A Work Project, presented as part of the requirements for the Award of a master's degree in Economics from the Nova School of Business and Economics.

CHERRY PICKING WORDS:
A TOPIC MODEL APPLICATION TO ECB SPEECHES

PIETRO FADDA

Work project carried out under the supervision of:

Paulo Manuel Marques Rodrigues

20/05/2022

Abstract

Communication is set to become vital for policy transmission. Growing textual data availability requires the development of more sophisticated tools. This report investigated the empirical application of Latent Dirichlet Allocation (LDA), a Machine Learning Topic Model, with the aim of revealing topics in ECB speeches for the period 1997-2022. Results corroborate comparable studies and produce significant novelties: Topic dynamics captured structural macroeconomic events; ECB speakers differ in topic allocation and frequency coverage; Topic weights correlate with underlying variables and the model well perform on unseen documents. However, outcomes should be treated with caution as LDA lies between correlation and causality.

Keywords: Machine Learning, ECB, Topic modeling, LDA model, Central Bank communication

Introduction

Unravelling how the European Central Bank (ECB) communicates its policies is critical to understand how the central bank operates and is made accountable. Increasing ECB disclosure calls for additional effort to build more complex quantitative and qualitative models to make use of this new data. The research at hand outlines an empirical application of Latent Dirichlet Allocation (LDA), a Machine Learning Topic Model currently attracting considerable interest, aiming to automatically discover topic structure both at the document level as well as the word level, hidden under the ECB speeches' semantic. In addition, the study seeks to shed light not only on what semantically changed over the 1997-2022 period, but also on how different speakers tend to predilect certain topics. With the latter not yet receiving sufficient attention. This work project is structured as follows: *Section 1* introduces a brief recap of the literature concerning the importance of effective central bank communication and a few examples of previous Topic Model applications to the field of monetary economics; *Section 2* describes the theoretical model as well as its time-consuming implementation and settings. In *Section 3* the data goes through exploration, cleaning, and preprocessing. Following, the extensive results obtained in *Section 4* report the findings both for speakers and aggregate level. An inter-topic and topic correlation with macro trends is also investigated. Finally, the estimated model is applied to unseen documents for validation while *Section 5* provides the conclusions.

1. Literature review

This work project touches two main branches of literature. The first one relates to the importance of effective central bank policy communication to the public. The second tries to answer questions related to how LDA models could shed a light on how the ECB's communication has evolved. The research at hand aims to contribute to the latter.

According to Blinder et al (2008) and Vayid (2013), Central Bank communication evolved

over time. For instance, its frequency rose (Rzonca (2018)) and both quality and power of communication also improved (Blot (2018)). The ECB's communications increasingly affected the risk taking in the markets according to Scarpari (2015) and Horvath (2018). In addition, Moraes (2018) emphasized the value connected to disclosure as a macroprudential tool. Angino (2021) highlighted that higher clarity results in higher engagement from mass media and the public.

However, communication suffers some potential limits because of the trade-off between central bank independence and volatility. According to Meltzer (1986), the higher the independence the noisier are the announcements and most of all, "the longer it takes to build up the stock of credibility" especially if the ECB aims to promote European values and trust in European institutions (Assenmacher (2021)). Regarding transparency, Glas (2021) showed that it is steadily increasing. However, improvements are still insufficient to allow a capillary transmission of monetary actions. All studies point to the importance of increasing clarity and transparency as an effective way to manage economic agents' expectations.

Finally, monetary policy also operates through non-monetary news (Cieslak (2019)). For instance, Cieslak (2019) found that in press conferences the non-monetary component accounts for most of the policy debate and Q&A because the language is less technical. As a result, a wide variety of topics which are not strictly related to monetary actions may be covered. Additional information availability requires more complex models for processing it. The ECB itself is already implementing Machine Learning (ML) and Artificial Intelligence (AI) techniques in at least four segments: statistical information; macroeconomic analysis; prediction power and the monitoring of indicators (Kinywamaghana (2021)). Hence, ML and AI may help designing more targeted policies and increase the efficiency of current ones. For textual data mining purposes and topic modeling, one of the most interesting models is the Latent Dirichlet Allocation (LDA) developed by Blei et al. (2003).

According to Isoaho (2021), topic modeling enables scholars to apply policy theories and concepts to much larger sets of data. Saret and Mitra (2016) claimed that market watchers may benefit from LDA models. Other reasons for valuing LDA are that discontinuity in the semantic is associated with rising volatility according to Ehrmann (2020). So, different economic scenarios require different words and narratives.

Carboni et al. (2020) compared ECB to FED speeches over the period 2007-2019 and Luangaram and Wongwachara (2017) analyzed data from around 15 different central banks. More recently, Hansson (2021) focused on the Swedish Central Bank. While Keida (2019), analyzed regular press conference documents of the Bank of Japan, comparing governors' speeches over time.

2. Theory

2.1. Model and notation

Among all probabilistic topic models, the most popular is the LDA. LDA is an unsupervised machine-learning algorithms which seeks to detect the hidden, or "latent" structure behind words. First proposed by Blei et al. (2003), it is based on the idea that a set of documents is a mix of topics in varying proportion and each topic is in turn a mix of words with varying probabilities. In this work project the original Blei et al. (2003), notation is used. Specifically,

- By *word* it is meant the basic unit from a vocabulary indexed by $1, \ldots, \mathrm{V}$. Words are stored as a single component vector *w*, such that for every two words *u* and *v* such that $u \neq v$ then $w^v = 1$ and $w^u = 0$.
- A *document* is a series of *N* words $w = (w_1, w_2, \ldots, w_N)$ with w_n as the n^{th} word.
- A *corpus* is a set of *w* documents denoted by $\mathrm{D} = w_1, w_2, \ldots, w_m$.
- *Topic* refers to the gathering of words that tend to appear in a discussion and more frequently co-occur whenever the topic is discussed.
- α is a *k*-vector with components $\alpha_i > 0$ which represents the document-topic density.

Later, a document's topic distribution is randomly sampled from a Dirichlet distribution with alpha hyperparameter (α). Each document *w* has a multinomial distribution θ^w of independently but not identically distributed topics.

- β (later called η) is the topic-word density and is represented by $\beta_{ij} = p(w_j = 1|z_i = 1)$.
- The topic's word distribution is randomly sampled from a Dirichlet distribution with hyperparameter η (Treude (2018)) . Each topic z has a multinomial distribution ϕ^z of words (Rajani (2014)).

Following Blei et al. (2003), the joint distribution of a topic mixture θ (see *Formula A1* in the *appendix)*, with a set of N topics z, and a set of N words w is given by:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta).$$

From the above, the marginal distribution of a document could be shown to be the integration over θ and sum over z derived as

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \prod_{n=1}^{N} \left(\sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\,\theta.$$

The probability of a corpus is then

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \prod_{n=1}^{N_d} \left(\sum_{z_n} p(z_n|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\,\theta_d$$

Following Blei et al (2012), the LDA algorithm consists then of a pair of steps. The initial phase requires to randomly choose a distribution over topics while the second involves the allocation of each single word in each document to the topic distribution. This ensures that all documents will always share the same number of topics although in different proportions.

2.2. Determining the number of topics.

To train an LDA model a fixed number *k* of topics needs to be provided. However, citing Grimmer (2013), there is no such thing as an ideal *k.* In general, a modest one could result in too general outcomes while a large *k* might lead to unnecessary, disproportionally small and

overlapping topics, although it can help reveal further subtopics. There are programs which learn the technical optimal *k;* however, those results rarely make sense to humans. Navigating and balancing between different techniques is then highly desired. Roberts (2019) warns that the final decision should not be based solely on a single quantitative measure, but be accompanied by human judgment, given the research question under analysis.

A common way is to compute and compare coherence and perplexity scores for different numbers of topics. *Topic Coherence* scores the semantic similarity between high scoring words in the topic. A higher score is preferable, *ceteris paribus* for human topic interpretation purposes. *Perplexity*, instead, captures how the model performs when run on unseen data. It is measured as the normalized log-likelihood of a held-out test set (Zumbach (2021)). A lower value is preferable *ceteris paribus*. However, it may also occur that low perplexity or high coherence scores make human topic interpretation harder. One should weight the two scores and usually be willing to forego some perplexity to enhance coherence and vice versa, depending on whether the resulting model performs better or not. Note that the econometric parsimony principle also applies. A humanely interpretable model with a decent number of topics but not finest perplexity and coherence may be preferable to one with tempting scores but far more topics.

A second approach is checking whether the model gives the desired predictive value.

For instance, the *"pyLDAvis"* Python package is a straightforward tool. What it does is plotting an interactive web-based visualization of each topic semantic distance from each other. For instance, if two topics are overlapping it means that they have similar semantic, but they are not necessarily discussing the same topic. One should try to slightly decrease the topic number and see if the model does in fact combine them or not (note that all topics will change as well). Another, interesting feature is that *pyLDAvis* shows both saliency and relevance of topic specific keywords. Where *saliency* refers to how much the word is

informative of the referred topic (the size of the figure increases in the marginal topic distribution) while *relevance* is a measure of the word importance for the topic compared to the overall topics (estimated term frequency within the selected topic over the overall term frequency). Finally, the *pyLDAvis* allows to tune the "Lambda" parameter with $\lambda \in (0, 1)$. By changing λ is possible to rank and extract topic keywords in such a way as to give higher weight to words which are topic exclusive rather than globally frequent. λ is then entering the following *Relevance Score equation* where the relevance *r* of the word *w* to topic *k* for a provided λ is given by

$$r(w, k|\lambda) = \lambda \, log \, (\phi_{kw}) + (1 - \lambda) \, log \, (\frac{\phi_{kw}}{p_{kw}})$$

Where ϕ_{kw} refers to the probability of word *w* in topic *k* and $\frac{\phi_{kw}}{p_{kw}}$ is the uptake in terms of probability within a topic to its marginal probability across the entire corpus (Ruchirawat 2020). Lowering λ gives higher weight to the second term so highlighting unique words.

However, if the only interest is the relative distance between low dimensional spaces representation it may be advisable to plot a t-distributed Stochastic Neighbor Embedding (t-SNE). Note that the represented distance is not in itself important. The method only allows for a quick direct human assessment of whether topics are clustered (so correlated with lower perplexity).

Another simple method is to plot word clouds for the topic's main keywords. However, its usefulness is limited if the most discussed words are shared by all topics.

Finally, an intuitive way to successfully deduct the topic names is to have a look at the documents where they have been mostly discussed. This procedure is particularly time consuming but extremely effective for topics which are similar.

2.3. Implementation

Nowadays, the state of the art for implementation of LDA models are the Python NLTK and Gensim libraries developed by Řehůřek and Sojka (2011). Although the process is still time consuming and coding intensive, most of the model can be developed by providing the following argument functions:

- The set of documents which we are willing to analyse called *corpus.* Here each word is stored as two numbers, the first identifying the word itself while the second reporting the word frequency in the document
  (i.e., *"major_mileston – road - european_integr – process – peac – stabil..."* is converted into duplets as [(0, 2), (1, 1), (2, 1), (3, 1), (4, 1), (5, 2) . . .]).
- The set of unique words from the corpus-based dictionary, also called *id2word.*
- The number of topics that the model will have to find - *k*.
- The *random state* to ensure replicability as the model employs a random component.
- The number of documents loaded into memory each time, so called *chunk size*.
- The number of training iterations through the entire corpus, so called *passes*.
- α and η hyperparameters, meaning a-priori belief on document-topic and topic-word distribution, respectively. On those, Griffiths (2004) suggests an $\eta = \frac{200}{V}$ and $\alpha = \frac{50}{k}$ where *V* is the vocabulary length and *k* is the number of topics. Conversely, Siklos (2018), criticized Griffiths (2004) stating that for smaller *k* (so not in the order of hundreds but tens) a value of $\alpha = \frac{10}{k}$ is optimal. Fortunately, now an "auto" option is available where α and η are automatically assigned.

2.4. Validation

When the model is sufficiently trained it can be tested on external documents for validation, or the existing corpus can be further investigated. For instance, it will be possible to check

the topic composition of each document (i.e., *Topic 1:0.4, Topic 2:0.35, Topic 3:0.25*). Additionally, the model settings allow only topics showing an arbitrary minimum threshold to be displayed. Words below that level will be assigned to higher weighting topics in the document depending on whether the word is common to them or not. Finally, it is possible to extract each word probability, which is useful for visualization purposes.

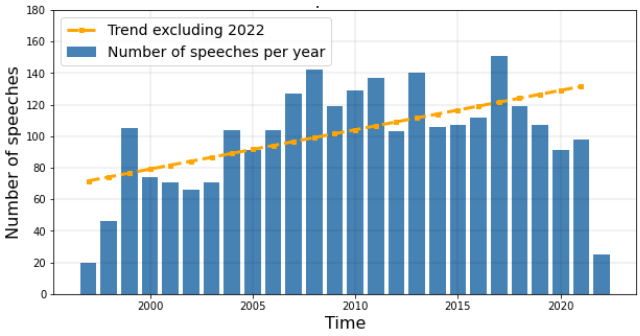
3. Data

3.1 Data exploration

The final corpus used in this work project consists of 2331 retrieved from ECB (2019). The dataset is updated monthly by the institution. Besides the speeches' scripts, it provides dates, locations, speakers, and event descriptions. The period considered is from Feb-1997 (first available observation) to Mar-2022. It originally totaled 2565 speeches. The number was later reduced by 80 documents which showed external slides' references and by 154 non-English speeches. References, and notes were also removed.

Figure 1 shows that the yearly number of speeches has been steadily increasing.

Figure 1: Number of speeches 1997-2022



Source: own elaborations with Python software

The set of speakers includes the President of the ECB, the President of the European Monetary Institute, the Vice-President, the remaining Members of the Executive Board and, finally, the Vice-Chair of the Supervisory Board. The former two as well as the latter two were treated similarly since they derive from the same environment.

Figures 2a and *2b* show speaker activity and relative weight over the years, respectively.

Figure 2a: Speakers over time

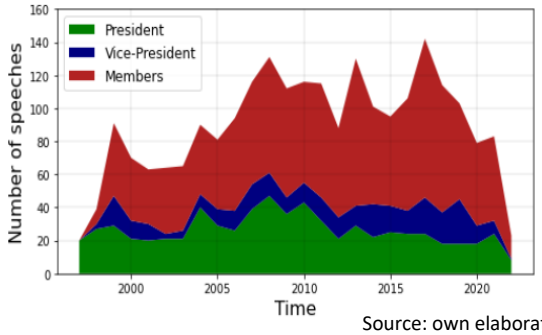
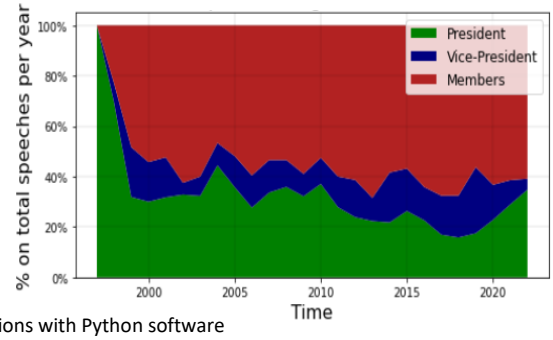


Figure 2b: Speaker weight over time



Source: own elaborations with Python software

On average, the President accounts for around 30% of total speeches while the Vice President and Members of the Executive Board to 13% and 57%, respectively. The President was more active during the Trichet term, but its relative weight has been gradually decreasing. Only recently does the trend seems to have inverted, a period which correlates with the beginning of the Lagarde term and the pandemic. The Vice-president's frequency has been constant till the Great Financial Crisis and slight increased since then. The leading speakers both in terms of frequency and weight are the Members of the Executive Board. It should not come as surprise that these are more numerous. However, speech frequency tells only part of the story since the documents have no fixed length and can change according to the speaker.

That is why also word distribution and word count matter. *Figure 3a* reveals that the President has the highest density at around 2000 words while the Vice-President exhibits a

Figure 3a: Word distribution by speaker

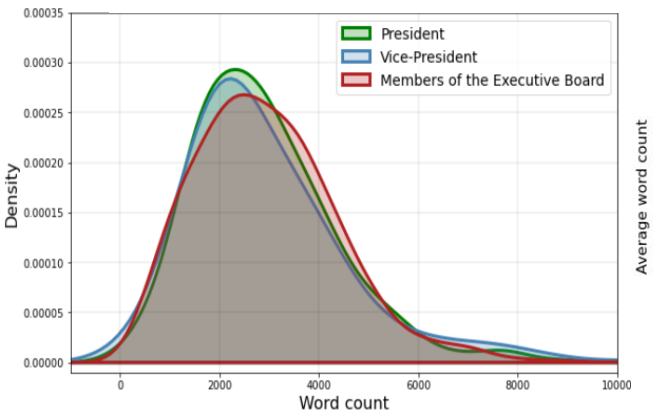
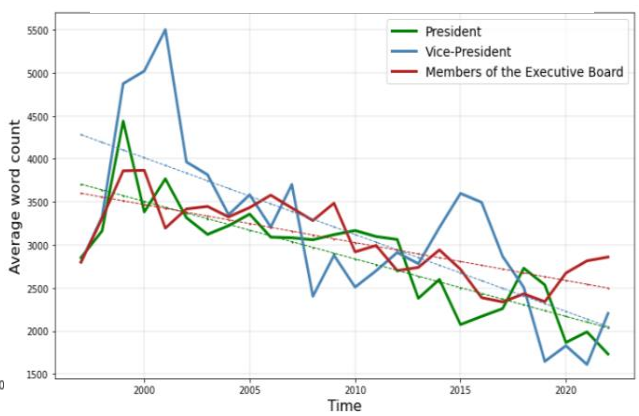


Figure 3b: Word count dynamic by speaker



Source: own elaborations with Python software

more pronounced fat right tail, meaning that there is a higher probability of longer speeches. Finally, the average length of documents shown in *Figure 3b*, seems to follow a downward trend with President and Vice-President declining at higher speed than Members of the Executive Board.

3.2. Data cleaning and preprocessing

Seven different filtering and transformation methods are applied to the original data (total words reduced by 83% and unique words by 77%). The procedure was particularly time-consuming and involved, in order, the removal of punctuation and words below 4-characters in length, the conversion to lower case, and the exclusion of stop and linking words (268 words).

Subsequently, only nouns were kept, so that verbs, adjectives, and adverbs were discarded. Nouns are more likely to expose and cluster topics and to facilitate the process of topic identification. Next, both *lemmatization* and *stemming* were applied. The first removed inflectional endings from the word and returns its dictionary form. For instance, it turned plurals into singulars; past, future tenses, and third person of the present tense to infinitive; superlatives and comparatives to basic adjectives. Stemming, instead, returned the morphological root of the word, so the irreducible segment. The rationale for applying them is that the LDA model will otherwise treat words with the same root as different so underestimating their importance and exposing the model results to bias. It may be argued that stemming the words alone would reach the same outcomes. In practice many technical words are less likely to show up if not previously lemmatized. In addition, stemming alone cannot relate words which have different forms based on grammatical constructs (i.e., good, better, and best). After the word transformation, bigrams and trigrams were generated since many words are likely to co-occur in the same order in different docs. What *n-grams* do is to bring together those words and treat them as one single entity. This is done through setting a

minimum frequency of occurrence and *threshold score*. Which is how much more likely are the words to co-occur than if they were independent. This is sensitive to unique combinations, so both the parameters should be carefully tuned to ensure relevance. This is a valuable step as it allows for technical word combinations to be highlighted. Overall, bigrams and trigrams accounted for 20% of total words and 45% of unique words at this stage. Finally, extremes are filtered. First words occurring in at least 50% of the documents are removed, second those that appear in less than 10 speeches are also removed. *Table 1* below shows the word count and number of unique words after every filtering step.

Table 1: Word count at each filtering step

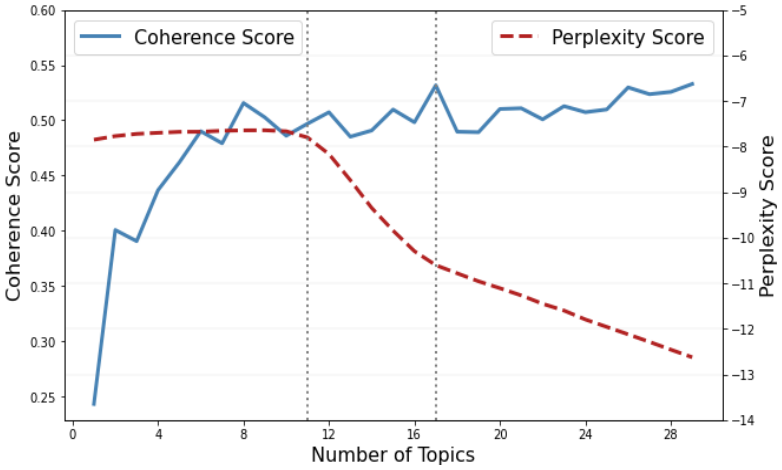
Words	Lower case	Small words	Stop and Linking words	Nouns	Lemmatization Stemming	Bigrams Trigrams	Filter Extreme
Count	6.396.247	4.092.164	3.156.109	2.206.662	2.192.729	1.789.607	1.091.048
Unique	29.963	28.657	28.389	11.962	7.586	13.838	7000

4. Results

4.1. Picked number of topics and settings

Figure 4 shows the Coherence and Perplexity Scores estimated for up to 30 topics.

Figure 4: Coherence and Perplexity scores



Source: own elaborations with Python software

The range between the 11th and 17th topics is of special interest. At $k<11$, the perplexity score is constant and maximum while at $k>17$ coherence score temporarily falls, and it takes a few

more iterations to reach a new high. However, the additional topics do not justify the small improvement. Another reason for investigating such range is that perplexity first accelerates and then suddenly slows down, creating two visible kinks in the graph.

A final decision is made, to set k=14 with a resulting coherence and perplexity score of 0.5098 and -9.8448 respectively *(*see *Table A1* in the *appendix).* In addition, pyLDAvis inter-topic distance map and t-SNE *(*see *Figures A2* and *A3* in the *appendix),* show both visible topics and clusters. In a similar study conducted by Luangaram and Wongwachara (2017) on 22 separate central banks, the number of topics was comparable (k=15 in that case).

However, many authors also have opted for k=10. For instance, Carboni et al. (2020) in a comparison between ECB and FED's Governor speeches, Küsters (2022) on ECB speeches and Greene and Cross (2015) on European parliament dialogues. The ECB itself found 58 sub-topics later clustered into 10 (Hartmann (2018)). Finally, Hansson (2021) found 12 local and 17 global topics in Swedish Central Bank speeches. However, in the latter case the author employed a Dynamic Topic Model (DTM), which differs from the LDA since it allows topic keywords to mutate over time.

The final LDA model was estimated with the parameter settings displayed in *Table 2 (*see *Table A2* in the *appendix* for the estimated *α).*

Table 2: Model settings

Chunk-size	Passes	α	η	Random seed	k
50	5	"auto"	"auto"	50	14

4.2. Naming topics and main keywords

To name topics it was necessary to balance keywords extracted at distinct λ values (λ of $0, 0.6,$ and 1) together with word-clouds (although the latter were only partially informative *(*see *Figure A4* in the *appendix)). Table 3* provides a selection of ten selected keywords for each topic.

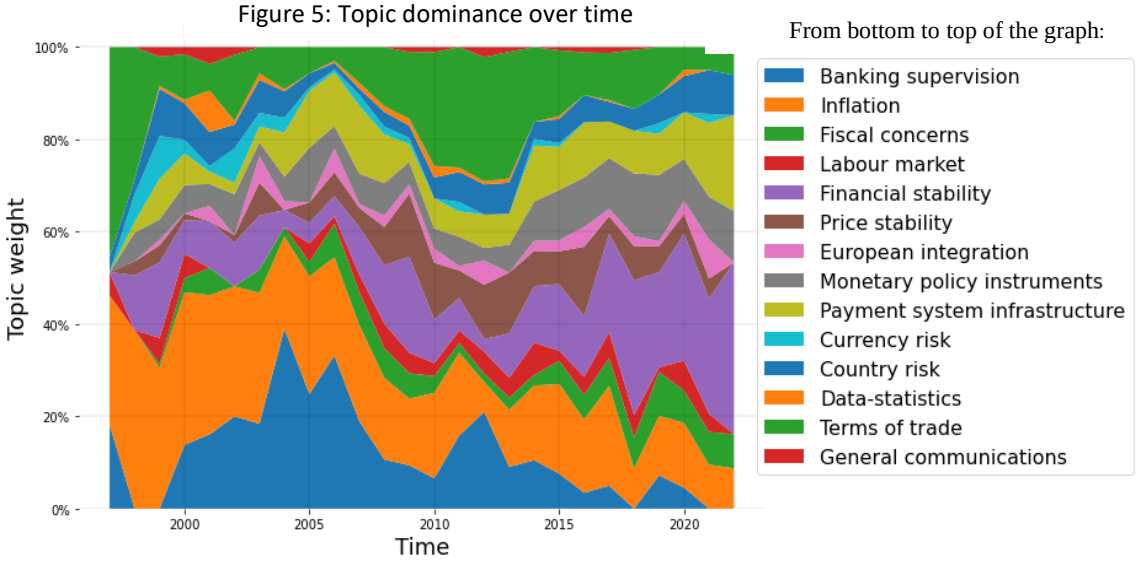
Table 3: Ten main keywords for each topic

Banking supervision	**Inflation**	**Fiscal concerns**	**Labour market**	**Financial stability**
bank_supervis supervisor risk_board sanction basel principl prevent rule regul requir	inflat_target inflat_expect inflat_forecast intermedi_target deviat_target shock expect central_bank_target central_bank_monitor asset_price_bubbl	asset leverage solvenc sovereign_debt lender_resort valuat real_estat buffer default deposit	wage labour_market_reform product_growth labour_forc labour_market unemploy_rate labour_suppli unit_labour_cost human_capit employ_growth	bank_sector capit_market risk_manag exposur disclosur vulner capit_requir perform_loan leverage_ratio contagion
Price stability	**European integration**	**Monetary policy instruments**	**Payment system infrastructure**	**Currency risk**
forward_guidanc inflat_outlook price_pressur headlin_inflat medium_term_inflat longer_term_oper inflat_dynam polici_accommodation excess-reserv rate_deposit_facility	climat_chang singl_market trust single_currenc european_integr european_citizen peac memori solidar closer_union	asset_purchas bond_issuanc money_market lend_rate purchas_program risk_free collater_framework swap_rate matur deposit_facil_rate	payment central_bank_money payment_system payment_service retail_payment payment_solut payment_solut digitalis user privaci	Risk_share Spillov Neg_shock Current_account_deficit Exchange_rate Cross_border_capit_flow Devalue Interest- _rate_differenti Real_converg Real_incom
Country risk	**Data-statistics**	**Terms of trade**	**General communications**	
exit_strategi prejudice_price_stabil interest_rate_reduct repair consum_good germani greek problem talk damag	data climate_risk statist action_plan methodology databas data_gap data_collect inform report	trade globalis manufactur valu_chain intern_trade intern_currenc world_trade capit_flow asian_economi currenc	gener reaction_function euro_banknot central-bank- commun commun democracy woman introduct_banknot career member_govern_council bridg	

4.3. Topics distribution over time

Figure 5 illustrates the share of speeches delivered having a specific dominant topic.

Figure 5: Topic dominance over time



Source: own elaborations with Python software

Except for less weighting topics such as "Fiscal concerns," "Labour market" and "European integration," topic proportions changed over the period.

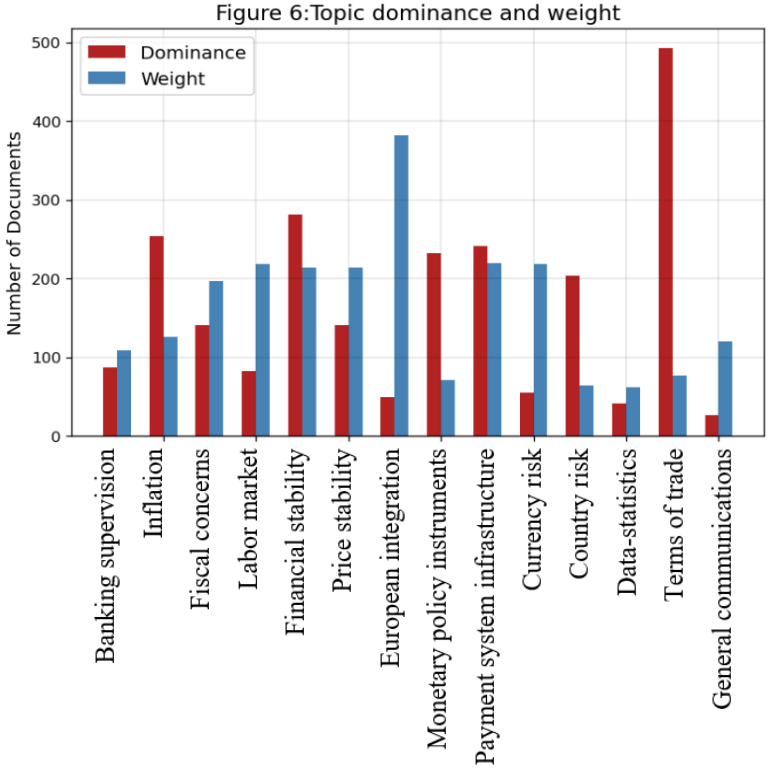
For instance, "Financial stability" and "Payment system infrastructure" appear to be increasingly discussed in recent years, a trend possibly amplified by the recent pandemic as digitalization is one of the most frequent topic words. It is not the case for "Terms of trade," which covers economic indicators and was largely covered during the Great Financial Crisis and the Sovereign Debt Crisis. Also worth mentioning are the spikes observable for "Banking supervision." They highly correlate with the final stages and implementation of Basel II, a word stressed in numerous occasions at the time. Hansson (2021) provided more details on the dynamics of the word. The rise since 2012 of "Monetary policy instrument" correlates with the time of the Quantitative Easing program implementation. However, for all topics, only the "what is changed" is observed (correlation) while the "how" and "why" (causality) cannot be inferred from LDA models (Matsui (2021)). Note that the persistence of a topic may still signal the desire of the ECB to decrease the uncertainty surrounding ECB communications with the aim of reducing market volatility specially in the sovereign-bond

market. For instance, Ehrmann (2020) observed that central banks frequently make slight changes to previous speeches and that sizable variations after similar reports result into higher market volatility. The effect is amplified since markets adjust to sensitivity (Ehrmann (2020)). It is not surprising that this autoregressive behavior captured the interest of macro-econometric scholars. For instance, Luangaram and Wongwachara (2017) used the communication indicator as an endogenous variable to construct a VAR to model monetary policy transmission. The power of their findings is however, weakened by the debate around the optimal Cholesky ordering.

4.4. Topic dominance and weight

Figure 6 compares *topic dominance* with *topic weight*.

Figure 6:Topic dominance and weight



Source: own elaborations with Python software.

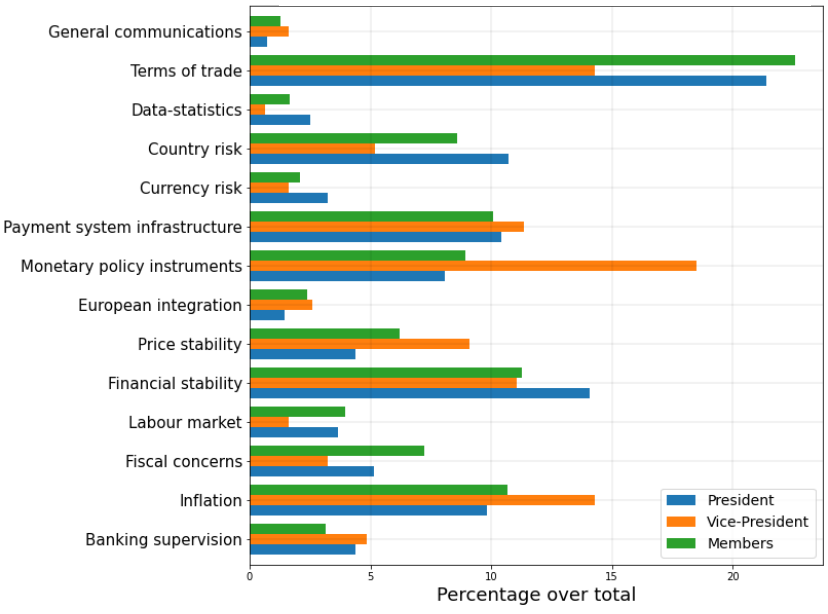
The difference between dominance and weight is that the former refers to the most weighting topic in the document while the latter is the sum of the actual weight contribution of each topic to the corresponding documents. Whenever topic dominance is higher than topic weight it means that such a topic, when discussed tends to dominate the document. Vice versa, when

the weight is higher than dominance, it implies that the topic is more likely to appear in combinations with others and play then a marginal role. “Inflation,” “Financial stability,” “Monetary policy instruments,” “Country risk” and “Terms of trade” tend to be the influential topics, with the latter dominating around five hundred documents or 20% of total speeches. Not surprisingly, “General communications” is the weakest topic. Its poor content mostly involves inaugurations and anniversaries (i.e., the issuance of new banknotes). The residual topics show less power but superior persistence. Among them, “European integration” is the most weighting topic of all, followed by “Labour market” and “Currency risk.” It makes economic sense since every speech provides some macroeconomic context at some point to prepare the audience for more technical matters.

4.5. Topic allocation by speaker

The following sections provides a comprehensive analysis of the different speakers. Something which finds weak evidence in the literature. For instance, Küsters (2022) compared different generations of ECB Executive Board Members while Siklos (2018) aimed at capturing Fed-official individual changes. *Figure 7* shows the relative topic allocation for each speaker, indicating which role is more committed to a certain topic.

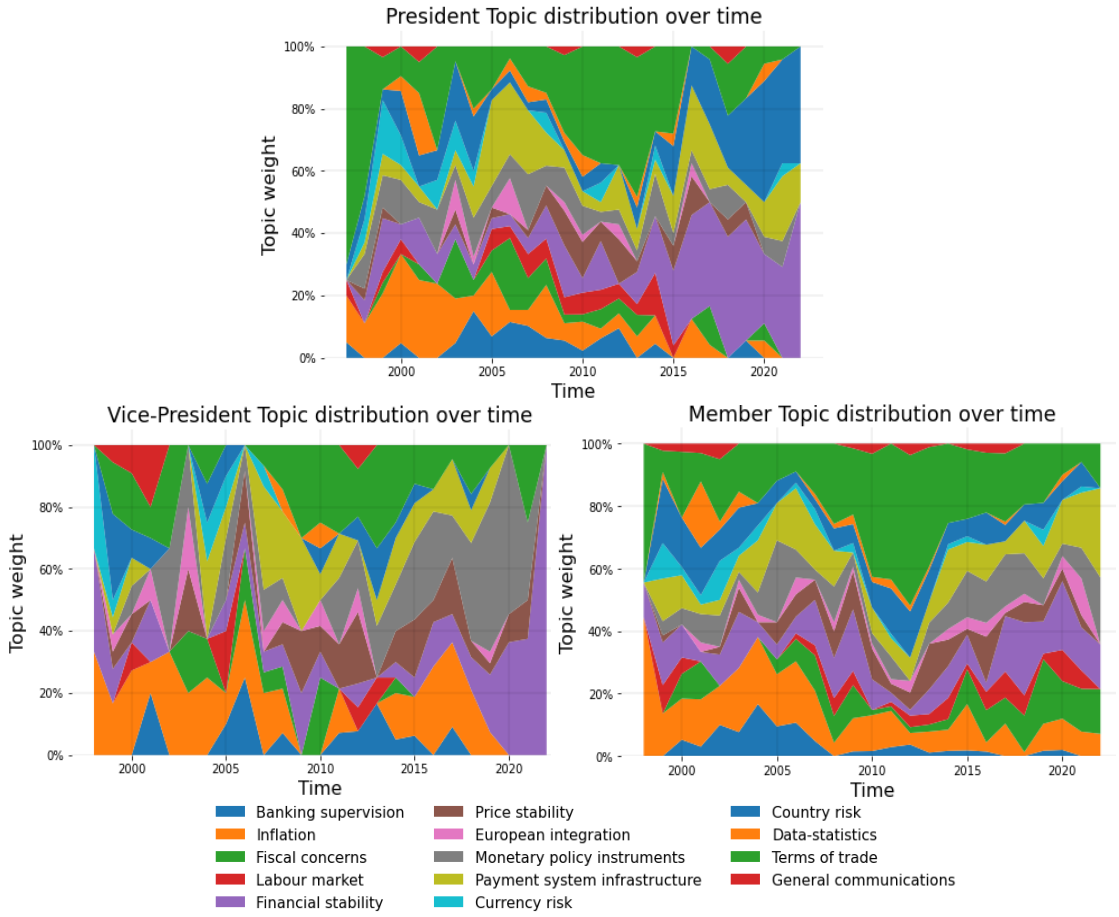
Figure 7: Topic allocation and weight by



Source: own elaborations with Python software.

Overall, the Vice-president (VP) of the ECB stands out in discussing "Monetary policy instruments" and "Inflation" which combined are around a third of the VP corpus content. It might be that the VP simply reaffirms what the President reported in other statements not covered by the corpus. The President, in fact, seems mostly concerned with "Financial stability," "Country risk" and "Currency risk." Members of the Executive Board appear to always lie between the President and Vice-president. There are only two exceptions. The first is the "Fiscal concerns" topic. Perhaps because more speakers could provide a better and persistence coverage of fiscal developments at member state level than what a single institutional figure is able or willing to. The second is "Terms of trade" which is also the most discussed topic by both President and Members. *Figure 8* shows the normalized topic distribution over time for each speaker, (an extension of *Figure 7).*

Figure 8: Topic allocation over time by speaker



Source: own elaborations with Python software

Legend from bottom to top of the graph.

First, the President narrowed its topics in the last decade. “Financial Stability” and “Country risk” confirm the trend anticipated in *Figure 7*. Also “Payment system infrastructure” is increasingly popular for both President and Members. A boost came undoubtedly from the gradual dismissal of paper money. Note that this topic should be understood in a broader manner to also include the debate around privacy, although it is not of special interest for the ECB. Interestingly “Inflation” and “Price stability” are only marginally covered by the president while persistently present in Members of the Executive Board’s speeches. Again, it may be that such topics are already largely discussed in other technical reports, such as policy decision speeches. The only topic predominantly covered by all speakers is “Terms of trade”. Both supply and demand shocks which directly affect the economy do impact trade as well and so are, not surprisingly, watched closely by central bankers.

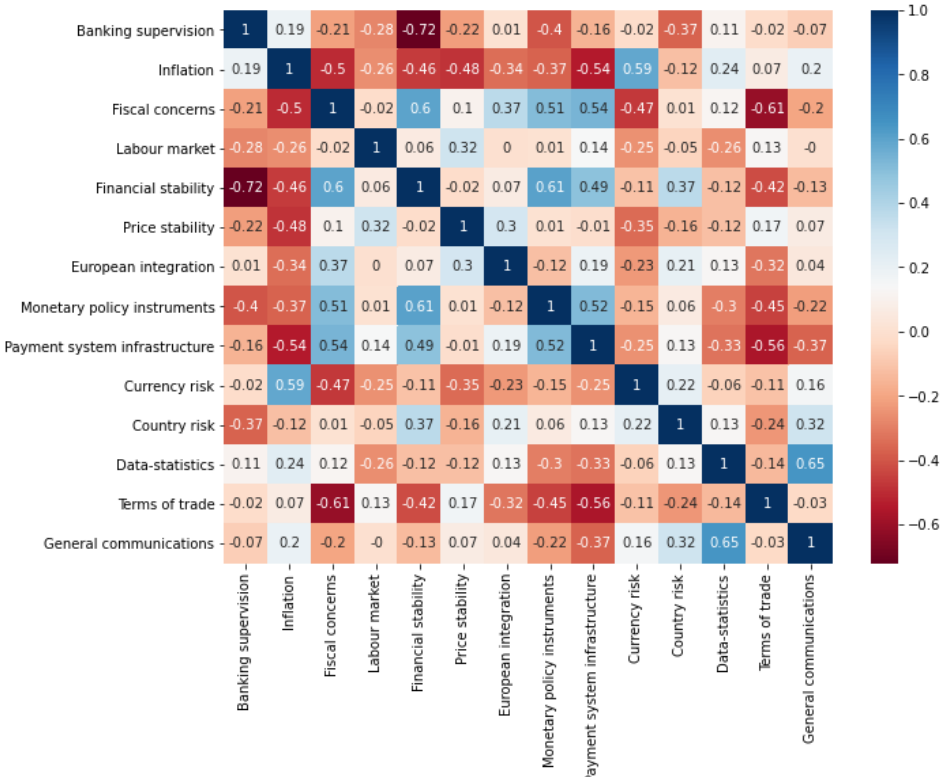
“Fiscal concerns” is covered mainly by the Members of the Executive Board, something already deducted from *Figure 7*. Finally, “Banking supervision” coverage is highly dependent on innovations in the field, both in terms of new regulations and new macroprudential policies put into place. Another reason for its dormancy might be the complexity of the topic which is of less interest to the public since it is targeting private financial institutions to which the ECB already communicates through other channels.

4.6. Inter-topic correlation

Some topics may have higher probability of being simultaneously covered than others. For instance, a weak or null correlation with all other topics may indicate that such a topic is either more likely to appear regardless of the existence of other themes, or that they tend to dominate. The former case is represented by “European integration” in *Figure 9* while the latter by “Labour market”. This Figure also shows that “Inflation,” “Fiscal concerns” and “Financial stability” tend to be linked to all topics. The strongest negative correlation is represented by “Banking supervision” and “Financial stability” (a value of -0.72). According

to Ferri et al (2018) LDA offers a good broad picture, but it fails to capture the underlying relationship between the defined topics. As a result, it is impossible to simply compartmentalize financial system stability as the two topics are intricately interdependent. For instance, Hansson (2021) shows that when the word "Basel" starts to decrease its frequency, the word "regulation" takes its place (the word "regul" appears in both topics).

Figure 9: Topic correlation

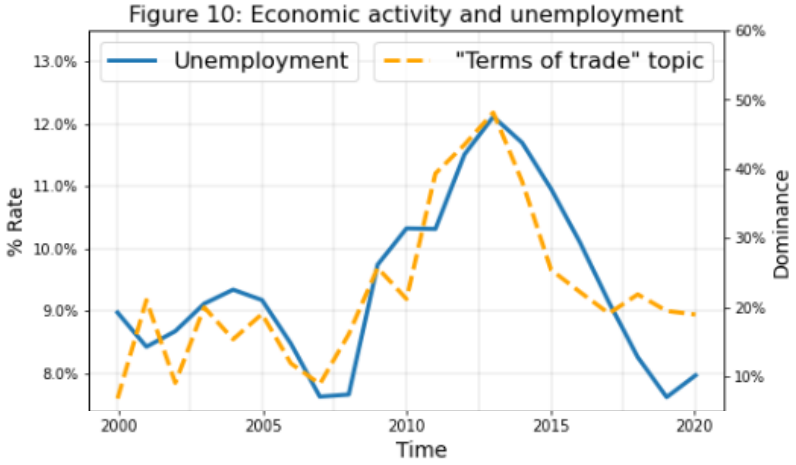


Source: own elaborations with Python software.

4.7. Topics and macro variables

As a last step, topic dominance probabilities are compared with the underlying macroeconomic variables. In other words, an increasing space devoted to a given topic may (hypothetically) suggest that the event is of ECB interest. Hansson (2021) tried to insert a set of controls to identify the shock. However, his model just converges to the *story-based* assumption that central banks narrative follows the occurring of events (Hansson (2021)). Nevertheless, the following analysis is relevant for the development of the field since LDA lies in the middle of simplistic correlation and causality as it provides more insights than mere word frequency (such as Google Trends).

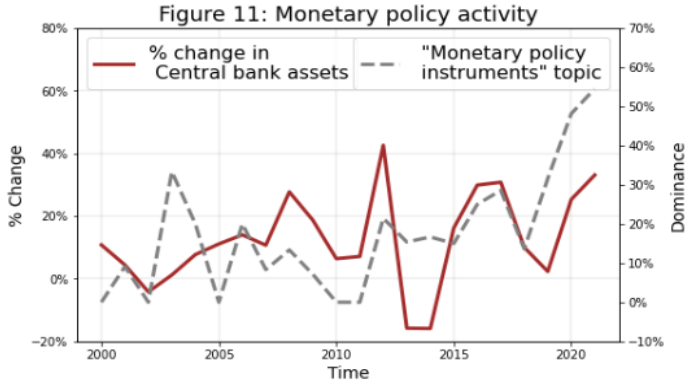
Figure 10 shows that the Members topic "Terms of trade" is highly positively correlated (0.804) to total rate of unemployment in the Euro area between 2000 and 2020.



Source: Own elaborations with Python software. Unemployment data retrieved from Federal Reserve Economic Data, time series code: LRHUTTTTEZM156S

Although it is not possible to infer causality, it makes economic sense that when supply and demand shocks occur, those events are largely discussed, being full employment also a focus of the ECB when consistent with price stability.

Figure 11 compares the Vice-president's "Monetary policy instruments" topic with actual changes in ECB assets. The intuition is that an expanding or shrinking balance sheet could be perceived as a direct consequence of monetary policy decisions which are referred to in the corpus at some point.

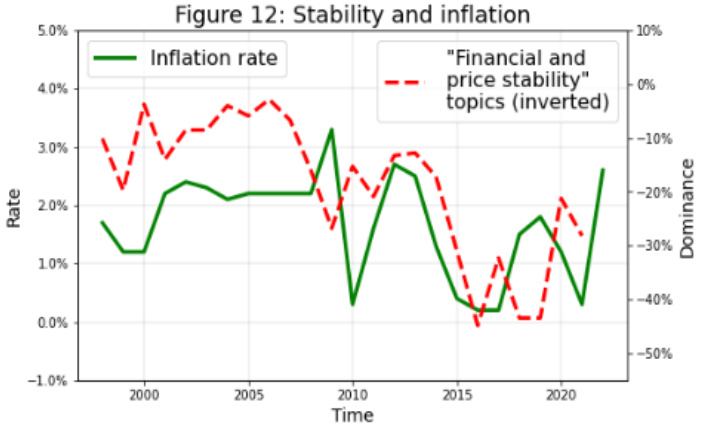


Source: Own elaborations with Python software. ECB Asset data retrieved from Federal Reserve Economic Data, time series code: ECBASSETSW_PCH

The correlation results are much weaker than in *Figure 10* (around 0.4). Still, balance-sheet dynamics does better reflect monetary policy action than interest rate changes due to the reaching of the zero-lower bound. In fact, the two variables are highly correlated when unconventional monetary policy action is in place.

Finally, *Figure 12* compares actual inflation dynamics with the President relative dominance sum of "Financial stability" and "Price stability" topics (the time series has been inverted for visualization purposes).

Figure 12: Stability and inflation



Source: Own elaborations with Python software. Inflation data retrieved from Federal Reserve Economic Data, time series code: CP0000EZ19M086NEST_PC1

Correlation is weakly negative at -0.46, meaning that when inflation decreases the space it takes in the ECB narrative increases. However, the interpretation of the sign could be misleading. The reason is that deflation was more a concern than high inflation in the period of interest. So, this Figure could wrongly suggest that when high inflation comes-in the ECB will talk less about it. Also interesting is that the topic seams to predict actual changes in inflation rate. However, further investigation of this intuition goes beyond the scope of this work project.

4.8. Model validation

The model is evaluated on the existing corpus and external new scripts. New documents should go through the same cleaning and preprocessing used for training the model.

Indeed, the model returned results which are consistent and correlated to the actual document contents and underlying narrative. *Figure 13* shows an example of how the model performed on a recent speech delivered by Fabio Panetta, member of the Executive Board, on April 25th, 2022, titled *"For a few cryptos more: the Wild West of crypto finance"* (ECB (2019)).

Figure 13: Word coloring and topic proportion of an external document

Source: own elaborations with Python software. Data retrieved from (ECB 2019)

Topics "Fiscal concerns" and "Price stability" weighted for a third of the document each. Words belonging to the former were for instance *"fear," "retail_payment," "protect_retail," "faith"* and *"protect."* They may reflect the concerns highlighted by Panetta in the actual speech regarding the criminal usage of cryptocurrencies and widespread frauds in retail anonymous accounts connected to cyber-attacks. In addition, words such as *"deposit," "wealth loss," "asset claim," "money free"* point to the fact that it is more difficult to maintain and monitor price stability if the traditional channels through which monetary policy operates decentralizes. Those worries are also extended to the general status of the economy through the "Terms of trade" topic which highlighted words such as *"lack public_sector rule," "research complex," "central data."* Calling for stricter regulations from authorities - *"commission," "European_parliament council."* Finally, the topic "Currency risk" weighted 17.5%. Words connected to this topic were mostly country names such as "*russia," "korea," "indonesia."* For instance, *"Russia"* was widely mentioned in the actual speech because of the attempt of Russian retailers to evade the depreciating value of the national currency following the invasion of Ukraine. Terms such as *"trade protection," "cross border," "spill"* and *"protection"* also support this argument.

5. Conclusions

The aim of the study was to understand ECB speech narrative by reducing text dimensions through the employment of an LDA Topic Model. The estimated fourteen topics are consistent both in number and substance to the ones found in previous studies.

Taken together, the results confirm that non-monetary topics are, as expected, largely discussed, with substantial space devoted to the status of the economy.

In addition, topic proportion over time is highly correlated to economic events such as the Great Recession, monetary policy implementation and regulation developments (also expected). An innovative investigation on the various speakers highlighted that President, Vice-President and Members of the Executive Board differ in both topic proportions and frequency.

The evidence from this study indicates that, for instance, narrative of Members is more stable and mostly covering the state of the economy, "Labour market" and "Fiscal concerns" at Member State level. The President favors "Financial stability", "Country risk" and "Currency risk" topics compared to other speakers. With the former two taking a larger proportion since the implementation of the Quantitative Easing. Finally, the Vice-President mainly covers monetary topics as "Monetary policy instruments" and "Inflation". Considerable progress has also been made in understanding the cooccurrence of some topics. For instance, "European integration" acts as a background topic, being the one showing the highest weight.

Finally, insights have been gained when relating topic dominance over time with the underlying macro variables. Although, correlation is significant, further research is needed to access its actual prediction power, especially whether topic coverage is anticipating (ex-ante) or reporting (ex-post) economic events. In terms of model validation, the trained model appears to be a powerful tool in evaluating external documents. Both in terms of topic identification and actual keywords. Finally, the model flexibility allows it to incorporate

additional documents, likely further improving model performance.

However, the results are potentially limited by the nature of the model.

First, the corpus may be subject to bias depending on the period and degree of filtering applied since it is subjectively chosen and there are still few rules of thumb and guidelines to follow. Indeed, the findings might not be generalized. These sources of bias were minimized since the period of interest was purposely chosen to provide narrative representative of different macroeconomic conditions: from supply and demand shocks, financial crisis, sovereign debt crisis, Brexit, Covid crisis and geopolitical tensions.

Another strong limitation of the model is that frequency of coverage although reasonably connected to macro events cannot imply causality as slight changes in the model settings can result in different outcomes and not necessarily the same themes. Overall, the research has raised many questions in need of further investigations, as enhancing the quality and power of the results.

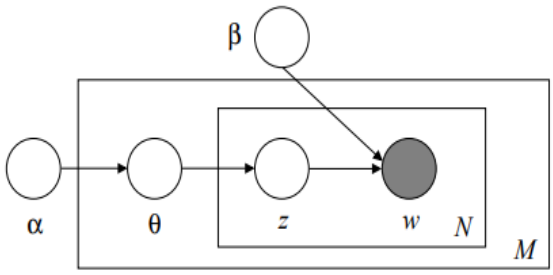
Appendix

Formula A1: Theta

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1},$$

Source: retrieved from Blei et al. (2003)

Figure A1: Graphical representation of the LDA model



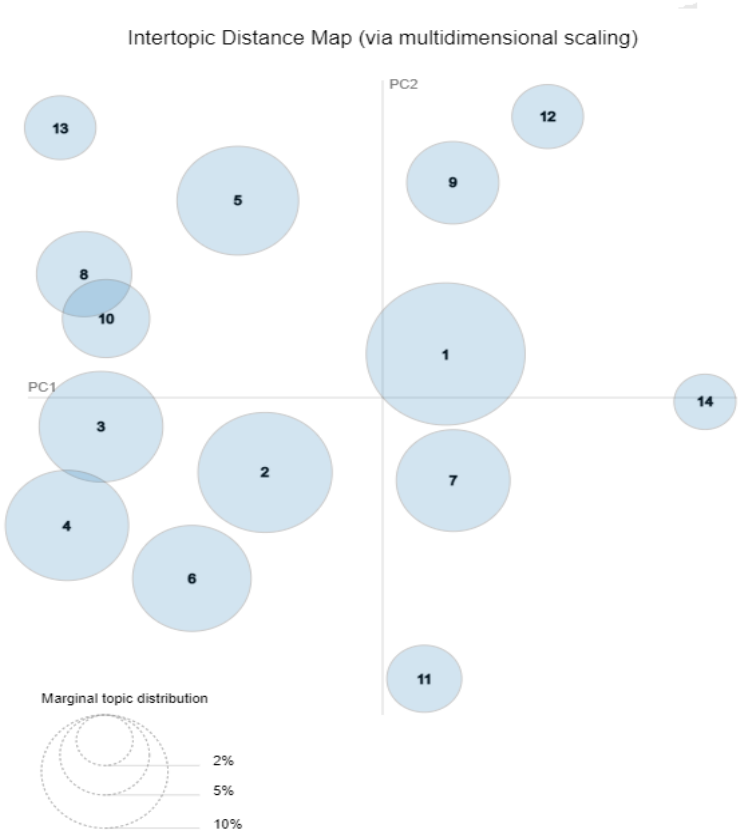
Source: retrieved from Blei et al (2003) (figure 5)

Table A1: Coherence and perplexity scores

TOPIC NUMBER	COHERENCE SCORE	PERPLEXITY SCORE
1	0.2432	-7.8498
2	0.4007	-7.7697
3	0.3906	-7.7215
4	0.4368	-7.6983
5	0.4620	-7.6772
6	0.4901	-7.6751
7	0.4792	-7.6537
8	0.5157	-7.6442
9	0.5027	-7.6419
10	0.4860	-7.6645
11	0.4970	-7.8009
12	0.5073	-8.1648
13	0.4852	-8.7376
14	0.4908	-9.3470
15	0.5098	-9.8448
16	0.4982	-10.2958
17	0.5318	-10.6032
18	0.4897	-10.7813
19	0.4893	-10.9570
20	0.5103	-11.1086
21	0.5109	-11.2672
22	0.5009	-11.4500
23	0.5128	-11.5988
24	0.5074	-11.7952
25	0.5099	-11.9577
26	0.5298	-12.1209
27	0.5237	-12.2866
28	0.5257	-12.4574
29	0.5329	-12.6245

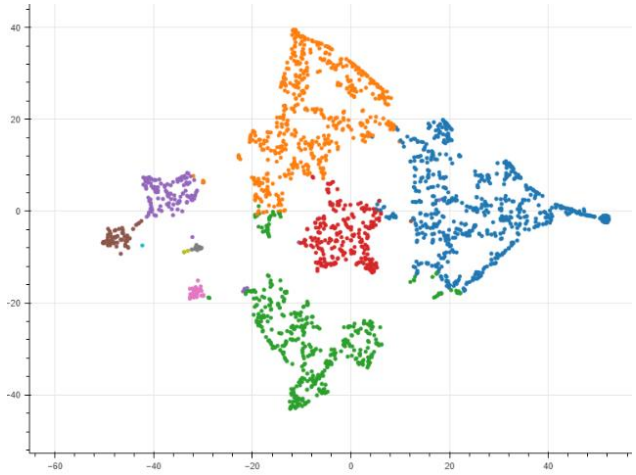
Source: own elaborations with Python software

Figure A2: pyLDAvis plot



Source: own elaborations with Python software

Figure A3: t-distributed Stochastic Neighbor Embedding



Source: own elaborations with Python software

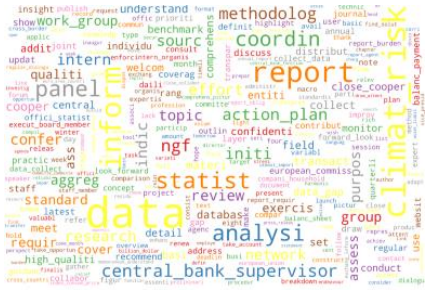
Figure A4: Word clouds

Currency risk

Banking supervision

Country risk

Data and statistics



Financial stability

Payment system infrastructure

Terms of trade

European integration

Monetary policy instruments

General communications

Fiscal concerns

Labour market

Price stability

Inflation

Source: own elaborations with Python software

Table A2: Estimated alpha parameter

Estimated Hyperparameter alpha
0.212754
0.541652
0.30765
0.430074
0.66545
0.817078
0.260469
0.681016
0.583602
0.21296
0.603577
0.182783
0.830462
0.288956

Table A3: Estimated relative topic dominance (aggregate)

	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Banking supervision	0.18	0.00	0.00	0.14	0.16	0.20	0.18	0.39	0.25	0.33	0.19	0.11	0.09	0.07	0.16	0.21	0.09	0.10	0.08	0.03	0.05	0.00	0.07	0.05	0.00	0.00
Inflation	0.28	0.39	0.31	0.33	0.30	0.28	0.28	0.20	0.25	0.21	0.21	0.18	0.14	0.19	0.18	0.06	0.12	0.16	0.19	0.16	0.22	0.09	0.13	0.14	0.10	0.09
Fiscal concerns	0.00	0.00	0.01	0.03	0.06	0.00	0.05	0.02	0.03	0.07	0.07	0.06	0.06	0.04	0.02	0.02	0.03	0.02	0.05	0.05	0.06	0.07	0.10	0.07	0.07	0.07
Labour market	0.05	0.00	0.06	0.05	0.00	0.00	0.00	0.00	0.04	0.02	0.04	0.05	0.04	0.03	0.03	0.05	0.04	0.07	0.02	0.04	0.06	0.05	0.01	0.06	0.04	0.00
Financial stability	0.00	0.12	0.16	0.07	0.10	0.09	0.12	0.04	0.04	0.04	0.10	0.13	0.21	0.09	0.07	0.03	0.10	0.12	0.14	0.13	0.21	0.29	0.21	0.28	0.25	0.37
Price stability	0.00	0.03	0.04	0.01	0.00	0.02	0.07	0.00	0.04	0.05	0.04	0.08	0.14	0.12	0.06	0.12	0.13	0.08	0.07	0.15	0.04	0.07	0.06	0.04	0.04	0.00
European integration	0.00	0.00	0.01	0.00	0.03	0.00	0.06	0.02	0.00	0.05	0.01	0.02	0.02	0.03	0.01	0.05	0.00	0.02	0.02	0.04	0.02	0.02	0.01	0.03	0.08	0.00
Monetary policy instruments	0.00	0.06	0.04	0.06	0.05	0.09	0.03	0.05	0.12	0.05	0.07	0.07	0.05	0.04	0.06	0.03	0.06	0.09	0.11	0.11	0.11	0.14	0.14	0.09	0.09	0.11
Payment system infrastructure	0.00	0.02	0.09	0.07	0.03	0.02	0.03	0.10	0.12	0.12	0.15	0.11	0.04	0.07	0.05	0.07	0.07	0.12	0.09	0.12	0.08	0.09	0.09	0.10	0.16	0.21
Currency risk	0.00	0.06	0.09	0.03	0.01	0.07	0.03	0.03	0.01	0.01	0.02	0.02	0.01	0.00	0.02	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.02	0.00	0.02	0.00
Country risk	0.03	0.03	0.10	0.08	0.07	0.05	0.07	0.06	0.03	0.01	0.01	0.03	0.03	0.05	0.07	0.07	0.07	0.04	0.05	0.06	0.04	0.05	0.06	0.08	0.10	0.09
Data-statistics	0.00	0.00	0.01	0.01	0.09	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Terms of trade	0.45	0.29	0.06	0.10	0.06	0.14	0.06	0.09	0.06	0.03	0.08	0.13	0.15	0.25	0.26	0.27	0.27	0.16	0.14	0.09	0.10	0.13	0.10	0.05	0.05	0.06
General communications	0.00	0.00	0.02	0.02	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.02	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00

Table A4: President of the ECB relative topic dominance

YEAR	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Banking supervision	1.0	0.0	0.0	1.0	0.0	0.0	1.0	6.0	2.0	3.0	4.0	3.0	2.0	1.0	2.0	2.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
Inflation	3.0	3.0	6.0	6.0	5.0	5.0	3.0	2.0	6.0	1.0	2.0	8.0	2.0	4.0	1.0	1.0	2.0	2.0	0.0	3.0	1.0	0.0	0.0	1.0	0.0	0.0
Fiscal concerns	0.0	0.0	1.0	0.0	1.0	0.0	4.0	2.0	2.0	6.0	4.0	4.0	1.0	1.0	2.0	1.0	2.0	0.0	0.0	0.0	3.0	0.0	0.0	1.0	0.0	0.0
Labour market	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	2.0	1.0	3.0	3.0	2.0	3.0	2.0	1.0	1.0	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Financial stability	0.0	2.0	5.0	1.0	3.0	2.0	1.0	2.0	1.0	1.0	2.0	5.0	6.0	2.0	5.0	0.0	3.0	4.0	6.0	8.0	8.0	7.0	7.0	4.0	7.0	4.0
Price stability	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	3.0	4.0	5.0	2.0	3.0	1.0	0.0	2.0	3.0	0.0	1.0	1.0	0.0	0.0	0.0
European integration	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	0.0	3.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Monetary policy instruments	0.0	3.0	3.0	3.0	1.0	3.0	1.0	5.0	2.0	2.0	7.0	3.0	4.0	4.0	1.0	1.0	1.0	3.0	1.0	1.0	1.0	2.0	0.0	1.0	2.0	0.0
Payment system infrastructure	0.0	1.0	2.0	1.0	1.0	0.0	1.0	4.0	8.0	6.0	8.0	5.0	2.0	2.0	1.0	3.0	2.0	1.0	3.0	5.0	5.0	1.0	1.0	2.0	5.0	1.0
Currency risk	0.0	2.0	5.0	2.0	0.0	2.0	2.0	2.0	0.0	0.0	0.0	3.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Country risk	1.0	2.0	1.0	3.0	2.0	2.0	4.0	7.0	1.0	1.0	1.0	2.0	1.0	2.0	2.0	0.0	2.0	1.0	4.0	3.0	5.0	3.0	5.0	7.0	8.0	3.0
Data-statistics	0.0	0.0	0.0	1.0	4.0	0.0	0.0	1.0	0.0	1.0	2.0	1.0	1.0	3.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
Terms of trade	14.0	13.0	3.0	2.0	2.0	7.0	1.0	8.0	4.0	1.0	5.0	7.0	9.0	15.0	12.0	8.0	13.0	6.0	7.0	0.0	1.0	3.0	3.0	1.0	1.0	0.0
General communications	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

Table A5: Vice-President of the ECB relative topic dominance

YEAR	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Banking supervision	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	1.0	3.0	0.0	1.0	0.0	0.0	1.0	1.0	2.0	1.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0
Inflation	0.0	1.0	3.0	3.0	1.0	1.0	1.0	2.0	1.0	3.0	3.0	2.0	0.0	0.0	2.0	0.0	0.0	3.0	2.0	4.0	6.0	4.0	2.0	0.0	0.0	0.0
Fiscal concerns	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	2.0	1.0	1.0	0.0	3.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Labour market	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Financial stability	0.0	1.0	2.0	0.0	2.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	2.0	1.0	0.0	1.0	0.0	1.0	1.0	2.0	2.0	2.0	5.0	4.0	3.0	1.0
Price stability	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	1.0	2.0	1.0	2.0	3.0	0.0	2.0	3.0	1.0	4.0	1.0	1.0	1.0	1.0	0.0
European integration	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
Monetary policy instruments	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	2.0	1.0	2.0	1.0	0.0	0.0	3.0	2.0	2.0	3.0	4.0	4.0	3.0	6.0	13.0	6.0	2.0	0.0
Payment system infrastructure	0.0	0.0	1.0	1.0	0.0	0.0	0.0	2.0	1.0	0.0	5.0	3.0	3.0	1.0	2.0	0.0	1.0	3.0	2.0	1.0	4.0	2.0	3.0	0.0	0.0	0.0
Currency risk	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Country risk	0.0	0.0	5.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	2.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
Data-statistics	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Terms of trade	0.0	0.0	3.0	2.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	2.0	3.0	3.0	4.0	2.0	4.0	5.0	2.0	2.0	1.0	3.0	2.0	0.0	2.0	0.0
General communications	0.0	0.0	1.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table A6: Members of the Executive Board of the ECB relative topic dominance

YEAR	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Banking supervision	0.0	0.0	0.0	2.0	1.0	4.0	3.0	7.0	4.0	6.0	3.0	0.0	1.0	1.0	2.0	2.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0
Inflation	0.0	4.0	6.0	5.0	5.0	5.0	8.0	9.0	7.0	11.0	10.0	3.0	7.0	7.0	8.0	2.0	6.0	4.0	8.0	2.0	10.0	1.0	5.0	5.0	4.0	1.0
Fiscal concerns	0.0	0.0	0.0	3.0	4.0	0.0	0.0	0.0	2.0	4.0	7.0	6.0	7.0	1.0	1.0	1.0	2.0	2.0	6.0	7.0	8.0	9.0	12.0	6.0	7.0	2.0
Labour market	0.0	0.0	4.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	4.0	3.0	0.0	1.0	2.0	3.0	4.0	1.0	4.0	8.0	5.0	1.0	5.0	3.0	0.0
Financial stability	0.0	1.0	6.0	4.0	1.0	4.0	7.0	2.0	2.0	3.0	9.0	9.0	13.0	6.0	2.0	1.0	7.0	6.0	5.0	2.0	17.0	18.0	6.0	11.0	7.0	2.0
Price stability	0.0	0.0	1.0	0.0	0.0	1.0	3.0	0.0	3.0	4.0	4.0	6.0	8.0	6.0	2.0	3.0	13.0	5.0	1.0	10.0	1.0	5.0	3.0	2.0	2.0	0.0
European integration	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	3.0	0.0	2.0	1.0	1.0	1.0	2.0	0.0	2.0	2.0	3.0	2.0	2.0	0.0	2.0	6.0	0.0
Monetary policy instruments	0.0	0.0	2.0	2.0	3.0	4.0	1.0	3.0	11.0	5.0	2.0	8.0	3.0	2.0	5.0	0.0	6.0	5.0	8.0	9.0	16.0	10.0	5.0	2.0	5.0	3.0
Payment system infrastructure	0.0	0.0	6.0	4.0	1.0	2.0	2.0	7.0	5.0	11.0	9.0	8.0	0.0	5.0	4.0	4.0	6.0	10.0	5.0	8.0	4.0	8.0	6.0	7.0	9.0	4.0
Currency risk	0.0	0.0	5.0	1.0	1.0	5.0	1.0	2.0	0.0	1.0	3.0	0.0	2.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	3.0	0.0	1.0	0.0
Country risk	0.0	0.0	9.0	6.0	5.0	4.0	5.0	3.0	3.0	2.0	2.0	5.0	4.0	5.0	10.0	8.0	9.0	4.0	3.0	7.0	5.0	4.0	5.0	3.0	4.0	0.0
Data-statistics	0.0	0.0	1.0	0.0	7.0	1.0	2.0	0.0	0.0	0.0	1.0	1.0	2.0	1.0	2.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
Terms of trade	0.0	4.0	3.0	8.0	3.0	8.0	6.0	8.0	5.0	5.0	10.0	18.0	14.0	24.0	30.0	26.0	34.0	15.0	12.0	13.0	21.0	15.0	11.0	5.0	3.0	2.0
General communications	0.0	0.0	1.0	1.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	0.0	2.0	1.0	0.0	1.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0

References

Angino, Federico Maria Ferrara and Siria. 2021. "Does clarity make central banks more engaging? Lessons from ECB communications." *European Journal of Political Economy* 102-146.

Assenmacher, K., Glöckler, G., Holton, S., Trautmann, P., Ioannou, D., Mee, S., & Taylor, E. 2021. " Clear, consistent and engaging: ECB monetary policy communication in a changing world."

Bennani, H., Fanta, N., Gertler, P., & Horvath, R. 2020. "Does central bank communication signal future monetary policy in a (post)-crisis era? The case of the ECB." *Journal of International Money and Finance* 104: 102-167.

Blei, D. M. 2012. "Probabilistic topic models." *Communications of the ACM* 55 (4): 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 993-1022.

Blinder, Alan S., Michael Ehrmann, and Marcel Fratzsche, and and David-Jan Jansen and Jakob De Haan. 2008. "Central bank communication and monetary policy: A survey of theory and evidence." *Journal of economic literature* 46 (4): 910–945.

Carboni, M.and Farina, V. and Previati, D. A. 2020. "ECB and FED Governors' Speeches: A Topic Modeling Analysis (2007–2019)." In *In Banking and Beyond*, 9-25. Palgrave Macmillan.

Christophe Blot, Paul Hubert, et al. 2018. "Central bank communication during normal and crisis times." *Monetary Dialogue.*

Cieslak, A., & Schrimpf, A. 2019. "Non-monetary news in central bank communication." *journal of International Economics* 118: 293-315.

ECB. 2019. *Speeches dataset.* October 25. https://www.ecb.europa.eu/press/key/html/downloads.en.html.

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., & Talmi, J. 2020. "Starting from a blank page? Semantic similarity in central bank communication and market volatility." *Journal of Monetary Economics* 111: 48-62.

Ferri, P., Lusiani, M., & Pareschi, L. 2018. "Accounting for Accounting History: A topic modeling approach (1996–2015)." *Accounting History* 23 (1-2): 173-205.

Gentzkow, Kelly, and Taddy, M. 2019. "Text as data." *Journal of Economic Literature* 57 (3): 535-74.

Glas, A., & Müller, L. S. 2021. "Talking in a language that everyone can understand?

Transparency of speeches by the ECB Executive Board."

Greene, D.,, and J. P. Cross. 2015. "Unveiling the political agenda of the european parliament plenary: A topical analysis." *ACM web science conference.* 1-10. Accessed June.

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. 2004. "Integrating topics and syntax." *Advances in neural information processing systems* 17.

Grimmer, J., and 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.

Hansen, S., M. McMahon, and A. Prat. 2014. "Transparency and Deliberation within the FOMC: A computational Linguistics Approach." *CEP Discussion Paper* (1276).

Hansson, M. 2021. "Evolution of topics in central bank speech communication." *Working Paper in Economics: University of Gothenburg* (881). Accessed November 5, 2021.

Hartmann, P., and Smets, F. 2018. *The first twenty years of the European Central Bank: monetary policy.* European Central Bank.

Horvath, Pavel Gertler and Roman. 2018. "Central bank communication and financial markets: New high-frequency evidence." *Journal of Financial Stability* 36: 336-345.

Isoaho, K., Gritsenko, D., & Mäkelä, E. 2021. "Topic modeling and text analysis for qualitative policy research." *Policy Studies Journal* 49 (1): 300-324.

Keida, M., & Takeda, Y. 2019. "The Art of Central Bank Communication: A Topic Analysis on Words used by the Bank of Japan's Governors." *RIETI.*

Kinywamaghana, A., & Steffen, S. 2021. "A Note on the Use of Machine Learning in Central Banking."

Küsters, A. 2022. " Applying Lessons from the Past? Exploring Historical Analogies in ECB Speeches through Text Mining, 1997–2019." *International Journal of Central Banking* 18 (1): 277-329.

Luangaram, P., and W. Wongwachara. 2017. "More than words: a textual analysis of monetary policy communication." *PIER discussion Papers* (54).

Matsui, A., Ren, X., & Ferrara, E. 2021. "Using Word Embedding to Reveal Monetary Policy Explanation Changes." *Third Workshop on Economics and Natural Language Processing.* 56-61. Accessed November.

Meltzer., Alex Cukierman and Allan H. 1986. "The credibility of monetary announcements."

Moraes., Helder Ferreira de Mendon¸ca and Claudio Oliveira de. 2018. "Central bank disclosure as a macroprudential tool for financial stability." *Economic Systems* 42 (4): 625-636.

Rajani, N. F. N., McArdle, K., & Baldridge, J. 2014. "Extracting topics based on authors, recipients and content in microblogs." *37th international ACM SIGIR conference on Research & development in information retrieval.* 1171-1174. Accessed July.

Řehůřek, Radimand, and Petr Sojka. 2011. *Gensim—statistical semantics in python.*

Roberts M. E., Stewart B. M. and Tingley, D. 2019. "Stm: An R package for structural topic models." *Journal of Statistical Software* 91: 1-40.

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

Ruchirawat, Nicha. 2020. *https://towardsdatascience.com/.* November 1. https://towardsdatascience.com/6-tips-to-optimize-an-nlp-topic-model-for-interpretability-20742f3047e2.

Rzonca., L Janikowski and A. 2018. "Central bank communication at times of non-standard monetary policies." *Monetary Dialogue.*

SARET, Jeffrey N., and Subhadeep. MITRA. 2016. " An ai approach to fed watching." *Two Sigma Insights.*

Scarpari, GC Montes and A. 2015. "Does central bank communication affect bank risk-taking?" *Applied Economics Letters* 22 (9): 751-758.

Siklos, Pierre L and St Amand, Samantha and Wajda, Joanna. 2018. "The Evolving Scope and Content of Central Bank Speeches." *Centre for International Governance Innovation.*

Treude, C., & Wagner, M. 2018. "Per-corpus configuration of topic modelling for github and stack overflow collections." *arXiv preprint arXiv:1804.04749*, 157-168.

user20160. 2018. *How to interpret t-SNE plot?* August 27. https://stats.stackexchange.com/questions/331745/how-to-interpret-t-sne-plot.

Vayid, Ianthi. 2013. "Central bank communications before, during and after the crisis: from open-market operations to open-mouth policy."

Zumbach, J., & Volbert, R. 2021. "What Judges Want to Know From Forensic Evaluators in Child Custody and Child Protection Cases: Analyzing Forensic Assignments With Latent Dirichlet Allocation." *Frontiers in Psychology* 12.