its rank is 1. It is possible to rank each word with regard to the referring speech or corpus. Figure 14 shows the Zipf's law, that translates the idea that the frequency of a term is inversely proportional to its rank. We note that it is almost identical for every speech (of comparable length). These distributions are typical of the language and are observable in any context.

Figure 14: Draghi's intervention examples: Zipf's Law



Inverse document frequency (idf) is a complementary numerical statistic reflecting the importance of a word to characterize a document among the corpus of texts. The tf–idf increases with the frequency but is offset by the number of documents in the corpus that also contain a given word. It is defined as follow:

$$idf(w,\mathcal{C}) = \log\left(\frac{\text{Total number of documents in the corpus } \mathcal{C}}{\text{Number of documents in } \mathcal{C} \text{ containing the word } w}\right)$$

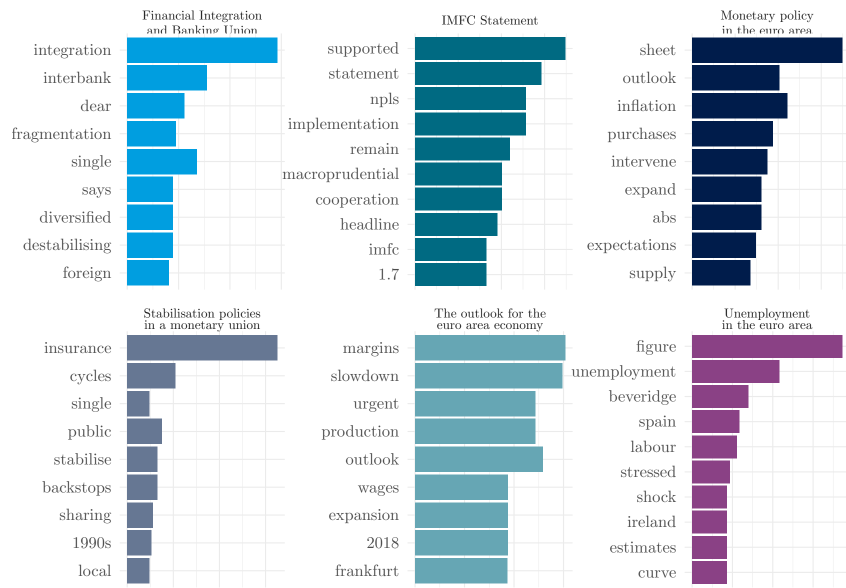
This procedure allow us to adjust the frequency of the words that appears in every document. This metric is interesting before implementing topic analysis. For example, if the world '*euro*' has a term frequency 0.0199, its *idf* in this corpus is null because this world is present in every document.

The $tf.idf$ is the metric defined for each word in every document as the product of the two:

$$tf.idf(w,d,\mathcal{C}) = tf(w,d) \times idf(w,\mathcal{C})$$

Figure 15 shows the tf-idf in a subset of six Mario Draghi's speeches. This metric allows to extract the information that are characteristic of every speech.

Figure 15: Draghi's interventions example: $tf.idf$



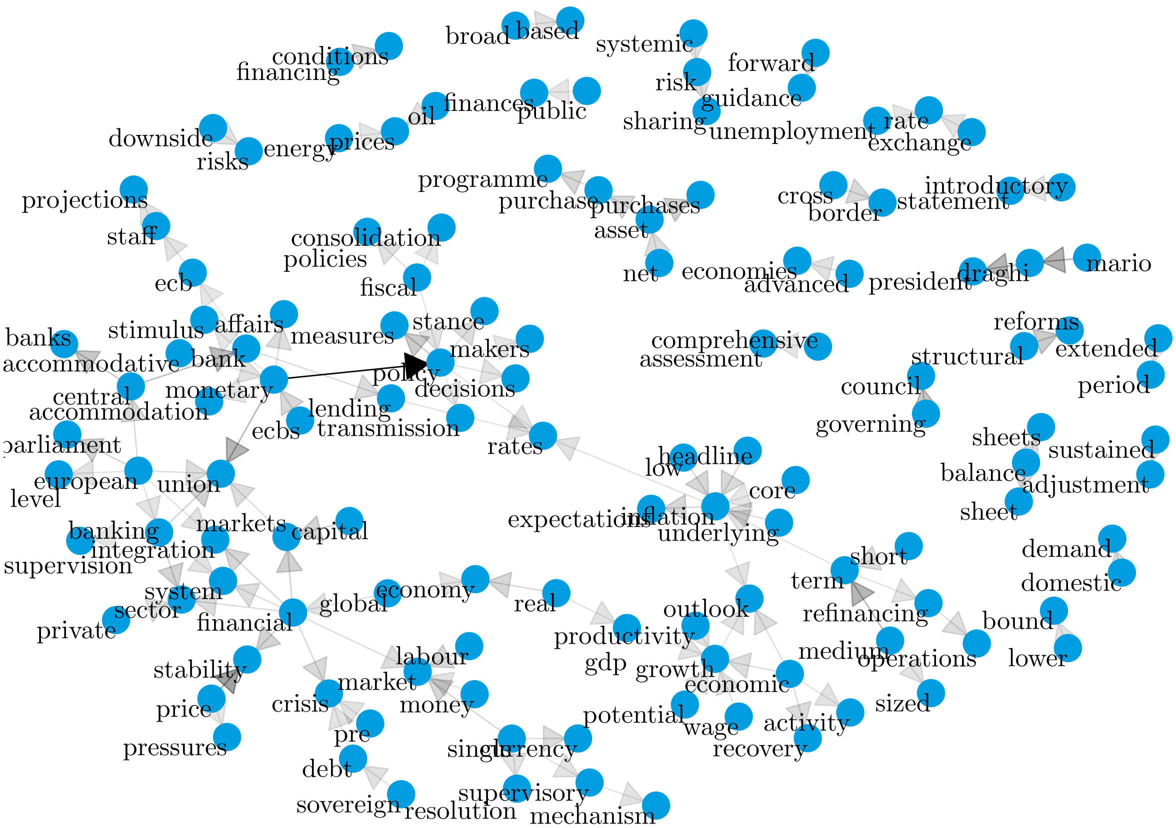
A.2 Exploration and contextual treatment

Combined words and common patterns: n-grams In text analysis, we can also observe and extract relationship between words. The common sequences of words are called n-grams. Figure 16 is a graphic interpretation of the most encountered words-relationships in Mario Draghi's speech database. We can note, for instance, that one of the main linkages is 'monetary policy', but also that 'comprehensive' is generally followed by 'assessment', 'cross' by 'border', 'downside' by 'risk'. In quantitative processes, n-grams are often retrieved automatically, using a frequency threshold and a brief qualitative check-up. In our paper, we chose to establish the list manually, an important procedure that was done in the preprocessing.

Text preprocessing This step is of prior importance. The preprocessing can be defined as a series of command transforming the original data (e.g.)speech) into an character easily convertible in informative time-series. These operations depend on the desired output. In general, it is common to remove punctuation, numbers[50] and '*stop-words*'. Stop-words are

[50]Except if is the idea is to extract numeric information from a report, for example.

Figure 16: Draghi's interventions n-grams



words very often found in texts and that have no informative value: 'the', 'a', 'that', 'of', 'in', 'but', 'also', 'by', etc. Note that the word 'not' is also an stop-word. Hence, before doing stop-worlds removal, one might establish the existance of 'not-words' otherwise it would be removed, and we would lose the information about the negation and hence, changing the context of the document analysed[51].

The second important step of the preprocessing is to integrate the contextual language common patterns or n-grams (generally we collect bigrams and trigrams). The desired output of this treatment is not for our learning algorithm to determine that the subject is indeed 'monetary policy' (without splitting the two words). We wish to extract information such as is the speaker is in favor of a strengthening or an ease of the policies? Optimally, we aim at assessing the speakers' confidence and opinion about future policies. Therefore, common language patterns must be treated ex-ante, or will bias our analyses. The operation is rather simple, and consists into sticking words together. Following the example given in the n-grammas above, 'monetary policy' becomes 'monetary-policy', 'cross border' becomes 'cross-border' and 'downside risk' becomes 'downside-risk'. After the preprocessing, we generally obtain smoother frequency distributions. Note that the recent topic and sentiment

[51]For instance, "not happy" would be defined as one world and become "not-happy".This procedure is important to keep the negation of the context. Once this is done, one can procedure with the removal of stop words.

algorithm proposed treat reasonably well these operations, but we consider results are cleaner, more readbale and of superior quality when supervised.

To obtain smoother words' distribution, it is required to lemmatize or stem the text. Lemmatization automatically groups words such as 'produce', 'production', 'product', under a common word: 'produce'. The stem operation provides the common root only, 'produc-'. Some of these operations have to be controlled in a '*non-naive*' process, not to affect specific designations (for instance 'global domestic produc-'). Other operations can be added, including sparsity treatment, translation of non-English words, etc. We opted to use lemmatization of the words as it gives better interpretability.

A.3 Sentiment analysis

The value-added of a text mining process lies on its ability to quantitatively assess speakers opinions and sentiments. This translates into numerous questions: How to detect the positiveness of a speech? Extract clues reflecting the confidence of the speaker or spot sarcasm in an interview? Is there a quantitative way to distinguish and classify speeches based on a text mining analysis? Answering all these questions require a advanced sentiment analysis model.

This type of model exists in the industry, and an increasing number of data providers now propose a set of metrics capturing sentiments. Naturally, their internal processes are confidential and therefore produce metrics hardly reproducible. On the academic side, we can find some example of initiative exploring the subject. Liu (2012) and Liu and Zhang (2012) present all the fundamental concepts of text mining and associated literature, including sentiment analysis. Some overviews of more advanced machine learning methods can be found in Feldman (2013). These concepts are sometimes associated with topic modeling (Lin & He, 2009). Finally, comprehensive implementation guides[52] (Naldi, 2019) are also made available to ease the development of internal processes. In this paper, we use solely open-source database lexicons and simple process. More advanced modeling techniques is not necessarily pertinent in our case, as central banks tend to be factual, neutral anf may (normally) use sarcasm during interventions.

A common method to classify speeches is to compare the words used to a preliminary classified dictionary or lexicon. For example, we can create small dictionary for negative locutions. There are of two main types of negative words: the '*not-words*[53]', that reverse the sense of the following word, and '*negatively charged word*[54]', that indicate that the speaker is talking about something undesirable. The words in the second list can retrieve a value score, for instance: 'banal' gets -1 or 0 while 'atrocious' and 'awful' -3 (AFINN score). Filtering the speeches the words connected to items in the '*not-words*' list is crucial to avoid misinterpreting the speech.

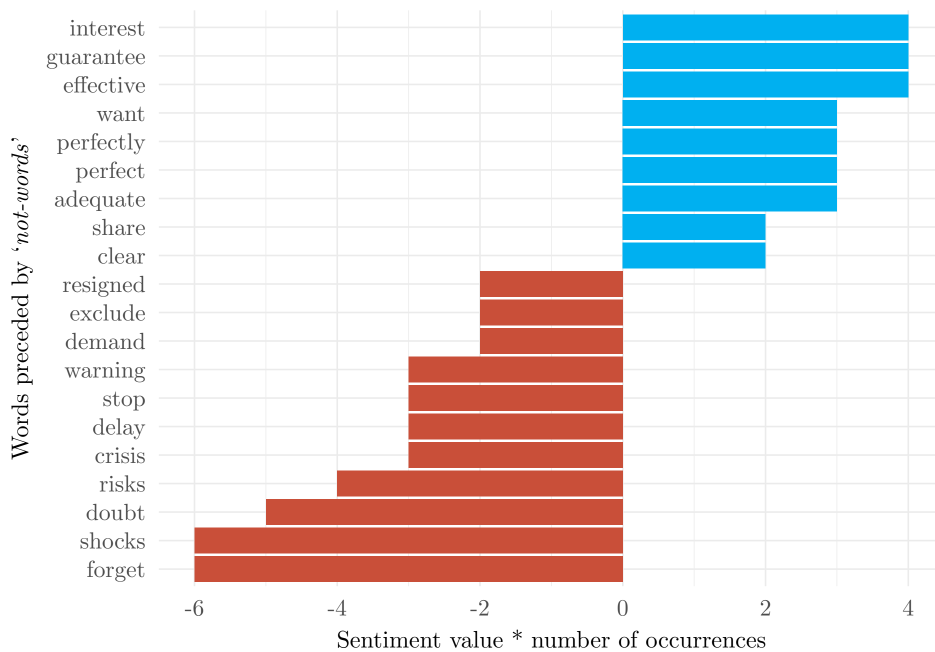
[52] Most sentiment detection functions are designed for social networks and informal speeches. Processes using punctuation detection or emoticons are less suitable when the use of formal language is the norm, such as central bank speeches and statements.

[53] Not-words list: 'not', 'no', 'neither', 'nor', 'negative', 'under', 'never', 'without', etc.

[54] Negative words: 'alarming', 'annoy', 'anxious', 'apathy','appalling', 'atrocious', 'awful','bad','banal','barbed', 'beneath','boring', 'broken','collapse', 'confused', 'contradictory', 'contrary', 'corrosive', 'corrupt','crazy', 'criminal', 'cruel','cry', 'cutting', 'cut', 'deformed', 'deny','deplorable', etc.

The lexicon can be complemented by positive words, to build an affinity index. In Mario Draghi's speeches, we filtered the affinity words. Figure 17 shows the words that are preceded by a negation and their affinity value provided by external open-source. This picture provides an indication of the preprocessing that will be required in the next step of our analysis. Indeed, we can see that some words positively charged such as 'perfect' or 'clear' are very often encountered preceded by a negation: 'not-perfect' and 'not-clear'. Inversely, some negative words are mitigated by not-words as well 'warning' and 'resigned' were in fact, 'no-warning' and 'not-resigned'. Therefore the sentiment value can be reverted or ignored (i.e. set to 0). Other step can be added to extract properly information from text (detect sarcasm or contrasted opinions).

Figure 17: Draghi's interventions: Negation and Affinity Scores



Sentiment lexicons time-series To simply attribute a score to the overall sentiment carried by a speech, we introduce lexicons and sum-up the scores of the words within the speech. In such a specific context, the general lexicons should be treated to avoid dedicated financial terms, such as *interest*, to be influential, as, it generally carries a sentiment, in this case, *a priori* positive (in AFINN only[55]). The generation of control series based on the sentiment derived from these lexicons could add value in the process. We use multiple sentiment indicators which increase the robustness of our analysis. In addition, we employ both, general and financial dictionaries, allowing to capture different language usage which may vary according to the communication tool employed.

The first general purpose lexicon we introduced is the affinity (AFINN) table contains contains a list of 2477 words scored from -5 to +5 (Nielsen, 2011). For each speech, the

[55]For that reason we had to remove this word from the list.

AFINN score is defined as follow:

$$\mathcal{S}_{\text{Affin}}(t) = \sum_{k=1}^{N} \frac{\text{AFINN}(\text{word} \in \text{speech})(t)}{N(\text{word} \in \text{speech})(t)}$$

where N is the total number of scored words in the speech given at time t. The BING list counts for 6786 words that are simply sorted as positive or negative. The score of each word is binary (+1 or -1). Similarly, the BING (Hu & Liu, 2004) score is the defined as the excess of either positive or negative locution in the speech:

$$\mathcal{S}_{\text{BING}}(t) = \sum_{k=1}^{N} \frac{\text{BING}(\text{word} \in \text{speech})(t)}{N(\text{word} \in \text{speech})(t)}$$

Finally, we introduce the crowd-sourced word-emotion NRC lexicon (Mohammad & Turney, 2013). In this lexicon words are associated to an emotion. It regroups 13 901 connections, for instance, abandon is connected to fear and more generally to a negative sentiment. The NRC signal is a 10-dimensional vector constructed following the similar process:

$$\mathcal{S}_{\text{NRC}}(t) = (\mathcal{S}_{\text{positive}}(t), \mathcal{S}_{\text{negative}}(t), \mathcal{S}_{\text{fear}}(t), \mathcal{S}_{\text{anticipation}}(t), \ldots, \mathcal{S}_{\text{trust}}(t))$$

where $\mathcal{S}_i(t)$ is the re-based count of words associated to one emotion with respect to the speech length N:

$$\mathcal{S}_i(t) = \sum_{k=1}^{N} \frac{I(\text{word} \in \text{speech})(t)}{N(\text{word} \in \text{speech})(t)}$$

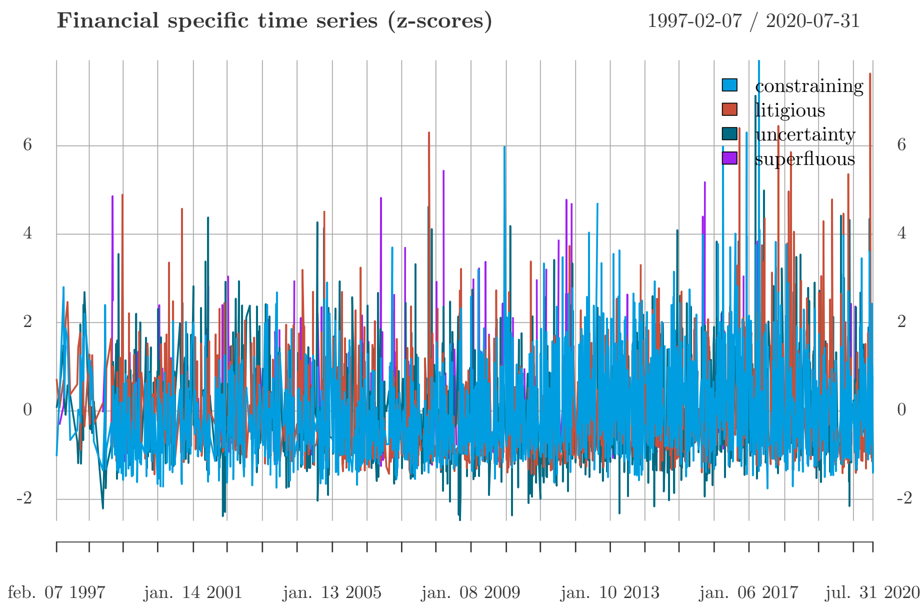
where $I(\text{word} \in \text{speech})(t) = 1$ if the word in the speech given at date t is in the list describing emotion I and 0 otherwise, and N the total number of words in the speech. Loughran and McDonald (2011) introduced a financial lexicon indicating for sentiment negative, positive, litigious, uncertainty, constraining, or superfluous. We represented the normalized detection of the Financial terms on Figure 18.

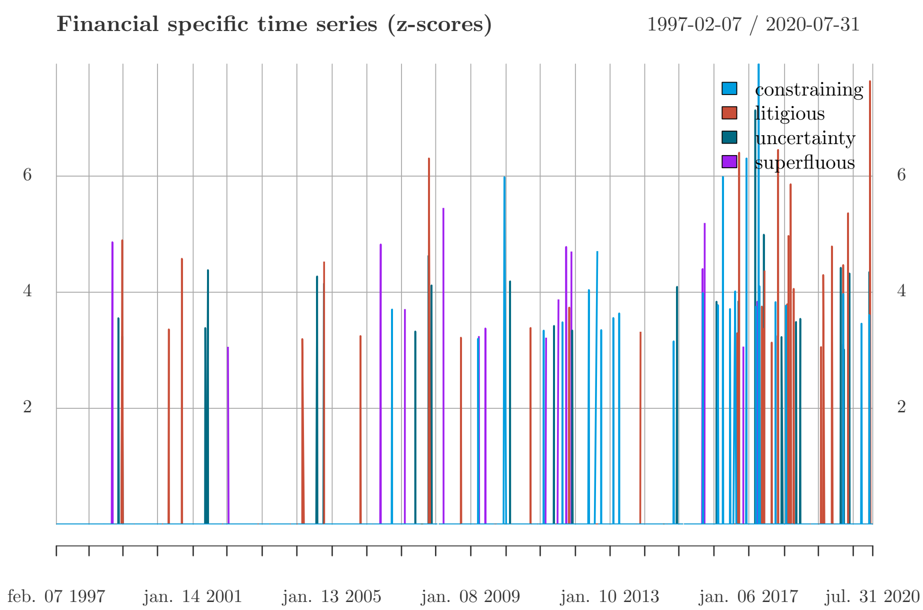
This process reacts to higher concentration of words in the corresponding lexicon. Ignoring positive and negative classification, already available in the general purpose lexicons, the four remaining dimensions presents series with little correlation to each other (Figures 18 and 19). Indeed, we can observe *spikes* in the concentration of terminology associated to litigious or uncertainty.

This process provide simple averaged word count signals. We tested different processing, for instance controlling the scores when the word were preceded by not words, and we found the simple average specification to be robust in this particular context. These aggregated scores from each of these open source lists provides time-series presenting redundancy.

Principal component analysis This method consists in creating representative objects, the principal components, carrying the variance of the full database. For instance, we constructed 18 open source time-series from the scores described above. The scree-plot Figure 20 shows that 5 dimensions allow to embed 77.2% of the variance. The 3 first dimensions alone, 66.4% and the 2 first, plotted Figure 8, 59.5%. We constructed the mapping given Table 3.

Figure 18: Loughran and McDonald (2011) detection over ECB's communication

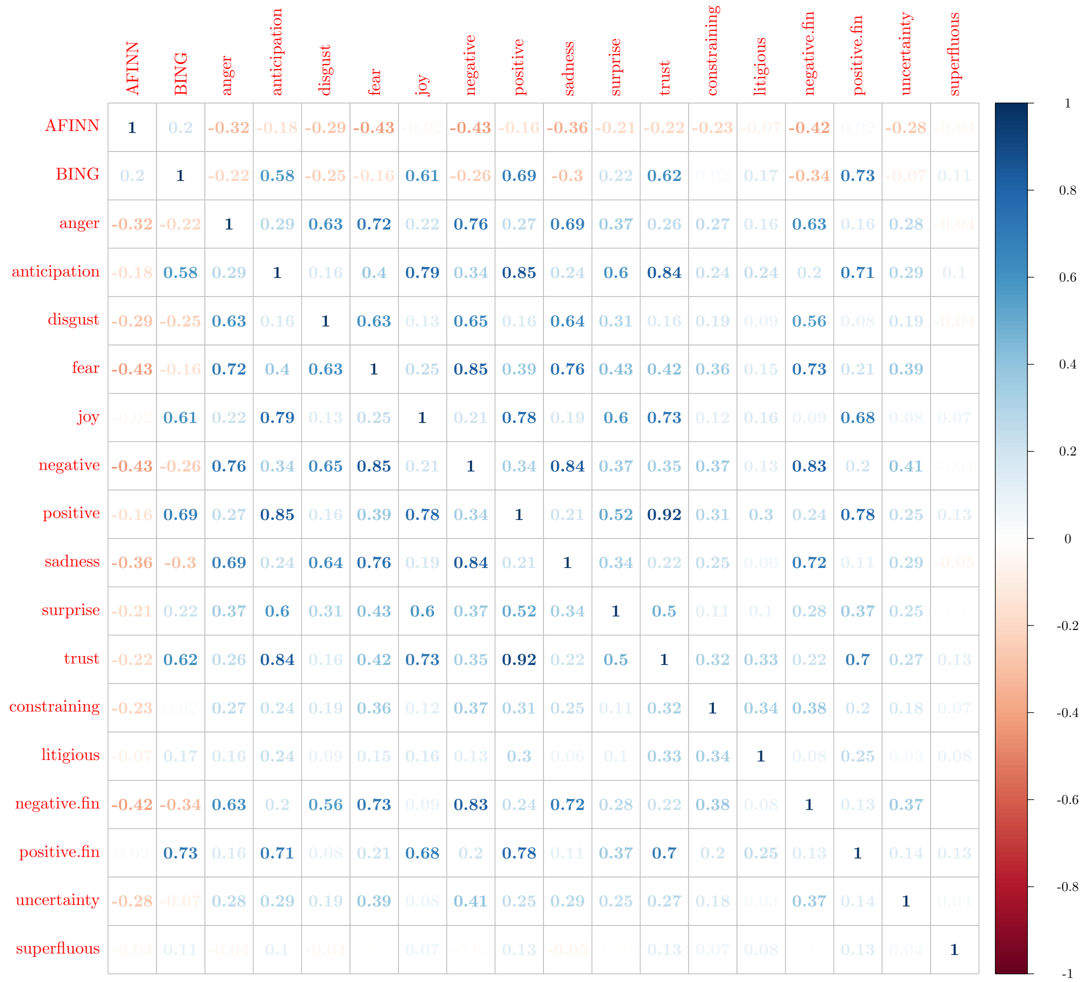




All series above and spots over 3 standard deviation bellow.

The positive and negative lexicons are highly correlated to each other. Figure 8 indeed shows similar projection on PC1 and PC2 of corresponding sentiment, the *polarity* being mostly characterized through the PC1 component discriminant contribution. AFINN appears perfectly anti-correlated (and therefore redundant) with negative dictionaries indicators and also present low correlation with the positive general emotions dictionary. Figure

Figure 19: Correlation sentiment score series



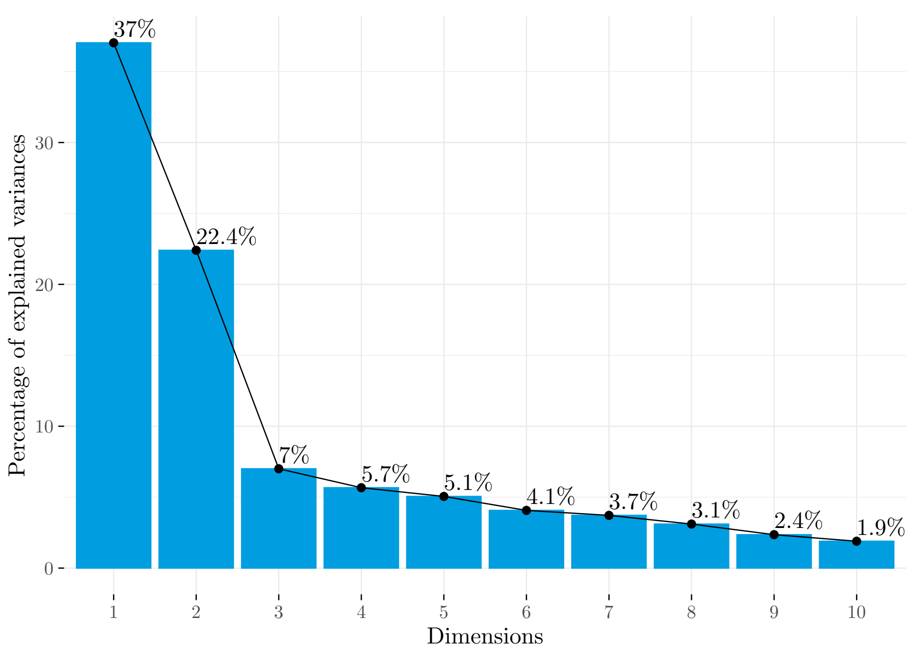
8 and Table 3 shows that the two first components of the PCA fails to capture Loughran and McDonald (2011) dictionaries: uncertainty requires PC4 to PC6, superfluous is also captured by PC5. These components are *hybrid* scores we enter the Lasso analysis in our exercise.

As a conclusion, under the current specification and given the restricted content used, principal components adds little value to the initial signals, but they are useful to understand the decomposition of the input signal variances. They show that there is little effect of specific emotions over the binary positive/negative component. This observation is made in the context of the full database, with no distinction between speech, interview or press conference. The context and level of formalism can explain this observation. On the hand, we also show that specific and dedicated vocabulary must be specified in a dictionary based process. Therefore, Loughran and McDonald (2011) based series could have a value-added on the prediction of the market variables.

Table 3: Component contributions (%)

	Score	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
	AFINN	2.59	2.84	2.63	14.29	7.02	50.66	4.59	8.81	0.72	0.02
	BING	0.84	19.91	0.05	0.20	0.00	0.54	1.65	3.13	2.55	0.11
NRC	anger	7.08	5.08	0.63	3.05	1.73	0.08	1.38	0.18	0.00	50.23
	anticipation	8.80	6.32	0.56	0.72	0.33	0.24	0.11	0.14	0.43	1.95
	disgust	4.85	5.83	1.87	4.24	5.79	0.26	0.92	1.62	65.51	4.87
	fear	9.64	4.36	0.03	0.02	0.09	0.16	0.11	0.08	0.88	2.03
	joy	6.20	8.19	5.05	0.47	0.87	1.21	0.02	2.20	0.56	0.12
	negative	9.43	6.47	0.02	0.04	0.20	1.49	0.51	0.58	3.52	0.18
	positive	9.11	7.63	0.05	0.06	0.15	0.07	0.96	0.94	0.00	0.01
	sadness	6.92	7.40	1.53	0.87	1.74	0.31	0.31	0.43	5.98	0.66
	surprise	6.31	0.61	9.94	0.46	0.07	10.12	11.14	37.99	0.07	5.74
	trust	8.85	6.45	0.52	0.25	0.61	0.19	0.44	1.21	0.03	1.83
Loughran	constraining	2.98	0.18	31.85	2.50	4.42	1.47	15.58	31.97	5.68	0.63
	litigious	1.37	0.67	31.13	17.05	0.24	3.31	37.89	3.18	2.39	1.82
	negative.fin	6.62	7.60	0.37	0.07	0.04	2.43	3.37	0.02	7.97	20.94
	positive.fin	5.44	9.17	0.00	0.51	0.55	5.64	1.98	6.33	0.10	8.74
	uncertainty	2.91	0.73	0.25	30.55	16.81	21.80	18.49	0.10	3.60	0.04
	superfluous	0.08	0.58	13.51	24.65	59.34	0.02	0.55	1.07	0.00	0.09

Figure 20: Scree plot



A.4 Topic analysis

Even-though the role of a central bank is highly specific, which restrict the possibilities in terms of possible topics, one must distinguish if the main subject is, for instance, energy price, unemployment or discussion about the potential causes of the next financial crisis. To perform an unsupervised topic analysis of a corpus, a standard approach is to use the Latent Dirichlet Allocation (LDA) model, introduced by Blei *et al.* (2003). Then, multiple methodological improvements have been developed to supervise and optimize results (Mcauliffe & Blei, 2008; Ramage *et al.*, 2009; Zhu *et al.*, 2012).

A.4.1 Structural Topic Model (STM)

In this study, we use M. E. Roberts *et al.* (2019) Structural Topic Model R package (M. Roberts *et al.*, 2020). This method adds richer structure to previous probabilistic models, notably the Latent Dirichlet Allocation (LDA) and its extensions. For instance, the Correlated Topic Model (CTM), which allows correlation between topics (Blei, Lafferty, *et al.*, 2007) and the Dirichlet-Multinomial Regression topic model (D. M. Mimno & McCallum, 2008), which attempted to incorporate general covariate information into topic model.

In essence, the STM, like other probabilistic topic models (e.g. LDA), is a mixture over words where each word has a probability of belonging to a topic. And a document is a mixture of topics. As such, the sum of topic proportions across all topics for a document is one, and the sum of word probabilities for a given topic is one. But the STM model uses a correlated topical model structure that can be enhanced by using external covariate information about the experimental design into the model (e.g. authors, gender, political party, location). This is a major innovation on topic modeling. In other wors, in this specification, topical prevalence and topical content can be a function of document metadata (i.e. additional information). One must distinguish these two concepts:

- *Topical prevalence* refers to how much a document is associated with a topic;
- *Topical content* refers to the words used within a topic.

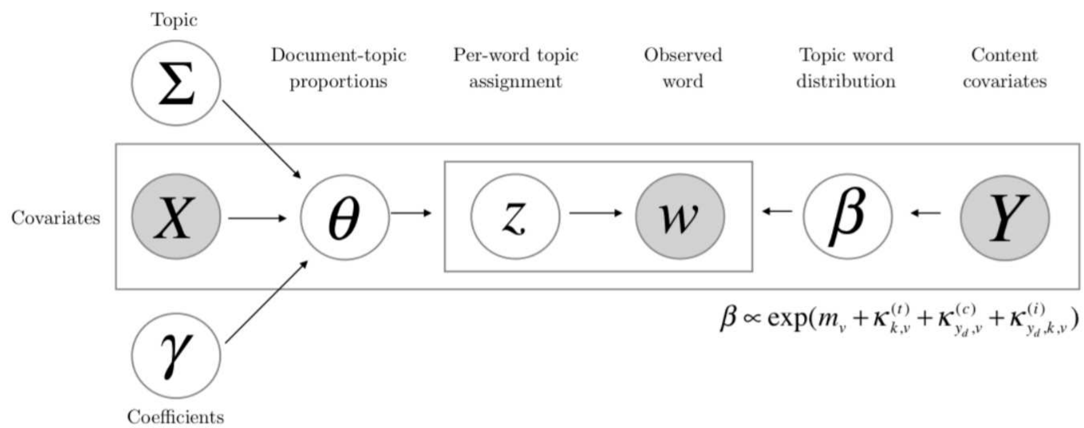
Hence, metadata associated to topical prevalence are referred by the authors as *topical prevalence covariates*. Variables associated to topical content are referred to as *topical content covariates*. That being said, the model is flexible: it allows using or not topical prevalence covariates, a topical content covariate, both, or neither. In the case of no covariates, the model reduces to a (fast) implementation of the Correlated Topic Model (CTM) (Blei, Lafferty, *et al.*, 2007). The STM is then general and suitable to a large number of areas of research.

The data generating process summarized bellow follows M. E. Roberts *et al.* (2019) and M. E. Roberts *et al.* (2016) and Figure 21 shows a graphical illustration. Essentially, each document indexed by d ($d \in 1,...$ D) with a vocabulary of terms, indexed by v ($v \in 1, ..., V$), given the number of topics K, can be expressed as:

1. Draw topic proportions in documents θ_d

$$\theta_d | x_d \Gamma, \Sigma \sim \text{LogisticNormal}\left(\mu = x_d \Gamma, \Sigma\right)$$

Figure 21: Graphical illustration of the Structural Topic Model (M. E. Roberts *et al.*, 2016)



Source : M. E. Roberts *et al.* (2016).

where x_d is a vector of document prevalence covariates size p, $\Gamma = [\gamma_1, \ldots, \gamma_K]$ is a matrix $(p, K-1)$ of coefficients of topical prevalence and Σ is the $(K-1, K-1)$ covariance matrix.

2. Given a document-level content covariate y_d, draw the document-specific distribution over words representing each topic (k):

$$\beta_{d,k,v} \propto \exp\left(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(r)}_{y_d,k,v}\right)$$

where:

- m_v is the baseline word distribution[56]
- $\kappa^{(t)}_{k,v}$ is the topic specific deviation
- $\kappa^{(c)}_{y_d,v}$ is the covariate group deviation
- $\kappa^{(i)}_{y_d,k,v}$ is the interaction between the two

Note that when no covariate is introduced in the model, the expression becomes $\beta_{d,k,v} \propto \exp\left(m + \kappa^{(t)}_{k,v}\right)$ or simply point estimated (this latter behavior is the default).

3. For each word in each document, (n $\in$ 1, ..., N_d):

   (a) Draw each word's topic assignment based on the document distribution over topics:

$$z_{d,n}|\theta_d \sim \text{Multinomial}(\theta_d)$$

[56]The authors highlight that m_v is specified as the estimated (marginal) log-transformed rate of occurrence of term v in the document collection under study (e.g. Airoldi *et al.* (2005)), but it can alternatively be specfied as any baseline distribution of interest.

(b) Conditional on the topic chosen, draw an observed word from that topic:

$$w_{d,n}|z_{d,n}, \beta_{d,k} = z_{d,n} \sim \text{Multinomial}(\beta_{d,k} = z_{d,n})$$

The STM approach to estimation builds on prior work in variational inference for topic models (e.g. Blei, Lafferty, *et al.* (2007)). In particular, the authors developed a partially-collapsed variational Expectation-Maximization algorithm which upon convergence gives the estimates of the models parameters. Regularizing prior distributions are used for γ, k and (optionally) Σ to help enhancing interpretation and preventing overfitting. Further technical details can be found on M. E. Roberts *et al.* (2016).

This model also computes the FREX metric (Airoldi & Bischof, 2012, 2016; M. E. Roberts *et al.*, 2019) to measure exclusivity balancing word frequency. FREX is the weighted harmonic mean of the word's rank in terms of exclusivity and frequency:

$$\underset{k,v}{\text{FREX}} = \left(\frac{\omega}{\text{ECDF}(\beta_{k,v} / \sum \beta_{j,v})} - \frac{1-\omega}{\text{ECDF}(\beta_{k,v})} \right)^{-1}$$

where ECDF is the empirical CDF and ω is the weight which set to .7 to favor exclusivity. The STM package allows to visualize topic correlation (see Figure 33), and provide cluster tool with respective FREX metrics.

A.4.2 Optimal number of topic

This step consists in comparing metrics in several topic models to maximize or minimize to find – the scale of magnitude of – the optimal number of topics and their specification. Several study revealed that their is no right answer, but several optimization process were developed providing a value for the optimal number of topics (Arun *et al.*, 2010; Cao *et al.*, 2009; Deveaud *et al.*, 2014; Griffiths & Steyvers, 2004)[57]. The optimal number of topics was determined by maximizing the coherence score, that are based on averaging measures of pairwise association between the most probable worlds in a topic (Newman *et al.*, 2010):

$$\hat{c}(k) = \sum_{i<j} F_s(w_i, w_j)$$

where w_i, w_j are the top words of the topic, and F_s is the coherence scoring function. There are multiple measures for coherence, we find several example of specifications on the Palmetto project platform[58]:

- C_V measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized point-wise mutual information (NPMI) and the cosine similarity (Röder *et al.*, 2015)

- C_P is based on a sliding window, one-preceding segmentation of the top words and the confirmation measure of Fitelson's coherence (Röder *et al.*, 2015)

[57] The comparison of the metrics are proposed by ldatuning R package (Nikita, 2016).
[58] https://palmetto.demos.dice-research.org/.

- C_{UCI} extrinsic measure is based on a sliding window and the point-wise mutual information (PMI) of all word pairs of the given top words (Newman *et al.*, 2010)
- C_{UMass} intrinsic measure is based on document co-occurrence counts, a one-preceding segmentation and a logarithmic conditional probability as confirmation measure (D. Mimno *et al.*, 2011)
- C_{NPMI} is an enhanced version of the C_{UCI} coherence using the normalized pointwise mutual information (NPMI) (Aletras & Stevenson, 2013)
- C_A is based on a context window, a pairwise comparison of the top words and an indirect confirmation measure that uses normalized point-wise mutual information (NPMI) and the cosine similarity (Aletras & Stevenson, 2013)

Formally, the most commonly used coherence score is the C_{UMass}, specified as follow:

$$C_{UMass}(k, \mathcal{V}(k)) = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \log \frac{D(w_m^{(k)}, w_j^{(k)}) + 1}{D(w_j^{(k)})}$$

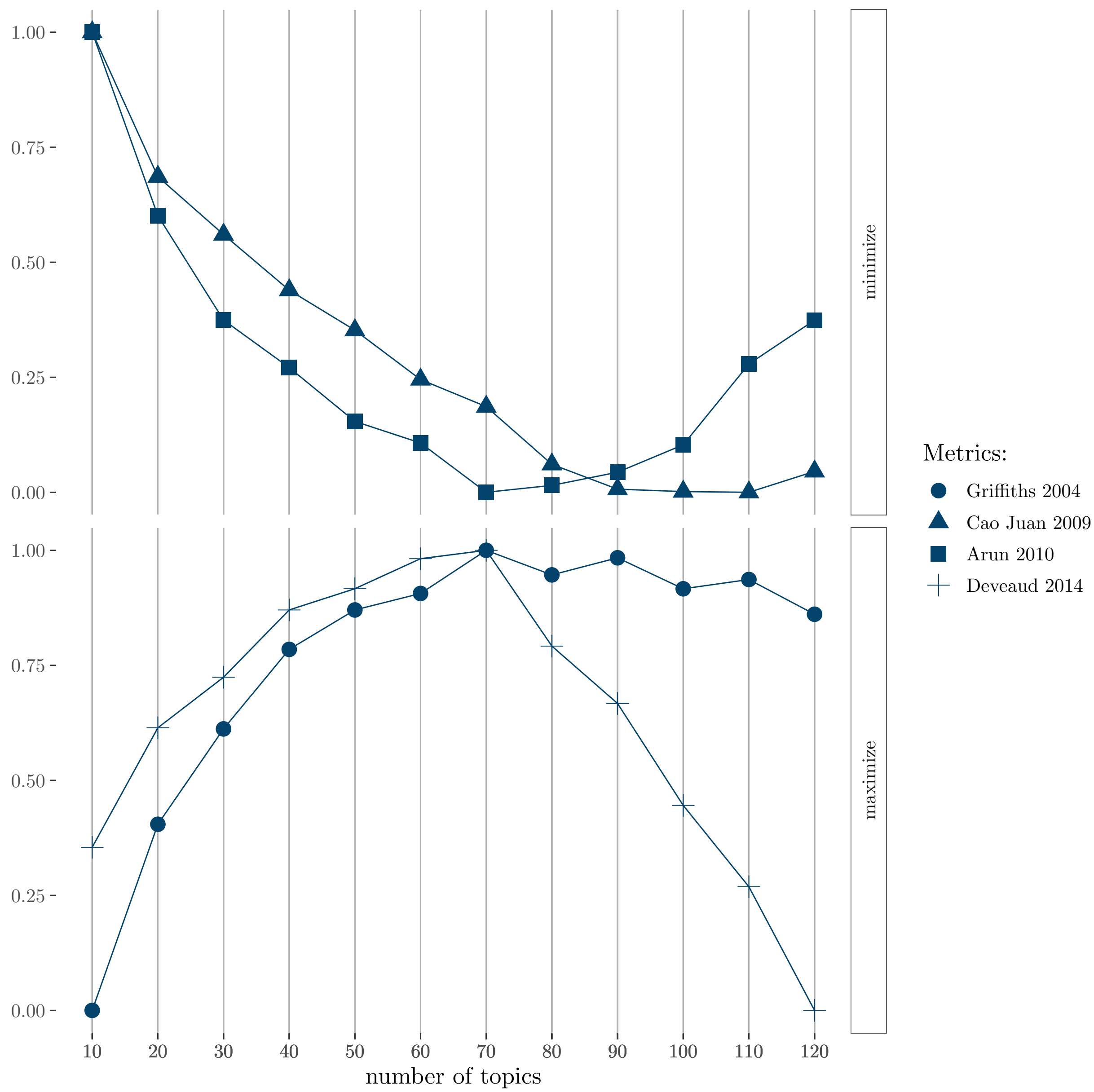
where $\mathcal{V}(k) = (w_1^{(k)}, ..., w_M^{(k)})$ is a list of the M most probable words in topic k and where "*$D(v)$ is the document frequency of word type v (i.e., the number of documents with least one token of type v) and $D(v, v')$ be co-document frequency of word types v and v′ (i.e., the number of documents containing one or more tokens of type v and at least one token of type v′)*"(D. Mimno *et al.*, 2011, p. 265). Comparing different metrics (Figure 22), we can see that the optimal number of topics generally is between 70 and 100. Additionally the spectral initialization (D. Mimno & Lee, 2014) of structural topic modeling (M. E. Roberts *et al.*, 2019) package converged to 87[59]. This number can vary depending on the preprocessing, input database (with or without interviews), beginning of the sample, but consistency of topics provided by the model remains perfectly acceptable.

A.4.3 Topic clustering

In this section we describe the distribution of the topics used in the analysis in time. We gathered the 87 topics in thematic clusters to be able to compare the change in general thematic discourse. We note that overall the coverage of themes is stable overtime, while underlying specific topics distribution are strongly concentrated in small time period. This suggest that using these topics to measure the impact on the communication on financial variable can be pertinent on rather small time period only. The Figure 5 on page 20 represents the percentage of the general communication represented by each cluster overtime. The following figures detail the breakdown of the subjects. The vertical axis gives the relative probability of presence of the topic, normalized per its maximum prevalence score. The normalize value follows: $\gamma_{k,t}/\max_t(\gamma_k)$.

[59] Among other values, according to the computer used, pre-possessing complexity, input database (with or without interviews), etc.

Figure 22: Optimal number of topics



Source: Implementation of Nikita (2016) *ldatuning* R package on data processed by the authors

Figure 23: Shift in '*Banking*' discourse probability of presence

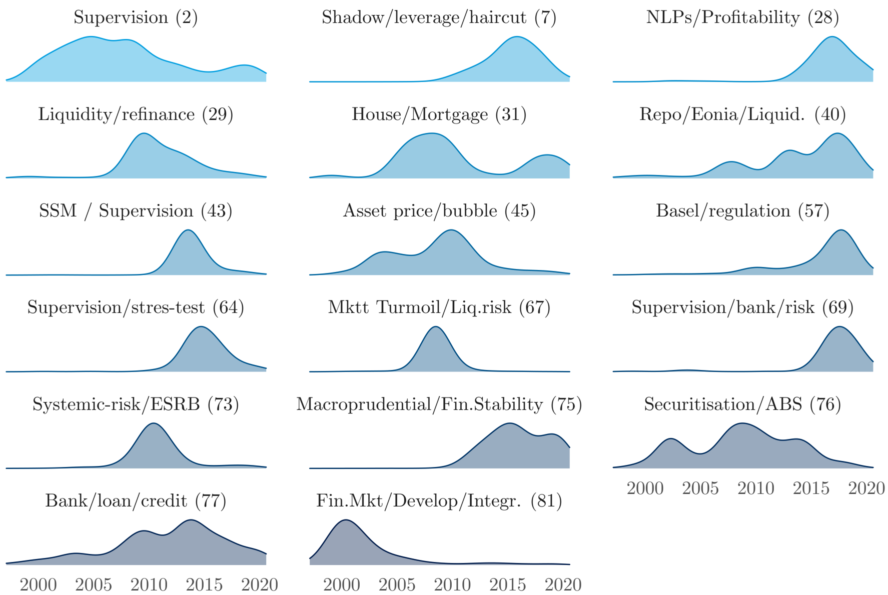


Figure 24: Shift in '*Monetary Policy and Inflation*' discourse probability of presence

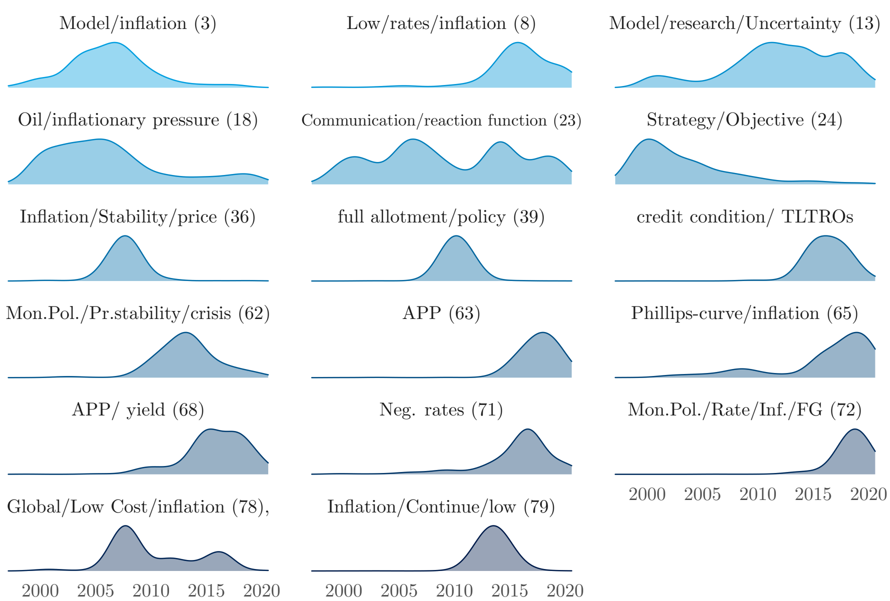


Figure 25: Shift in '*trade and international relations*' discourse probability of presence

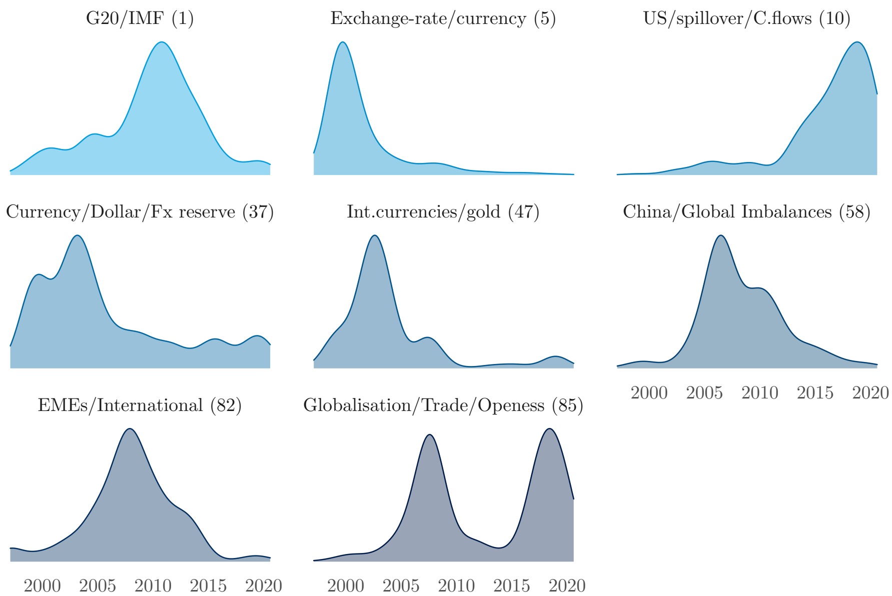


Figure 26: Shift in '*Payment, Settlement, Statistics*' discourse probability of presence

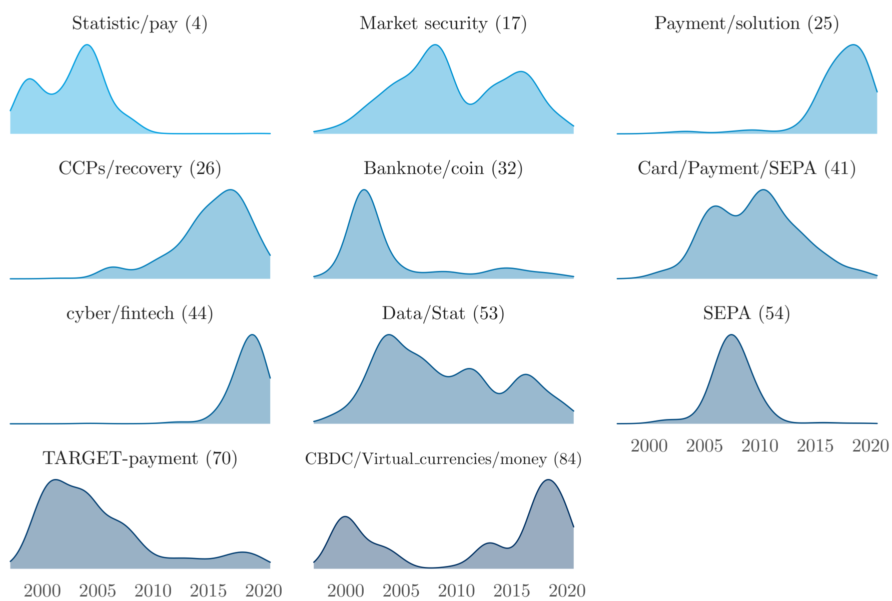


Figure 27: Shift in '*Fiscal Policy*' discourse probability of presence

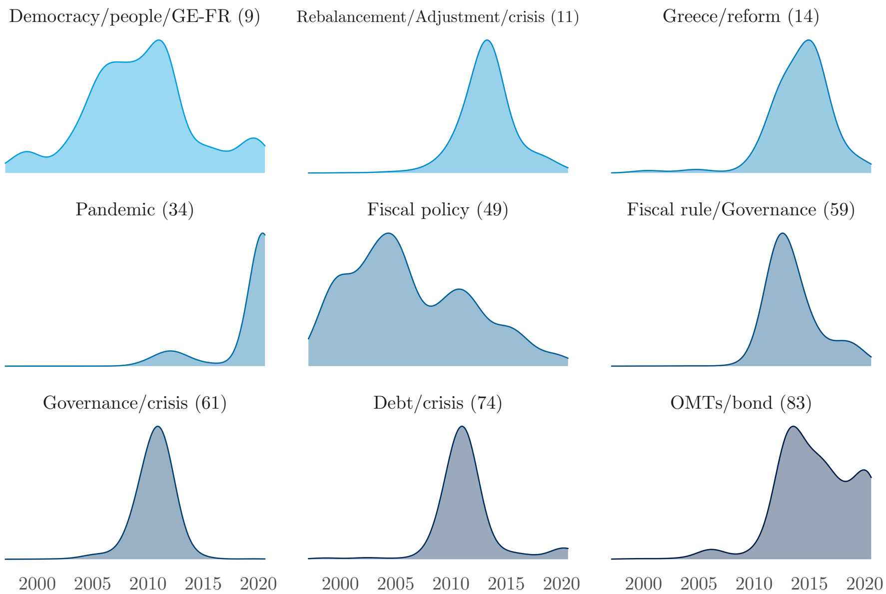


Figure 28: Shift in '*Growth and structural reforms*' discourse probability of presence

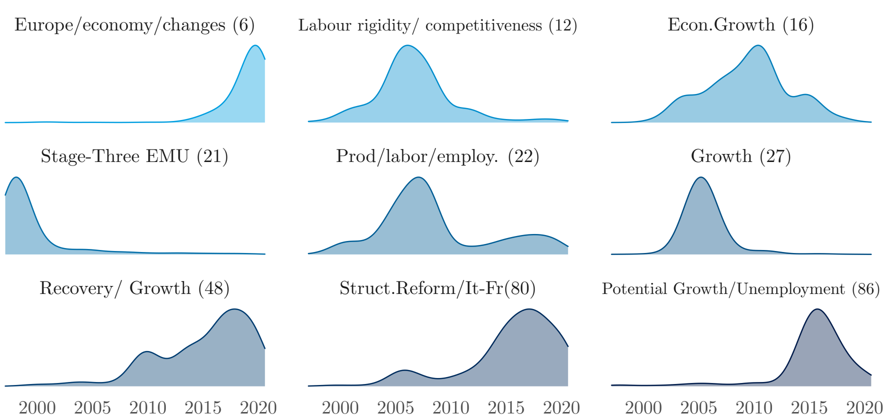


Figure 29: Shift in '*Euro construction and EMU*' discourse probability of presence

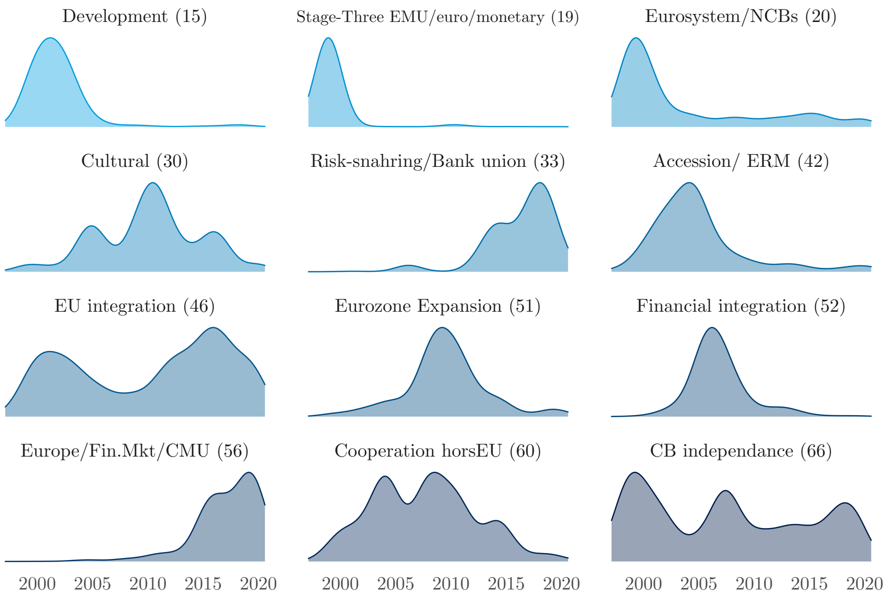
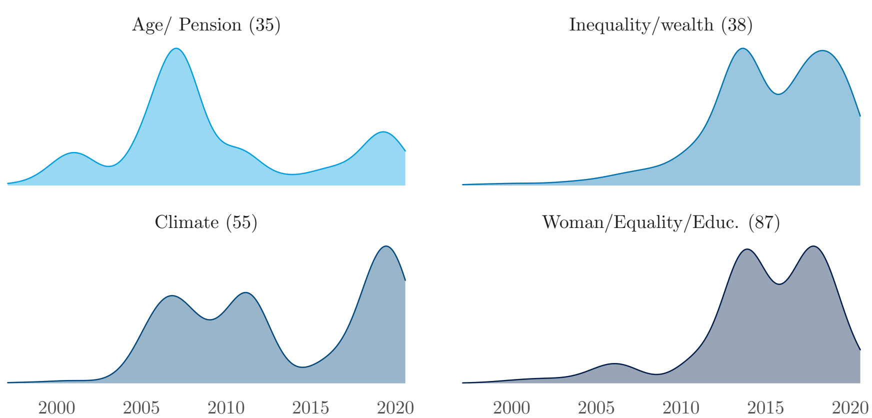


Figure 30: Shift in '*Extra-financial questions*' discourse probability of presence



B Notations

Indices and sets	
$j \in \mathcal{A}$	Country
$l \in \mathcal{M}$	Control variables
$k \in \{1, \cdots, K\}$	Topics
$s \in \{1, \cdots, S_m\}$	Sentiments
$v \in \{1, \cdots, V\}$	Vocabulary/Dictionary
$d \in \{1, \cdots, D\}$	Documents
$n \in \{1, \cdots, N_d\}$	Token (position or) index in document d
Variables	
EUR/USD	EUR/USD (euro) exchange rate
$y_{j,t}$	Bond yield time series of Country j
$\text{OIS}_{T,t}$	OIS rates time series at maturirity T for the euro area
$\text{FC}_{l,t}$	Control variables time series
$Sm_{s,t}$	Sentiment time series
θ_d	Topic mixture / proportion of each topic in document d
x_d	Vector of document topical prevalence covariates
y_d	Vector of document topical content covariates
Γ	Matrix coefficients for topic prevalence
Σ	Covariance matrix
$\beta_{d,k,v}$	Word probabilities for each topic
γ_k	Probability of presence of a topic
$w_{d,n}$	Observed word (in document d at position n)
$z_{d,w}$	Per-word topic assignment
m_v	Baseline word distribution
$\kappa^{(t)}_{k,v}$	Topic specific deviation
$\kappa^{(c)}_{y_d,v}$	Covariate group deviation
$\kappa^{(i)}_{y_d,k,v}$	Interaction between topic and covariate deviations
$c(k), (C)$	Coherence measure
Parameters	
T	Maturity
$\mathcal{A}$	Country
$\mathcal{M}$	Control variables
K	Topics
S_m	Sentiments
V	Vocabulary/ Dictionary
D	Documents

C Glossary

ABS *Asset-Backed Securities* It is an investment security (a bond or a note) which is collateralized by a pooll of assets such as loans, leases, credit card debt, royalties.

ACCBs *Accession Country Central Banks* It corresponds to the Central Banks of countries aiming at being part of the euro area.

ACP countries *African, Caribbean and Pacific countries* is a grouping of 79 African, Caribbean and Pacific states with which the European Union has a special relationship, particularly regarding trade.

Current trade arrangements are based upon Economic partnership agreements (EPAs). EPAs link the EU to ACP countries in seven regional groupings:

AnaCredit *Analytical Credit Datasets* It is a central credit register created by the ECB to harmonize statistical database on credit provided by financial institutions within the eurozone and/or that are part of the Single Supervisory Mechanism (SSM).

APP *Asset Purchase Programme* It is s part of a package of unconventional monetary policy measures on which central bank buys government bonds or other financial assets to support the monetary policy transmission mechanism and provide the amount of policy accommodation needed to ensure price stability. It was initiated by the ECB in mid-2014 and it comprises currently four different programms: 1) corporate sector purchase programme (CSPP); 2) public sector purchase programme (PSPP); 3) asset-backed securities purchase programme (ABSPP) and 4) third covered bond purchase programme (CBPP3). In response to the coronavirus outbreak, a temporary asset purchase programme of private and public sector securities was also launched, the *Pandemic Emergency Purchase Programme (PEPP)*, aiming at providing additional support to the euro area. .

BRRD *Bank Recovery and Resolution Directive* This Directive came into effect on 2015 with the purpose of creating a common framework in the European Union for the recovery and resolution of banks and investment firms. It is considered a key component of European efforts to end the "too big to fail" problem.

BVAR *Bayesian Vector Auto-regression* It is a statistical technique that uses Bayesian methods to estimate vector auto-regression (VAR) models.

CBDC *Central Bank Digital Currency* It is also called *Digital Base Money* or *Digital Fiat Currency*. It is the digital form of fiat money - a currency established as money by a government regulation, monetary authority or law.

CCPs *Central Counterparties Clearing* It corresponds to the Clearing members and/or clearing activities which are authorized to offer services in European Union.

CCyB *Countercyclical Capital Buffer* It is a regulatory capital adequacy ratio requirement applied to banks taking into account of the macro-financial environment in which they operate. The rate can vary (as a proportion of risk-weighted assets) and should be increased suring the upswing of the financial cycle and relaxed during a downturn.

CEBS *Committee of European Banking Supervisors* It was established with the objective to advise the European Commission on legislative/regulatory measures in the banking field, to contribute to the consistent implementation of Community Directives and to the convergence of Member States' supervisory practices throughout the Community. CEBS is comprised of high level representatives from the banking supervisory authorities and central banks of the European Union.

CFA Franc Zone It is an economic and monetary area bringing together France and 15 countries in Sub-Saharan Africa. Note that the CFA franc is one of two regional African currencies backed by the French treasury with pegging to the euro. It can refer to either the Central African CFA franc, or the West African CFA franc.

CIISI *Cyber Information and Intelligence Sharing Initiative* It is an scheme aiming at protecting the financial system by preventing, detecting and responding to cyber attacks.

CMU *Capital Markets Union* It corresponds to an European Union (EU) initiative which aims to deepen and further integrate the capital markets of EU member state.

CPMI *Committee on Payments and Market Infrastructures* It is an international standard setter that promotes, monitors and makes recommendations about the safety and efficiency of payment, clearing, settlement and related arrangements, thereby supporting financial stability and the wider economy. The CPMI also serves as a forum for central bank cooperation in related oversight, policy and operational matters, including the provision of central bank services.

CRE loans *Commercial Real State loans* It is a mortgage security by a lien on a commercial, rather than residential, property.

CRT *Credit Risk Transfers* They are instruments that can be used by banks for credit risk shedding (protection buying) and/or risk taking (protection selling) purposes. They can also make markets in credit derivatives aiming at running matched credit risk positions.

CSDs *Central Securities Depositories* They are financial organizations that specialize in holding securities, including equities, bonds and money market instruments.

DLT *Distributed ledger technology* They are also called *Distributed ledger* or *shared ledger*. It is a consensus of replicated, shared and synchronized digital data geographically spread across multiple sites, countries or institutions. There is no central administrator (e.g. the blockchain system).

EAA *Euro area accounts* It provides a comprehensive picture of how economic value is generated and distributed in the euro area economy on the basis of an analytical grouping of economic agents into institutional sectors based on the methodological framework established in the European System of Accounts 2010 (ESA 2010).

ECRB *Euro Cyber Resilience Board for pan-European Financial Infrastructures* It is a forum for strategic discussions between financial market infrastructures. Its objectives are to raise awareness of the topic of cyber resilience, foster trust and collaboration and catalyse joint initiatives to develop effective solutions for the market.

EMI *European Monetary Institute* It was the forerunner of the European Central Bank (ECB), operating between 1994 and 1997.

EMS *European Monetary System* It was an exchange rate regime set up in 1979 (and which ended in 1999) to foster closer monetary policy cooperation between the central banks of the Member States of the European Economic Community (EEC). The objective of the EMS was to promote monetary stability in Europe. It was built on the concept of stable but adjustable exchange rates defined according to the newly created European currency unit (ECU) – a currency basket based on a weighted average of EMS currencies. Within the EMS, currency fluctuations were controlled through the Exchange Rate Mechanism (ERM).

EMU *Economic and Monetary Union* The European Council confirmed the progressive realization of the Economic and Monetary Union (EMU) in June 1988. To achieve this objective, three evolutionary steps were proposed: i) Stage One (starting in July 1990): it enhanced, for example, the complete freedom for capital transactions. ii) Stage Two (starting in January 1994) which, for example, increased the coordination of monetary policies among countires and promoted the process leading to independence of national central banks. iii) Stage Three (Starting on January 1999): set the introduction of the euro and put into force the Stability and Growth Pact, among others. Further details on https://www.ecb.europa.eu/ecb/history/emu/html/index.en.html

EPC *European Payments Council* It was founded in 2002 aiming at developping the Single Euro Payment Area. It represents payment service providers, supports and promotes European payments integration and development, notably SEPA.

EPP *European People' Party* It is an European political party with Christian-democratic, conservative and liberal-conservative member parties. It includes major centre-right parties such as the CDU/CSU of Germany and The Republicans of France. It has been one of the largest party in the European Parliament since 1999.

ERM *Exchange Rate Mechanism* It was a system introduced by the European Economic Community on 13 March 1979, as part of the European Monetary System (EMS), in order to stabilise exchange rates and help Europe to become an area of monetary stability before the introduction of the single currency, the euro, which took place on 1 January 1999.

ERM - II *Exchange Rate Mechanism II* The Exchange Rate Mechanism (ERM II) was set up on 1 January 1999 as a successor to ERM to ensure that exchange rate fluctuations between the euro and other EU currencies do not disrupt economic stability within the single market, and to help non euro-area countries prepare themselves for participation in the euro area. The convergence criterion on exchange rate stability requires participation in ERM II.

ESCB *European System of Central Banks* It comprises the ECB and the national central banks (NCBs) of all EU Member States whether they have adopted the euro or not.

ESM *European Stability Mechanism* It is an instrument designed to safeguard financial stability in the euro area. Like its predecessor, the temporary European Financial Stability Facility (EFSF), the ESM provides financial assistance to euro area countries experiencing or threatened by financing difficulties.

ESRB *European Systemic Risk Board* It was established on 2010 in response to the financial crisis. It is tasked with the macroprudential oversight of the financial system within the European Union in order to contribute to the prevention or mitigation of systemic risks to financial stability in the EU.

Eurosystem It is the Monetary Authority of the euro area. It comprises the European Central Bank (ECB) and the national central banks of the Member States whose currency is the euro.

FCIs *Financial Conditions Indexes* The FCIs are constructed as weighted averages of different financial variables. The ECB FCIs, in particular, include the 1-year OIS, the 10-year OIS, the Nominal Effective Exchange Rate (NEER) of the euro vis-à-vis 38 trading partners, and the Euro STOXX Index.

FMIs *Financial Market Infrastructures* It is defined as a multilateral system among participating institutions, including the operator of the system, used for the purposes of clearing, settling, or recording payments, securities, derivatives, or other financial transactions.

FSAP *Financial Services Action Plan* is a key component of the European Union's attempt to create a single market for financial services. Created in 1999 and to last for a period of six years, it contained 42 articles related to the harmonisation of the financial services markets within the European Union. It was scheduled to be completed by the end of 2004.

IEA *International Energy Agency* It is an intergovernemental organization acting as a policy advisor to help countries provide secure and sustainable energy for all.

IMS *International Monetary System* It refers to the system and rules that govern the use and exchange of money around the world and between countries.

IMS *International Monetary System* It refers to the system and rules that govern the use and exchange of money around the world and between countries.

IOSCO *International Organization of Securities Commissions* It refers to the system and rules that govern the use and exchange of money around the world and between countries.

IPN *Inflation Persistence Network* The IPN is a research team consisting of economists from the European Central Bank and the national central banks of the Eurosystem conducting a coordinated research project on the patterns, determinants and implications of inflation persistence in the euro area and in its member countries.

LCBGs It is an acronym used to refer to *Large and Complex Banking Groups*.

LCR *Liquidity Coverage Ratio* It is one ratio used as policy response for the liquidity crisis. It aims at promoting the short-term resilience of the liquidity risk profile of banks. It does this by ensuring that banks have an adequate stock of unencumbered high-quality liquid assets (HQLA) that can be converted easily and immediately in private markets into cash to meet their liquidity needs for a 30 calendar day liquidity stress scenario.

Libra *Libra Digital Currency* It is a blockchain-based payment system proposed by Facebook.

LMU *Latin Monetary Union* It was a system (established in 1865 and disbanded in 1927) that unified several European currencies into a single one and that could be used in all the member states. Many countries minted coins according to the LMU standard even though they did not formally accede to the LMU treaty.

LTV ratio *Loan-to-value ratio* It is an assessment of lending risk regurlarly used by financial institutions before approving a mortgage. Typically, high LTV ratios are considered higher risk loans.

M3 *Monetary Aggregates* They comprise monetary liabilities of monetary financial institutions (MFIs) and central government, vis-à-vis non-MFI euro area residents excluding central government. **M1** corresponds to the sum of currency circulation and overnight deposits; **M2** is the sum of M1, deposits with an agreed maturity of up to two years and deposits redeemable at notice of up to three months; **M3** is the sum of M2, repurchase agreements, money markets fund shares/units and debt securities with a maturity of up to two years.

MaRs *Macroprudential Research Framework* It is an internal network for macroprudential research launched in Spring 2010 by the European System of Central Banks. It was created in response to the 2007-2008 Global Financial crisis which put in evidence that macroprudential aspects of financial supervision and regulation had to be significantly strengthened.

MNEs It is an acronym used to refer to *Multinational Enterprises*.

MTOs *Medium-term budgetary objectives* It refers to a country budgetary target in accordance with the Stability and Growth Pact (SGP) with the objective of supporting EU governments in achieving their commitments on sound fiscal policies. They take into account the need to achieve sustainable debt levels while ensuring governments have enough room to manoeuvre and a safety margin against breaching the EU's fiscal rules. All EU countries are expected to reach their medium-term budgetary objectives (MTOs), or to be heading towards them by adjusting their structural budgetary positions at a rate of 0.5% of GDP per year as a benchmark. MTOs are updated every 3 years, or more frequently in the case of a country that has undergone a structural reform which significantly impacted its public finances.

NBH It is an acronym used to refer to *National Bank of Hungary*.

NBRM It is an acronym used to refer to *National Bank of the Republic of Macedonia*.

NCBs It is an acronym used to refer to *National Central Banks*.

NPLs *Non-performing loans* It consists of a loan in which the borrower is default.

OCA *Optimum Currency Area* It corresponds to a geographical region that would maximize its economic efficiency having the entire region sharing a single currency.

OMTs *Outright Monetary Transactions* It is a program of the ECB launched in 2012, in response to the financial fragmentation stemming from the sovereign debt crisis, under which the central bank could make purchases ("outright transactions") in the secondary sovereign bond markets, under certain conditions, of bonds issued by Euro area member states. One of the most relevant technical features of OMTs was no ex ante quantitative limits would be considered for this transactions.

PEPP *Pandemic Emergency Purchase Programme* It is a temporary asset purchase programme of private and public sector securities to counter the serious risks to the monetary policy transmission mechanism and the outlook for the euro area posed by the outbreak and escalating diffusion of the coronavirus, COVID-19. It was launched in March 2020 with an amount of €750 billion and extended by EUR 600 billion to EUR 1.35 trillion in June 2020. The ECB also extended the life of PEPP from the end of 2020 to June 2021.

PES *Party of European Socialists* Social-democratic European political party. It comprises national-level political parties from all member states of the European Union (EU) in additon to Norway and the United Kingdom.

Pfandbriefe *Pfandbrief bond market* is a type of bond issued by German mortgage banks that's collateralised by long-term assets (property mortgages or public sector loans, for example). Pfandbriefe are a classic funding instrument employed by Pfandbrief banks and constitute one of the largest segments of the European bond market.

PFMI *Principles for Financial Market Infrastructures* International standards for payment systems, central securities depositories, securities settlements systems, central counterparties and trade repositories. The CPMI and IOSCO monitor their implementation and its member committed to adopting their principles and responsibilities.

PSD *Payment Services Directive* It is an EU directive, administrated by the European Commission to regulate payment services and payment services providers throughout the European Union and European Economic Area (EEA).

PSPP *Public Sector Purchase Programme* It was unveiled by the ECB on 22 January 2015 and consisted of the purchase of bonds issued by euro-area governments, agencies and European institutions.

RTGS *Real-time gross settlement (TARGET 2)* It also known as TARGET 2, where the processing and settlement of payments takes place in real-time within the euro area. Each transfer is also settled individually.

SDD *SEPA Direct Debit* It is a payment method fully deployed and operated in the euro area member states.

SEPA / eSEPA *Single Euro Payments Area* It is a payment-integration initiative of the European Union for simplification of bank transfers denominated in euro. The eSEPA refers to the new technologies evolving the SEPA.

SNA *System of National Accounts* It is the internationally agreed standard set of recommendations on how to compile measures of economic activity.

SPEs *Special Purpose Entities* It is a subsidiary created by a parent company to isolate financial risk.

SREP *Supervisory Review and Evaluation Process* It is a supervisory activity on which there is an assessment on the risks faced by banks and on their capabilities on managing such risks effectively. Specifically, the SREP shows where a bank stands in terms of capital requirements and the way it deals with risks. In the SREP decision, which the supervisor sends to the bank at the end of the process, key objectives are set to address the identified issues. The bank must then "correct" these within a specific time.

SRM *Single Resolution Mechanism* It applies to banks covered by the single supervisory mechanism (SSM) and it is the second pillar of the banking union. The purpose of the SRM is to ensure an orderly resolution of failing banks with minimal costs for taxpayers and to the real economy.

SSM *Single Supervisory Mechanism* It refers to the system of banking supervision in Europe. It comprises the ECB and the national supervisory authorities of the participating countries.

TLTROs *Targeted Longer-term Refinancing Operations* They correspond to Eurosystem operations that provide financing to credit institutions and it was first launched in 2014. They are 'targeted' operations, as the amount that banks can borrow is linked to their loans to non-financial corporations and households. The third series of these operations (TLTROs - III) were announced in March 2019 and, similarly to TLTRO II, the interest rate to be applied is linked to the participating banks' lending patterns. The more loans participating banks issue to non-financial corporations and households (except loans to households for house purchases), the more attractive the interest rate on their TLTRO III borrowings becomes. The primary objective of these operations is to preserve favorable borrowing conditions for banks and stimulate bank lending to the real economy.

ULC *Unit Labor Costs* They are labor costs (e.g. wages and remunerations) per unit of added value produced.

D Data Description

All the variables used in our model have daily frequency and the sample starts on 01 January 2002 to 31 July 2020.

Overnight Index Swap (OIS) rates It is a fixed-for-floating interest rate swap with a floating rate leg tied to the index of daily interbank rates, that is the EONIA for the euro area. It is considered the best proxy for a risk-free yield curve in the region. We collected data with 1, 3, 6, 9 month, 1, 2, 3, 5, 10 year maturities. For each maturity, we calculated the daily return in bps. The series were obtained in Datastream.

Sovereign Yields Our data comprises sovereign bond yields with 2 and 10 year maturities for Germany, France, Italy, Spain and Portugal. For each of them, we calculated the daily returns in bps. For each maturity, we also defined two major groups. The "*core*" yields, which is the average yield of Germany and France and the "*peripheral*" which is the average yield of Italy, Spain and Portugal. Then, we also calculated the daily returns in bps. This grouping proved particularly effective during the Lasso selection. When bond yields for two or more countries in one of these groups were selected, we chose to use the variable as a group (hence, core or periphery) and not the yield related to a particular country. The series were obtained in Datastream.

Spreads Interest rate differentials between two countries play a key role in the currency markets.Thus, we calculated the spreads for short and long-term sovereign bond maturities (2Y and 10Y) between the United States and Germany. Then, we measured the daily returns in bps. The same calculus was also made for government yield spreads within the euro area, which indicates financial stress in the region. Specifcally, we measured the spreads for Italy, Spain, Portugal and also the aggregate Peripheral group against the Germany 2Y and 10Y bund yields.

ECB's Non-standard Monetary Policy Measures (ECB. NSM) This is a dummy variable which captures the main non-standard monetary policy measures announced by the ECB. They are highlighted in Figure 1 and further information can be found in the ECB's website.

Euro We considered the daily returns of the EUR/USD. Data was obtained in Datastream.

WTI We calculated the daily returns of crude oil price. Data was obtained in Datastream.

STOXX50 We considered the daily returns of the Euro area stock price index. It provides a blue-chip representation of super-sector leaders in the region. The index covers 50 stocks from 9 Eurozone countries: Belgium, Finland, France, Germany, Ireland, Italy, Luxembourg, the Netherlands and Spain. Data was obtained in Datastream.

iTRAXX These are tradable credit default swaps indices. We considered two measures in our analysis.The 'Main', composed of the most liquid 125 CDS referencing European investment grade credits and the 'Crossover', comprises the 75 most liquid sub-investment grade entities. We calculated the daily returns for the indices. Data was obtained in Datastream.

VSTOXX It is the implied volatility of the STOXX50, capturing financial stress. We considered daily changes of the index. Data was obtained in Datastream.

VDAX It is the implied volatility of the German DAX stock market index, which could also be a proxy of financial stress in the Euro area. We considered daily changes of the index. Data was obtained in Datastream.

Citigroup Economic Surprise Indices These are a good proxy of high-frequency macro indicators. These indices track how economic indicators are coming in ahead of or below markets expectations. Data is provided at daily frequency and calculated as the normalized deviation of the actual economic data release from the market consensus prior to the data publication. We took data for the United States and the Euro area from Datastream. We considered daily changes of these indicators.

E Factor picking Lassos

Figure 31: ΔEUR/USD – Pre-crisis Lasso regression

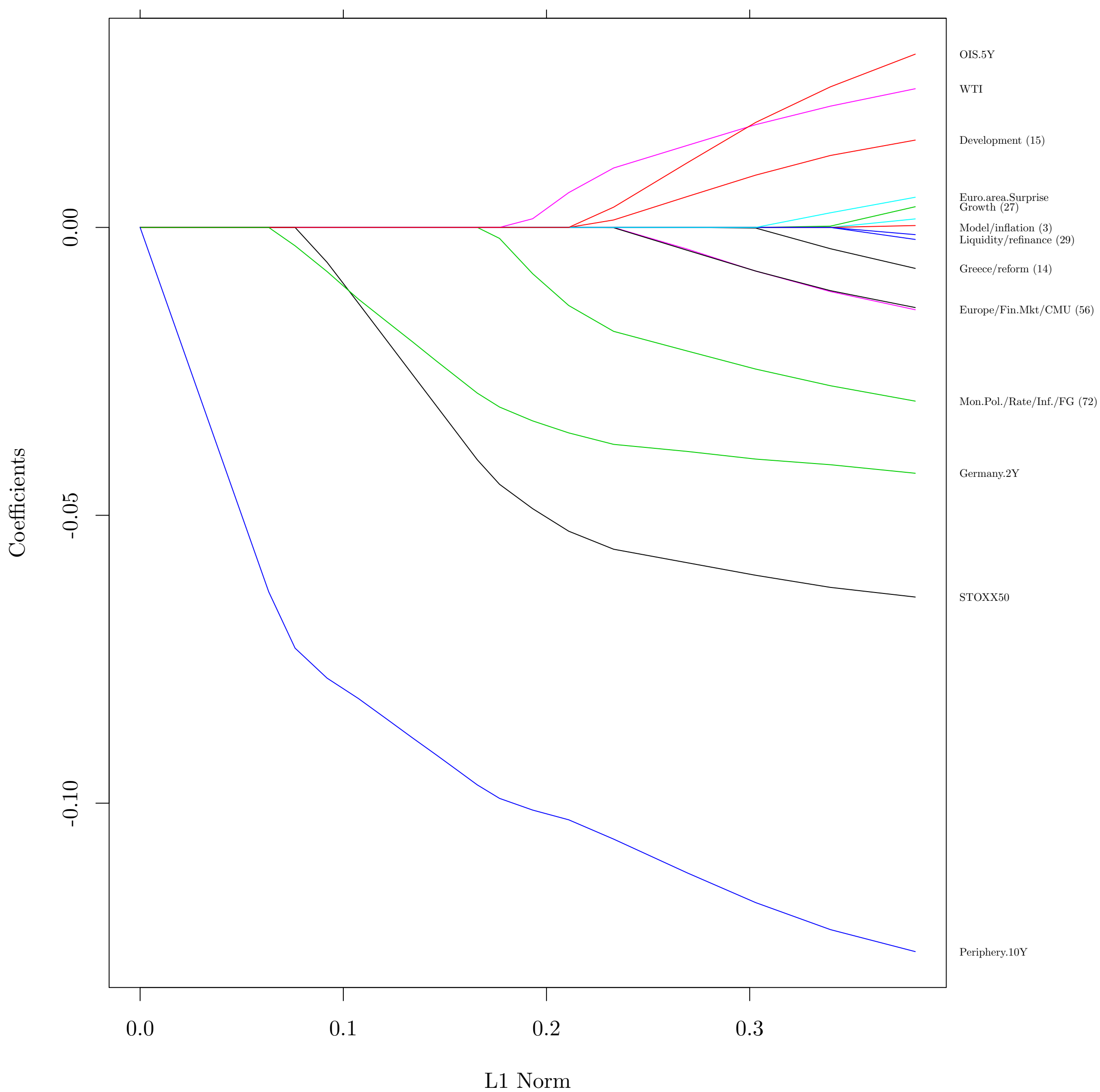
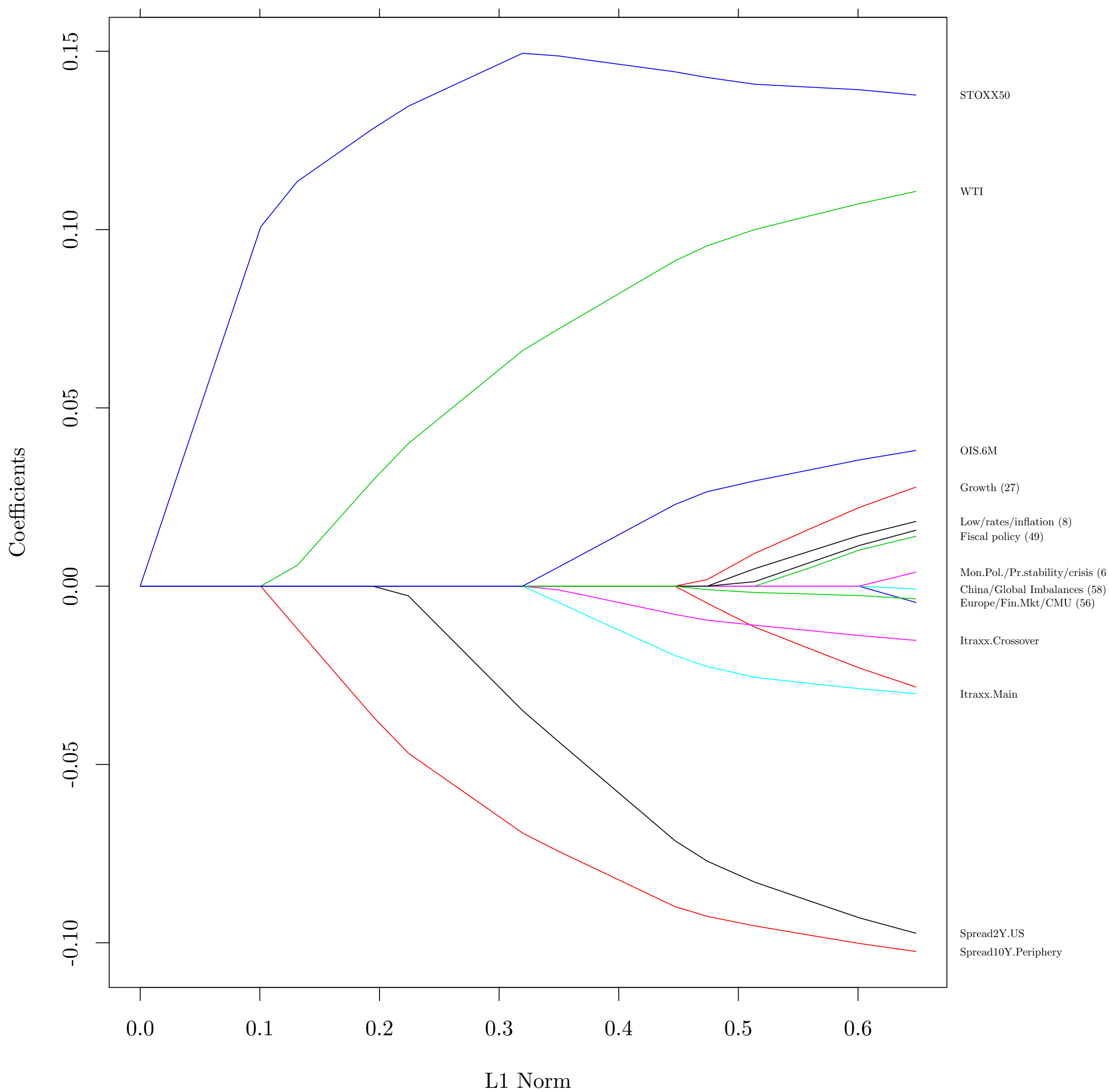


Figure 32: ΔEUR/USD – Crisis Lasso regression



F Complementary materials

Figure 33: Topic correlations

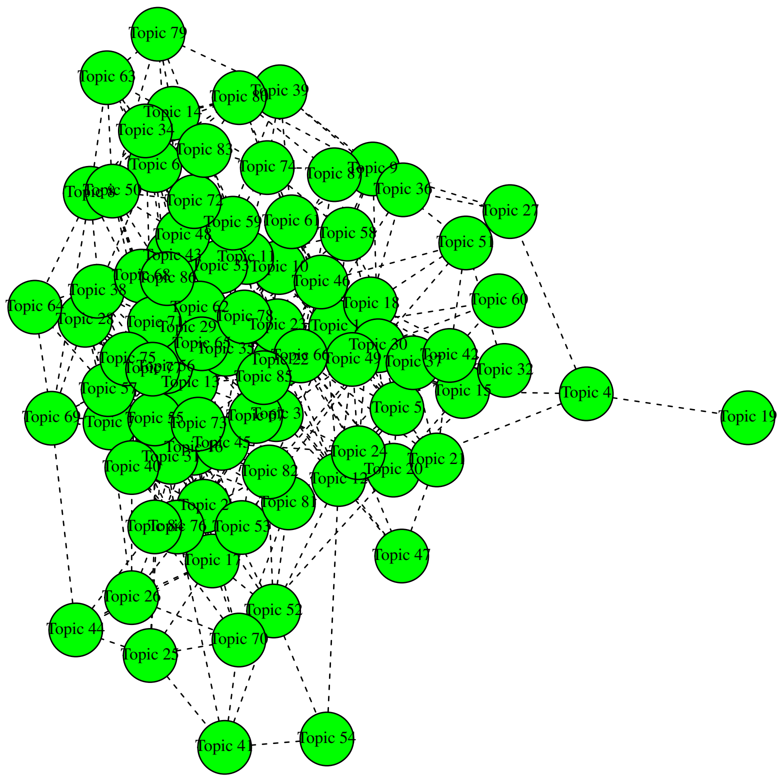


Figure 34: Topic Summary

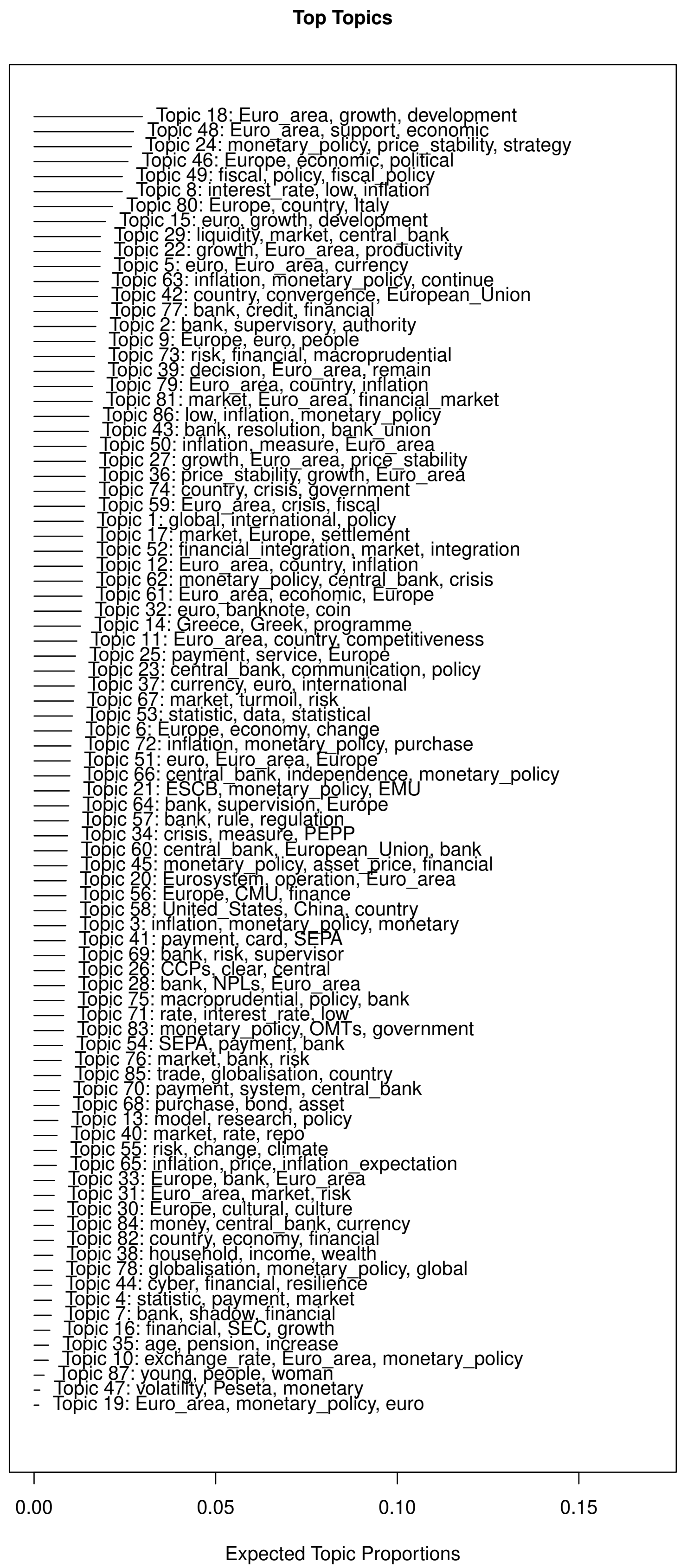
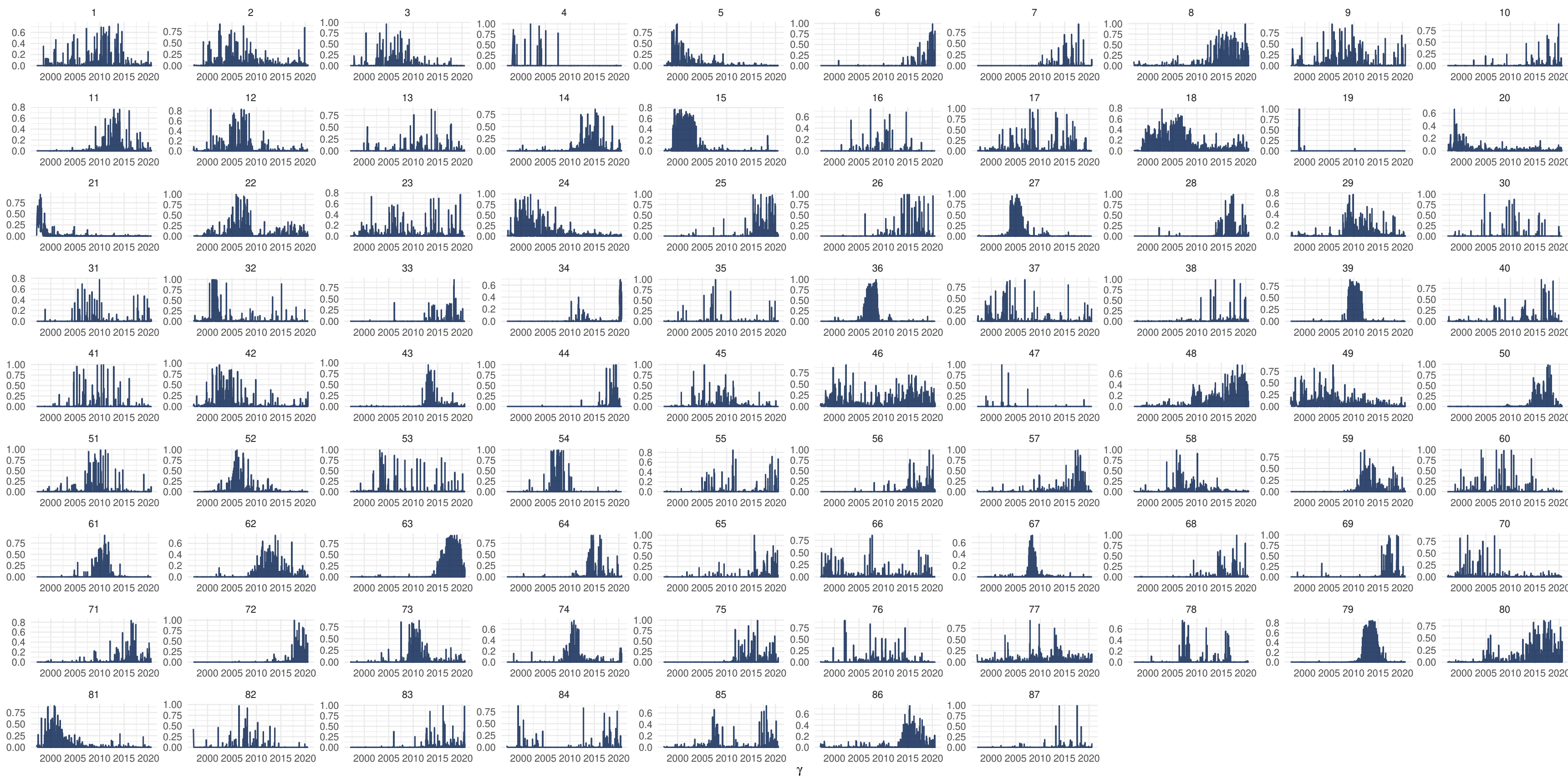


Figure 35: Topics top betas (influential words)

Structural topic model (stm), Spectral initiation (K=0), manual preprocessing

Figure 36: Topics gamma distribution (probability over each speech)



Structural topic model (stm), Spectral initiation (K=0), manual preprocessing