Interdisciplinary Project - TU Vienna & WU Vienna

Project Report - Understanding Central Bank Communication: A Machine Learning Analysis of ECB's Public Outreach

Paul von Hirschhausen (01453360)
az. Prof.in Dr.in Claudia Wutscher (Supervisor - WU Vienna)
Univ.Ass. Dr. Gábor Recski (Co-Supervisor - TU Vienna)

07.09.2025

Contents

1. Introduction

Central bank communication is a core policy instrument. For the European Central Bank (ECB), transparent and consistent messaging underpins both its monetary policy transmission and its role as banking supervisor. The ECB explicitly frames communication as part of accountability and good governance, stressing regular engagement with stakeholders across markets, politics, academia, and civil society to contextualize economic and financial dynamics as well as broader societal developments (European Central Bank, 2025).

Beyond market signaling, communication serves to anchor expectations, reduce uncertainty, and facilitate policy effectiveness. As the ECB has broadened its outreach—e.g., blogs, explainer pages, and social channels—its audience extends well beyond financial specialists. This diversification raises a substantive empirical question: how and when does the thematic focus of ECB communication shift, particularly around salient policy domains such as climate and the environment?

Prior work has analyzed the content of central-bank speeches, press conferences, and policy accounts, often using Latent Dirichlet Allocation (LDA) and related methods, and typically linking communication ⇒ market or policy outcomes market or policy outcomes such as Fortes & Le Guenedal (2021) or Kaminskas & Jurkšas (2024). In most cases the underlying corpora are restricted to official speeches and closely related formal channels rather than the ECB's broader communication arsenal, including blogs and explainer pages. Comparatively less attention has been given to the inverse perspective, policy, social, and macro events ⇒ communication, which is the perspective emphasized here. In particular, we study whether the prominence of climate-related themes in ECB communication exhibits structural changes over time.

Topic modeling offers a scalable way to quantify thematic structure in large, evolving text collections. However, purely unsupervised approaches can struggle with domain-faithful labeling and stable evaluation. This report therefore combines supervised and semi-supervised topic modeling with time-series change-point detection to (i) recover interpretable, policy-relevant themes and (ii) test for statistically significant shifts in climate-related communication.

To translate this motivation into concrete steps, we set the following research objectives: first we construct a comprehensive corpus of ECB communication (press releases, blogs, speeches) with maximal temporal coverage and rich metadata, secondly we recover and validate topic structure using both unsupervised (BERTopic-style) and supervised (multilabel) methods, with an emphasis on policy-relevant labels such as climate/environment, third we quantify the temporal dynamics of climate-related communication and detect structural breaks using statistically principled change-point methods and fourth we compare modeling choices, compraing embedding backends, supervised versus unsupervised with post-hoc mapping, and validation-time threshold calibration, with respect to predictive performance and interpretability.

To make these aims precise, we address the following research questions:

- RQ1: What are the dominant topics in ECB communication across channels and authors, and how consistently can they be identified using supervised vs. unsupervised methods?
- RQ2: How does the share of climate-related communication evolve, and do we observe statistically significant change points?

This Interdisciplinary project contributes along four dimensions in a single, integrated workflow: we first build a scraped ECB communication dataset spanning press, blogs, and speeches, retaining ground-truth labels where available and producing model-predicted labels elsewhere to enable full-coverage analysis; we then deliver a head-to-head comparison of unsupervised BERTopic-style pipelines (with topic $\rightarrow$ label mapping) and supervised multilabel pipelines (fine-tuned XLM-RoBERTa and CatBoost on document embeddings), including validation-time threshold optimization for F1; and lastly we quantify temporal structure in climate-related communication via a CUSUM/Galeano-style change-point analysis that yields dated structural breaks suitable for policy interpretation.

Finally, regarding scope and interpretive stance, our empirical focus is on communication dynamics: what changes, and when. We refrain from causal attribution of the detected breaks, and the qualitative interpretation of potential drivers (policy decisions, exogenous shocks, institutional initiatives) is left to the reader and to future work that integrates external event data. The full code and reproducible pipeline will also be accessible via the GitHub repository `https://github.com/Habenzu/ECB_Topic_Chronicles.git`.

2. Related Work

XXXXX This chapter situates our study in three strands: ECB communication with a focus on climate, topic modeling for policy text, and change point detection for communication time series. We only sketch the methods here to keep the narrative flowing. By topic modeling we mean approaches that infer recurring themes from large text collections, ranging from probabilistic models like LDA to embedding plus clustering methods such as BERTopic; details follow in the next section.

ECB communication and climate

The ECB treats communication as a policy instrument that supports accountability, transparency, and effective transmission of decisions. Its guiding principles explicitly emphasize clear, timely engagement with both specialised and general audiences, and a two-way dialogue that spans speeches, press events, and web-based outreach (European Central Bank, 2025). In parallel, international guidance frames communication as part of the central bank's toolkit and recommends using a broad mix of channels, including press releases, press conferences, websites, and social media, with content tailored to different audiences (Casiraghi & Perez, 2022).

Most empirical studies build on official speeches and a narrow set of formal events, then link content or sentiment to financial markets or policy stance. Hansson (2021) analyses speeches

from nine major central banks with dynamic topic models and reports that central banks cover a wide range of themes whose shares are strongly persistent over time. For the ECB specifically, Fortes & Le Guenedal (2021) extract topics and sentiment and show that these signals improve models of the EUR/USD exchange rate. Kaminskas & Jurkšas (2024) extends the scope by quantifying sentiment across press conferences, accounts, and Executive Board speeches, and by examining the relationship between tone and selected market indicators while also capturing media reactions. Overall, this strand treats communication as a driver of outcomes in markets and expectations rather than as an object whose structure might itself change in response to events (Fortes & Le Guenedal, 2021; Kaminskas & Jurkšas, 2024).

A growing literature focuses on how central banks communicate about climate change. Using a large corpus of speeches from over one hundred institutions, Arseneau et al. (2022) document a sharp rise in climate communication, a wide breadth of covered themes from macroeconomic impacts to sustainable finance, and relatively infrequent explicit discussion of macroprudential policy, together with more speculative language than in other topics. For the ECB context, these findings support treating climate as a distinct and evolving theme within official communication (Arseneau et al., 2022).

Three gaps follow from this evidence. First, broader outreach formats remain underrepresented relative to speeches, even though official and international guidance designate websites and social channels as core tools for reaching diverse audiences (Casiraghi & Perez, 2022; European Central Bank, 2025). Second, the reverse direction is comparatively underexplored: whether policy or social events coincide with structural changes in the communication record itself, especially for climate-related discourse, while existing work for the ECB mainly evaluates effects on markets and stance (Fortes & Le Guenedal, 2021; Kaminskas & Jurkšas, 2024). Third, large scale analyses often rely on unsupervised models applied to speeches, which complicates label alignment and evaluation when extending coverage to additional channels or languages (Hansson, 2021).

In response, this study assembles a multi-channel corpus that combines formal events with additional outreach hosted on the ECB website, recovers policy-relevant topics that include climate, and tests for statistically identifiable structural changes in climate-related communication over time. This design complements speech-based studies and follows institutional guidance to reach multiple audiences through multiple channels (Casiraghi & Perez, 2022; European Central Bank, 2025).

Topic modeling for policy text

To analyze themes consistently across channels and years, we require topic representations that are both scalable and aligned with policy labels. We therefore review topic modeling approaches next, with a focus on their suitability for ECB communication and their evaluation constraints.

Latent Dirichlet Allocation provides a generative foundation for discovering mixture-of-topics structure in text, and Dynamic Topic Models extend this idea to capture smooth topic

evolution over time. These models are historically important, but they rely on bag-of-words representations that struggle with short, technical, and multilingual policy text where meaning depends on context and phrasing. In multilingual settings, vocabulary fragmentation and polysemy further degrade topic quality and cross-time comparability, which motivates moving beyond pure bag-of-words (**Blei2003LDA?**; **BleiLafferty2006DTM?**).

Modern pipelines first produce contextual document embeddings, then cluster these embeddings, and finally derive topic descriptors. BERTopic follows this pattern and introduces class-TF-IDF to obtain compact, human-readable topic terms from clustered documents, which often improves semantic coherence for policy text. Contextualized Topic Models integrate transformer representations inside a probabilistic topic model, including cross-lingual variants that are useful when documents span multiple ECB languages. These methods capture context, idioms, and interlingual similarity better than bag-of-words, but they also raise practical questions about evaluation and label alignment across time and channels (**Grootendorst2022?**; **Bianchi2021CTM?**).

Automated coherence scores are convenient, yet human studies show a gap between such metrics and what domain experts perceive as interpretable or policy-faithful topics. For applied work, this means model selection based purely on coherence can be misleading, especially when topics must align with a predefined policy taxonomy. In addition, unsupervised topics may drift over time or split or merge in ways that complicate longitudinal comparisons (**Chang2009RTL?**; **Roder2015Coherence?**).

When ground-truth labels exist for part of the corpus, unsupervised clusters can be mapped to that label set after the fact, for example via regression from topic activations to labels or via optimal one-to-one assignment. This can improve comparability but may be unstable when labeled data are sparse or when cluster structure shifts across time. The Hungarian method gives a principled formulation of the one-to-one mapping problem that is often used as a baseline for such alignment. Supervised topic models provide an alternative that couples topic discovery with predictive targets, which strengthens alignment with policy-relevant labels (**Kuhn1955Hungarian?**; **BleiMcAuliffe2007?**).

To obtain policy-aligned topics at scale, we adopt two complementary supervised approaches. First, a fine-tuned multilingual transformer with a multilabel classification head directly predicts topic labels, which is attractive for multilingual ECB text given strong cross-lingual transfer of XLM-RoBERTa. Second, we freeze embeddings and train a strong tabular learner such as CatBoost on top, which is robust, fast to iterate, and often straightforward to calibrate. For both families, we calibrate decision thresholds on validation data to maximize F1, and we consider logit adjustment when label frequencies are highly imbalanced (**Conneau2019XLMR?**; **Prokhorenkova2018CatBoost?**; **Guo2017Calib?**; **Lipton2014F1?**; **Menon2021LogitAdj?**).

In our comparison, we evaluate three pipelines on the same train, validation, and test splits with an identical label set and reporting: an unsupervised BERTopic-style pipeline with post-hoc topic-to-label mapping, an embedding-plus-CatBoost classifier, and a fine-

tuned XLM-RoBERTa multilabel classifier. We report micro- and macro-F1 as primary metrics and discuss calibration and label-imbalance treatments that affect deployment to the unlabeled portion of the corpus (**Grootendorst2022?**; **Prokhorenkova2018CatBoost?**; **Conneau2019XLMR?**; **Guo2017Calib?**; **Lipton2014F1?**; **Menon2021LogitAdj?**).

Once documents carry reliable topic labels and calibrated scores, we can track the prevalence of climate related communication through time. We now review change point detection methods that allow us to test for statistically significant structural changes in that prevalence.

Change point detection for communication time series

In our setting the observable is either the prevalence of climate related topics over time or the sequence of climate topic events implied by document counts, with temporal aggregation chosen to balance variance and temporal resolution. Daily aggregation captures fine grained shocks but can be sparse, weekly aggregation stabilizes counts for moderate volumes, and monthly aggregation is often appropriate for policy communication where cycles are slower. When we convert counts into event times we can treat the resulting process as a point process and analyze deviations from a constant intensity benchmark (**Truong2020Review?**).

Multiple offline change point detection methods provide principled tools for this task. CUSUM type procedures test for changes in level or intensity by accumulating deviations from a reference model and comparing the resulting fluctuation statistic to a calibrated critical value. Exact dynamic programming approaches such as PELT solve the global segmentation with a penalty for the number of changes and achieve near linear cost under mild conditions, which makes them attractive for large corpora. Nonparametric energy distance methods like e divisive detect general distributional shifts without strong parametric assumptions and therefore serve as robust baselines when changes affect variability or higher moments rather than only the mean (**Truong2020Review?**; **Killick2012PELT?**; **Matteson2014EDivisive?**).

Our primary method follows a cumulative sum approach tailored to event style data. We compute a CUSUM statistic on event times to detect departures from a constant rate, obtain critical values by Monte Carlo under the null of a homogeneous Poisson process, and apply a binary segmentation routine that repeatedly tests the most prominent candidate split, subject to a minimum segment length. To control family wise error, we adjust the per test size during recursion, which limits spurious detections while retaining sensitivity to genuine shifts. This procedure is well suited to document arrival data where intensity changes correspond to communication bursts or lulls and where the ordering of events carries more signal than raw count magnitudes (**Galeano2007CUSUMPoisson?**; **Vostrikova1981BinarySeg?**).

For robustness we consider complementary specifications. First, we can run PELT on percentage time series of climate topic share rather than event times, which targets piecewise constant means in a directly interpretable metric. Second, we can apply e divisive to residualized signals after removing seasonal effects or long run trends, which increases power against broader distributional changes. Across these variants we control false positives through penalty calibration or resampling, and we choose temporal resolution by cross checking that

detected breaks persist across daily, weekly, and monthly aggregations (**Killick2012PELT?**; **Matteson2014EDivisive?**; **Truong2020Review?**).

Detected change points should be read as descriptive markers in the communication record. They locate statistically significant structural breaks, but they do not on their own identify causal drivers. Linking breaks to policy decisions, macro shocks, or institutional initiatives requires external event data and is outside the scope of this section (**Truong2020Review?**).

3. Data Collection & Preprocessing

In this chapter we will explain the general data collection and preporcessing pipelines whcih were used for this project and how it differs to the data collection by other research, and will also touch on the challenges when collecting this data.

In the first iteration of the project we focussed on the unsupervised learning part, by using already collected data accesible via Kaggle (**Olofaro2020ECBData?**), this dataset had the advantage of bein updated once a week by the creator of the dataset, also it included all speeches. After some initial experimentation with uunsupervised learning on this dataset, the generated models were able to edetect latent topics but we felt the need to evaluate the performance more than just based on unsupervised topic modelling metrics. Since evaluation of this amount of data in a qualitative manner is rather time consuming, we decided to build a dedicated scraper to colelct out own data. We focussed on the main channel for the ECB to publish any kind of communication (outside of social media channels) https://www.ecb.europa.eu/press/pubbydate/html/index.en.html. There all communication from 1997 up to now is accessible and thankfully some part of it are already categorized into topics. What then came apparent was that scraping via the filter type of `topic` we were only able to scrape 36% of the total roughly 10.000 news and publications, this was due to the "continuous scroll" setup of the page. What worked, to get all publications, was to use the filter type `year`, with this we were able to collect all publications, where in total we were able to get 36% already labeled. We then deduplicated the entries, created aggreations of topics, since one publication can be assigned to multiple topics, and we then parsed the downloaded html files with an sopecific parser. Due to the structure of the news & publication page of the ECb we were also able to collect various metdata which heklps in the further analysis. The distribution of the colledcdted samples per year can be viewed in the follwoing figure.

The final raw dataset consists of 9525 publications, where 3494 of them are labeled with in total 102 different topics. For each entry we colelcted the following data: * `date`: Publication date * `category`: ECBs assigned categopry to the publication (eg. Governing Council statement, Research Bulletin, Speech, ...) * title: Title of the publication * filter_type: The tyupe of filter used during scraping and downloading of the html (`year` or `topic`) * filter_value: Either the topic or the year used when scraping * url: the url of the downloaded html * authors: list of authors * text: the parsed text from the downloaded html file

This pipeline can be found in the file `all_new_publication_scraper.py` in the repository.

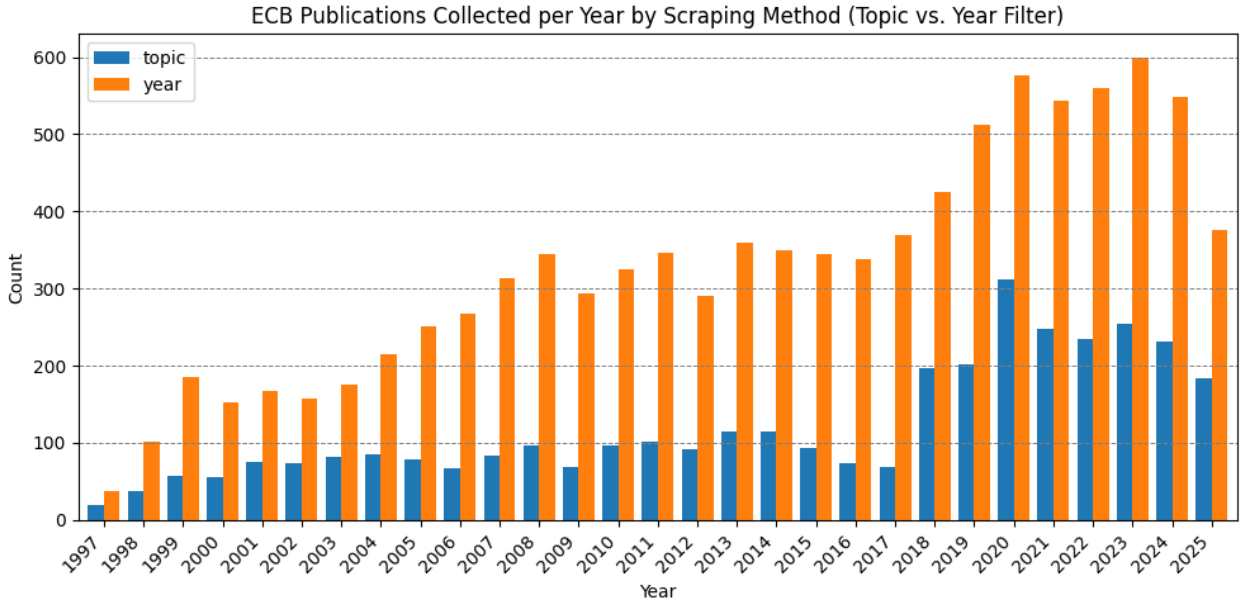


Figure 1: ECB Publications Collected per Year by Scraping Method (Topic vs. Year Filter)

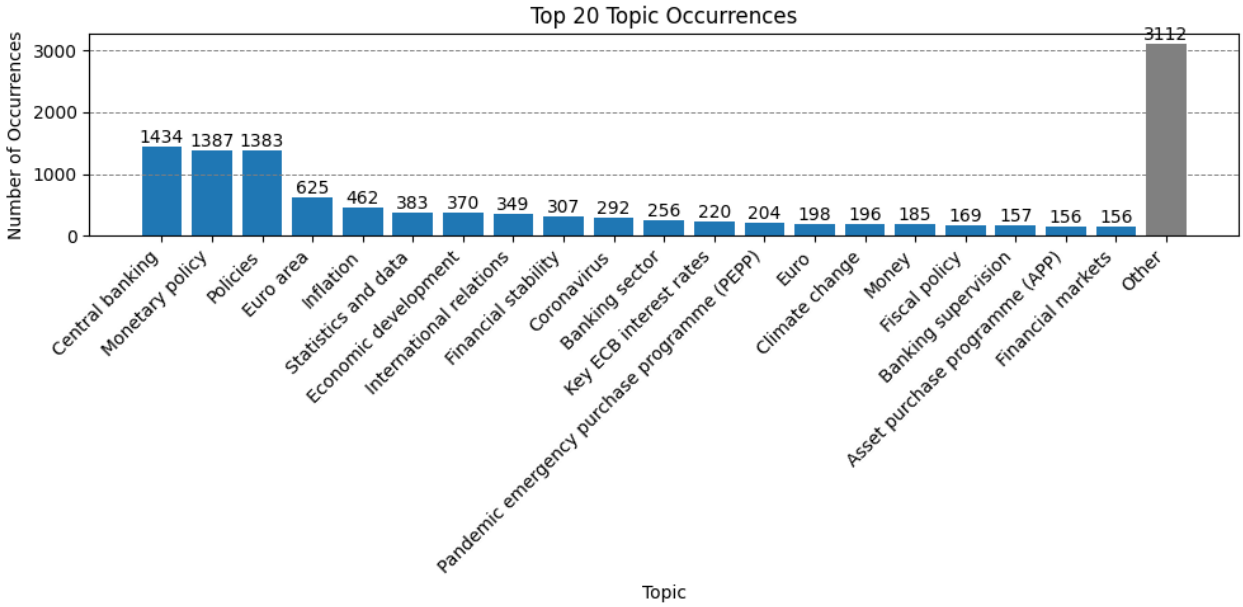


Figure 2: Top 20 Topic Occurrences

In the next step we dropped some rows which were not suitable for the analysius (eg. empty texts or pdf files) and split the labeled data into train (70%), validation (15%) and test (15%).

After that depending on the backbones embedding models context window length we applied token absed chunking and embedded the chunks with the mebdding models. the embnedding models used in this project are one propritory model OpenAIs `text-embedding-3-large`, and two open source models from the `sentence-transformers` package `all-MiniLM-L12-v2` and `all-mpnet-base-v2`. teh decision to use is to investigate how open source is operformaning against closed source and how the context window length contributed to the performance in such settings, and openais model is one of the most widly used embedding models in practical applicatio, while the other two are the top performing allrounder embvedding models easilyu accesible.

Once we have the data prepared, which means, we have text chunks with respective embeddings we can train the different models. This will be explained in the next chapter

4. Methodology

In this chapter we will first dive deeper in the final trainngs pipelines, what kind of archtectures and approaches were used to finally select a best perfromuing model we copuld then further use to label the unlabeled 64% of the dataset. After that we will briefly discuss the applied change point detection method.

Supervised and Unsupervised Training

As already mentioned we have trained multiple supervised and unsupervised models, with different backends. In general we trained the following models to benchmark the best model for topic modelling on this datset.

Unsupervised

We trained three different unsupervised BERTopic models, with the three embedding backends`text-embedding-3-large`,`all-MiniLM-L12-v2` and `all-mpnet-base-v2`. We decided to only investigate this on embeddings based approach because research has shown that BERTopic beause of its transformer based architecture to outperform more classical approaches, furthermore due to this architecture data preprocessing (tokenization, stemming etc.) is not needed. In general BERTopic uses the BERT word embedding vectors for topic modeling. A document is first to be embedded into word vectors. The high-dimensional word vectors then go through dimensionality reduction in order to be clustered into topics. BERTopic has a sequence of five modular components, normally with BERT as the embedding backend, but each of the components can be exchanged with other approaches as well.

In our approach we adapt this pipeline by additonally using a the OpenAI transformer embdding model to the two BERT based models, and also by adding a mapping step after the

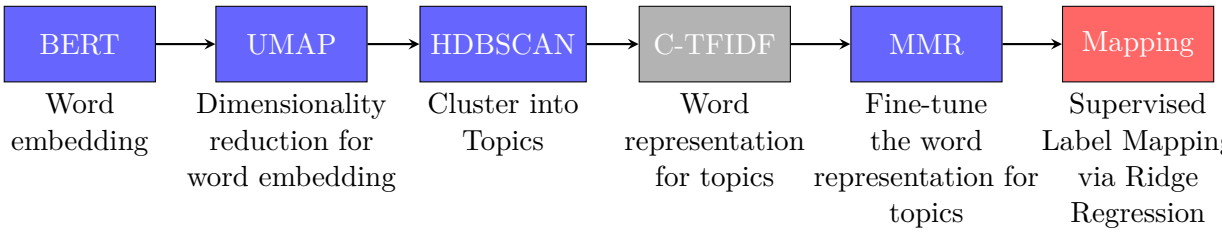


Figure 3: Overview of the BERTopic pipeline (blue: default, red: additional components, gray: theoretically not needed component).

MMR step opf the default BERTopic pipeline. Furthermore we applied a limited the number of clusters to 102 which, which is the total number of topics in our supervised dataset, this is necessary for the mapping to work.

First we embed the documents in the latent vector space created by the transformer based model (OpenAI embedding model and the other two share similar architectures). Dense document representations are constructed by first tokenizing the input and mapping each token and its position to a high-dimensional embedding vector; these initial embeddings are then processed through multiple layers of transformer blocks—each consisting of self-attention, feedforward networks, and normalization—with each layer refining the contextual meaning through learned transformations. Throughout this process, the model integrates contextual information across the entire sequence, resulting in rich, contextualized embeddings where each vector captures the nuanced semantics of its token within the document (Jurafsky & Martin, 2025).

After generating dense document embeddings, BERTopic applies Uniform Manifold Approximation and Projection (UMAP) to reduce their dimensionality while aiming to preserve the local and global structure of the data. UMAP operates by constructing a graph of nearest neighbors in the original high-dimensional space and optimizing a low-dimensional embedding that maintains these neighborhood relationships. We set the parameters to `n_neighbors=15`, `n_components=5`, `min_dist=0.0` and `metric='cosine'`, which indicates that each point considers its 15 nearest neighbors, projects data down to 5 dimensions, allows points to be packed together as closely as possible (min_dist=0.0), and measures distances using cosine similarity (Kuo, 2023; McInnes et al., 2020).

In the next step BERTopic creates clusters from the reduced embeddings using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). This density-based clustering algorithm identifies clusters of varying shapes and densities without requiring a predefined number of clusters, making it robust to noise and outliers. The parameters used here: `min_cluster_size=15` (`min_cluster_size=5` for `text-embedding-3-large` because with a cluster size of 15 we got to little clusters) sets the minimum size for a cluster, ensuring clusters contain at least 5 documents; `min_samples=1` controls how conservative the clustering is with respect to labeling points as noise; `metric='cosine'` measures similarity based on cosine distance, suitable for text embeddings; `cluster_selection_method='eom'` uses the

excess of mass method to select clusters; and `prediction_data=True` enables later assignment of new points to clusters. This approach avoids forcing every point into a cluster, improving the quality of topics by filtering out noise (Campello et al., 2013; Kuo, 2023).

In the default setup of BERTopic a topic representation creation step would be needed now and would follow the clustering. In our approach we still left it in the pipeline, but since we are then doing a mapping between our unsupervised predicted topic ids and the ground truth topics within the validation set we, theoretically dont need this step. Because we chunked the documents to fit into the context window of the embedding model, we create multiple predictions per document, which can then be used to create histogram feature vectors as follows:

$$h_d = \frac{1}{\sum_{j=1}^{K} \sum_{c=1}^{C_d} \mathbf{1}\{t_{d,c} = j\}} \left(\sum_{c=1}^{C_d} \mathbf{1}\{t_{d,c} = 1\}, \ldots, \sum_{c=1}^{C_d} \mathbf{1}\{t_{d,c} = K\} \right).$$

Where d is the document index, C_d is the number of chunks in the document, K is the total number of topics (in our case 102), and $t_{d,c}$ is the predicted topic ID of chunk c in document d. Now we can learn a non-negative linear mapping from topic histograms to labels using ridge regression. Let

$$T_{train} \in R^{N_{tr} \times K}$$

denote the topic histogram representations of the training documents and let

$$y_l \in \{0, 1\}^{N_{tr}}$$

be the binary indicator vector for label $l \in \{1, \ldots, L\}$. For each label, we solve the optimization problem

$$w_l = \arg\min_{w \geq 0} \ \|y_l - T_{train} w\|_2^2 + \alpha \|w\|_2^2,$$

where $\alpha > 0$ is the ridge regularization parameter. The solution yields a weight vector

$$w_l \in R^K,$$

which specifies how strongly each topic contributes to the prediction of label l. Stacking these vectors gives the mapping matrix

$$W = [w_1, \ldots, w_L] \in R^{K \times L}.$$

The non-negativity constraint ensures interpretability, as topics can only contribute positively to label activations.

In the second step, we transform the validation set representations T_{val} into label scores by

$$S_{val} = T_{val} W.$$

where $s_{d,l}$ denotes the score of label l for document i. Since these scores are continuous, they must be thresholded to produce binary predictions. A fixed global threshold (e.g. $t = 0.5$) is generally suboptimal because the marginal distributions of scores differ across labels, particularly under class imbalance. We therefore calibrate an individual threshold τ_l for each label by maximizing the F1-score on the validation set:

$$\tau_l = \arg\max_{t \in G} \; F1(y_{val,l},\; 1\{S_{val,l} \geq t\}) \,,$$

where G is a predefined grid of candidate thresholds. At test time, the final prediction rule is given by

$$\hat{y}_{d,l} = 1\{t_i^\top w_l \geq \tau_l\},$$

with t_d denoting the topic histogram of document d.

This procedure yields a simple mapping from topics to labels, while the calibrated thresholds ensure that binary predictions are aligned with the evaluation metric and robust to label imbalance. And we can use this to compare our unsupervised topic clusters to the uspervised ones, which mehtods we are explaining now.

Supervised

The first three models we are trianing are the once suimilar to the BERTopic pipeline, but instead of applying dimensionality reductiona or clusterting to our embedding vectors we are training a boosted random forest on those embedding vectors as feature vectors. For this we are using the `CatBoost` framework, which for this comparison here has multiple adfvnatages, firstly its considered and benchmarked to be for the most cases the most performing approach on tablar datasets (Shmuel et al., 2024) and its rather easy to set up, because of the boosted tree notion we do not have to take care about any data preprocessing and can for simplicity reasons use the out of the box default parameter settings.

CatBoost is a gradient boosting decision tree algorithm specially designed to handle categorical features efficiently without requiring extensive preprocessing. It builds an ensemble of symmetric trees sequentially, where each tree is trained to correct the errors of the previous ones, improving model accuracy iteratively (Prokhorenkova et al., 2019). Since the CatBoost classifier only predicts one lable we integrate this model into a `OneVsRestClassifier` which

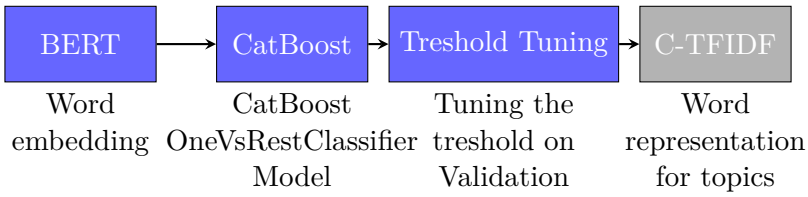


Figure 4: Supervised Learning Pipeline to match BERTopic approach (gray: theoretically not needed component)

then can return prediction scores for each topic. Since we are predicting multiple scores per chunk, and we have multiple chunks per document, we aggregate the chunk-level scores to the document level by taking the maximum across all chunks of a given document, i.e. a document receives a high score for a label if at least one of its chunks has a high score for that label. The thresholds used here are created by a treshiold tuning on the validation set to maximize the F1 score. Additonally we implemented a creation for the term representation, as implemented in BERTopic, but this is generally not needed and justy for investigative purposes.

The second batch of supervised models uses a transformer-based architecture fine-tuned for multi-label classification, employing the `xlm-roberta-base` backbone configured to predict 102 labels. The model applies mean pooling over token embeddings, which aggregates token-level contextual representations into a single vector per document chunk, and includes a 0.2 dropout for regularization to reduce overfitting. To handle long documents exceeding the model's maximum input length, texts are chunked into overlapping segments of 512 tokens with a stride of 64, and chunk-level predictions are aggregated by taking the maximum score per label to produce document-level predictions. Training is performed for 8 epochs with a batch size of 8, using a learning rate of 2e-5, weight decay of 0.01, and a linear warm-up schedule over 6% of total optimization steps, helping stabilize training early on.

For the multi-label prediction task, the loss function used is `BCEWithLogitsLoss`, which combines a sigmoid activation with binary cross-entropy loss in a numerically stable manner. This loss function independently models the presence or absence of each label by treating multi-label classification as multiple binary classification problems and calculating the sum of individual binary cross-entropy losses. The use of `BCEWithLogitsLoss` is theoretically grounded in optimizing the likelihood of observing the true label set under the assumption of label independence, while the sigmoid function converts raw model logits to probabilities for each label (Fallah et al., 2022).

Evaluation of the Topic Modelling Approaches

In order to properly compare the different approaches, and to investigate the performance of the trained models, we report a broad set of metrics that capture both per-document and per-label performance. Example-based metrics (example precision, recall, and F1 as well as Jaccard similarity) treat each document as a set of topics and assess the overlap

between predicted and gold labels, with subset accuracy representing the strictest case where the entire set must match. Hamming loss complements this by counting average label-wise mismatches per sample. Micro-averaged scores (precision, recall, F1) pool all predictions and are dominated by frequent labels, while macro-averaged scores average performance equally across all labels, making them sensitive to rare topics; we also include macro balanced accuracy to correct for label imbalance. Weighted scores sit between these extremes, weighting label-wise metrics by support. Together, these measures provide a nuanced view of prediction quality: from exact correctness at the document level to fairness across common and rare topics, enabling a balanced comparison of model performance. The comparisons was done on the test set of 15% of the total labeled dataset (525 documents across 102 topics).

Change Point Detection

After training and evaluating the model, we predicted the unlabeld part of the dataset inr oder to increase the data coverage of the topics. After that we use those predicted topics for out change point detection part in wich involves preprocessing temporal event data aggregated by topic counts and percentages across configurable intervals (e.g., daily, weekly, monthly). The data is cleaned and aligned into consistent time steps by filling missing periods with zeros, ensuring robust temporal continuity. Individual event timestamps are reconstructed by evenly distributing aggregated counts within each period, which allows application of time-based change detection methods. The core detection algorithm computes a cumulative sum (CUSUM) statistic adapted from Galeano (2007), measuring deviations from an assumed uniform event distribution—a signature of structural breaks or change points. Critical values for the test statistic are estimated via Monte Carlo simulations tailored per segment length and significance level to control type I error rates rigorously. Detected change points are identified using a binary segmentation procedure that recursively partitions the timeline at statistically significant deviations while controlling the family-wise error rate. This nonparametric, distribution-free method enables sensitive and interpretable detection of shifts in event occurrence dynamics. Visualization of CUSUM deviations further supports validation and interpretability of detected change points. This methodological pipeline, combining careful temporal aggregation, event reconstruction, and powerful statistical testing, facilitates robust analysis of temporal dynamics in topic-related event data (Killick et al., 2012).

6. Results

In this chapter we are comparing the different modelling approaches, from there derive the best performing model which is then used for the change point detection part and then we can answer the research questions appropriatly.

Topic Modelling

As previously mentioned we created two sets of models, unsupervised and supervised models end evaluated each model on example based and weighted metrics.

Table 1: Performance comparison of models for multi-label topic prediction. For each metric, the best result is highlighted in green and the second-best in blue. The prefix *usv* refers to unsupervised BERTopic pipelines, whereas *cb* denotes CatBoost-based models.

metric	usv-text-embedding-3-large	usv-all-mpnet-base-v2	usv-all-MiniLM-L12-v2	cb-text-embedding-3-large	cb-all-MiniLM-L12-v2	cb-all-mpnet-base-v2	finetuned-xlm-roberta-base
n-labels	102	102	102	102	102	102	102
n-samples	525	525	525	525	525	525	525
example-f1	0.491	0.523	0.473	**0.726**	0.486	0.643	**0.654**
example-precision	0.449	0.503	0.393	**0.748**	0.381	0.575	**0.670**
example-recall	0.674	0.684	0.759	0.765	**0.947**	**0.881**	0.711
hamming-loss	0.054	0.045	0.060	**0.019**	0.099	0.046	**0.029**
jaccard	0.397	0.430	0.364	**0.653**	0.375	0.548	**0.581**
macro-balanced-accuracy	0.622	0.673	0.672	0.648	**0.768**	**0.734**	0.692
macro-f1	0.226	0.254	0.251	**0.329**	0.247	0.309	**0.330**
macro-precision	0.241	0.246	0.221	**0.397**	0.177	0.244	**0.357**
macro-recall	0.289	0.382	0.400	0.306	**0.644**	**0.516**	0.404
micro-f1	0.403	0.466	0.425	**0.699**	0.383	0.551	**0.601**
micro-precision	0.317	0.382	0.310	**0.722**	0.241	0.404	**0.548**
micro-recall	0.553	0.598	0.679	0.677	**0.939**	**0.864**	0.665
subset-accuracy	0.177	0.200	0.116	**0.419**	0.130	0.307	**0.370**
weighted-f1	0.480	0.536	0.566	**0.668**	0.491	0.599	**0.632**
weighted-precision	0.487	0.540	0.520	**0.690**	0.353	0.478	**0.641**
weighted-recall	0.553	0.598	0.679	0.677	**0.939**	**0.864**	0.665

From those metrics we can rather clearly see and not very much surpiosignly can see that supervised modelling directly (not via unsupervised modelling) is outperformuing the unsupervised approach. In general the CatBoost model based on the `text-embedding-3-large` embeddings performs the best on most metrics. In close second the finetuned `xlm-roberta-base` model is also performing rather strong on most metrics, though as the `cb-text-embedding-3-large` the model has rather low recall metrics.

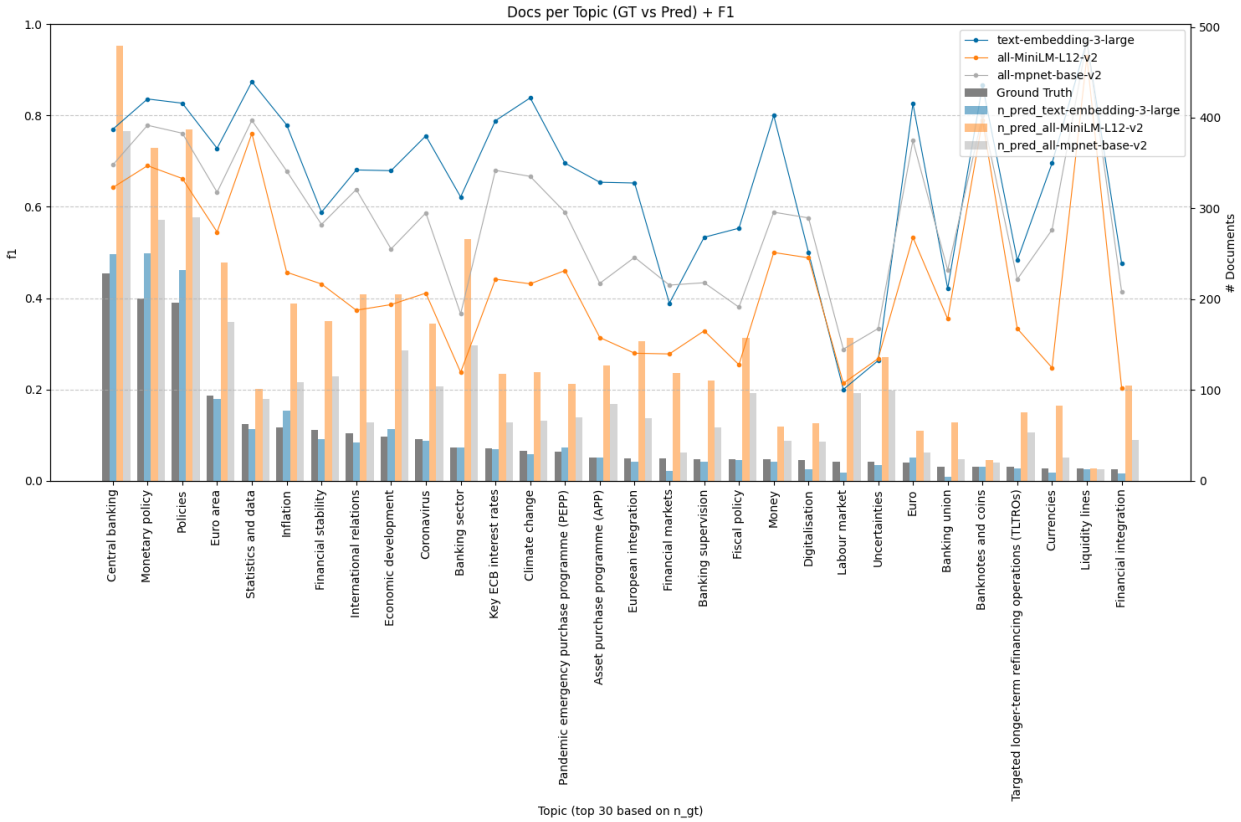
Interestingly when looking at the unsupervised topic model approaches, we can see the significance in this pipeline for smaller chunks. Since we created as large as possible chunks, the `text-embedding-3-large` does not create a lot of chunks, thus especially due to the nature of BERTopic which does intrinsically does not support multilabel topics, the perfromance is lower than for the embedding models with a smaller context window.

From this we can conclude that the large embeddings created by `text-embedding-3-large` (3072) carry proper information and value in the prediction of document topics, though apparently the two times lartger embeddings of `allmpnet-base-v2` (768) are based on this evaluation here not more informative than the `all-MiniLM-L12-v2` (384) embeddings.

Furthermore we can also see that finetuning an pretrained model like `xlm-roberta-base`, though not performing bad, is in some situations not worth the computatuional effort, since the finetuning took much more time compared to the catboost approaches.

Regarding the per topic performance of different models we can also see some similarities in trends of which topics are easier to predict and harder to predict swith varying degrees of performance byt the general performance of the model. The below figure examplaritly shows the performance for the catboost based modelling approaches on the test set. We can clearly see that the `text-embedding-3-large` backend almost consistently outperforms the other emebdding backends, while especially good for out use case, has a very strong performance of

predicting the topic “Climate Change”.

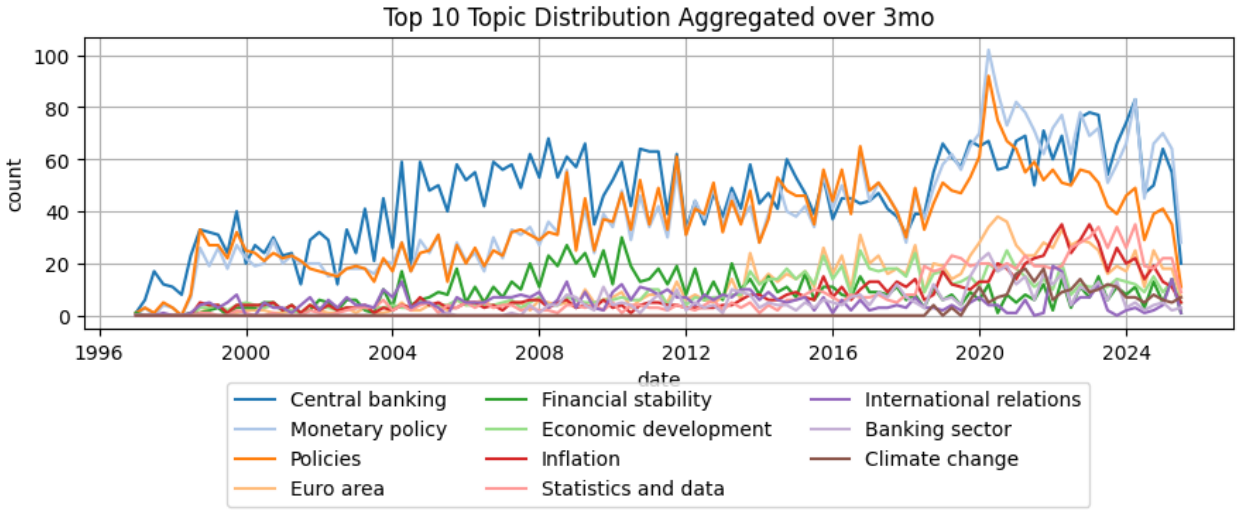


Topic Distributions

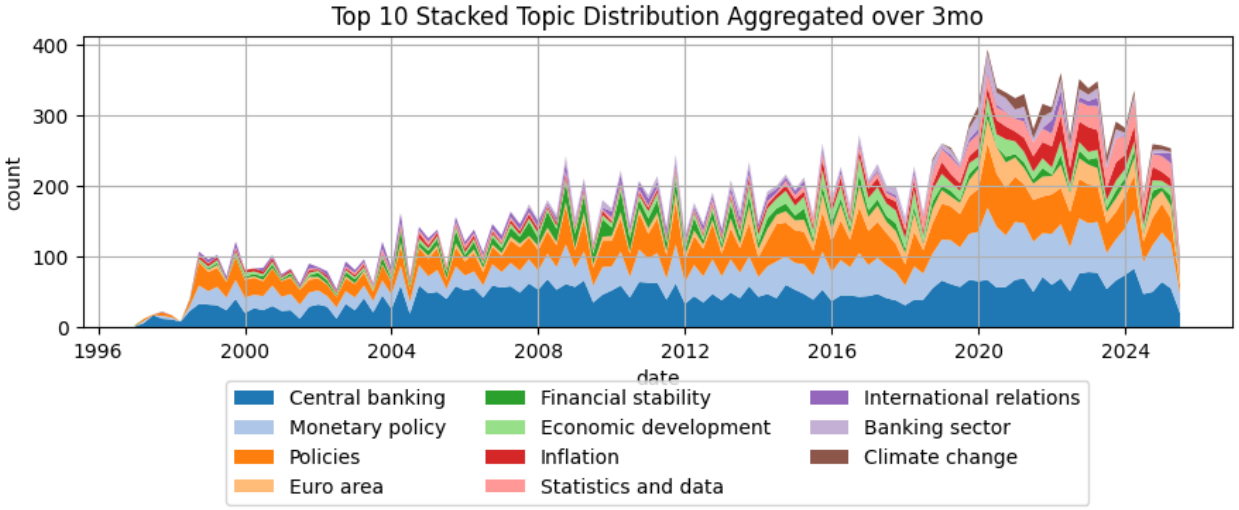
After we have now predicted the unlabeled dataset we can have a deeper look into the gernal dristribution of topics over time, see topic preferences of authors and can see how topics are distriobuted over the categories.

First we look at the general distribution of topics over time. We can clearly see that the overall number of publications is definietly increasing, with especially a rather large increase after around 2018/2019, and then a very small decline in number of publications. Generally the most promminent topic of “Central banking” is like a underlying base with rathjer little fluctuations, even when iothjer topics see a upwards or downwards trend. Furthermore we can see that the topic distribution is getting more diverse, at the beginning of the 2000s “Central Banking”, “Monetary Policy” and “Policies” are outweighing by far all other topics.

[OVERALL DIST, stacked and normal] - Note climaet change is actually top 22, but still in-



cluded



Looking at the topic distribution of "Climate Change" quite intyerstingly the whole topic first appeared in in October in 2018 with the speech "Ten years after the crisis – risks, rules and supervision" held by Sabine Lautenschläger. After this first speech there is quite a significant increase in climate change related publicationsm, with a peak in the end of the first quarter of 2021, since then it has a slight downward trend but still on a medium high level.

As already mentioned each publication can be assigned multiple topics at once so it makes inherent sense to look at the topic co-occurence. Since we dont just want to look at the raw coocurance counts, and instead want to see how much two topics co-appear across documents while neutralizing popularity, where the similarity is calculated as

$$cos(i,j) = \frac{\#\text{ docs with both i and j}}{\sqrt{(\#\text{ docs with i}) \cdot (\#\text{ docs with j})}}.$$

Here we can see that for example the most dominant topis "Central banking", "Monetary

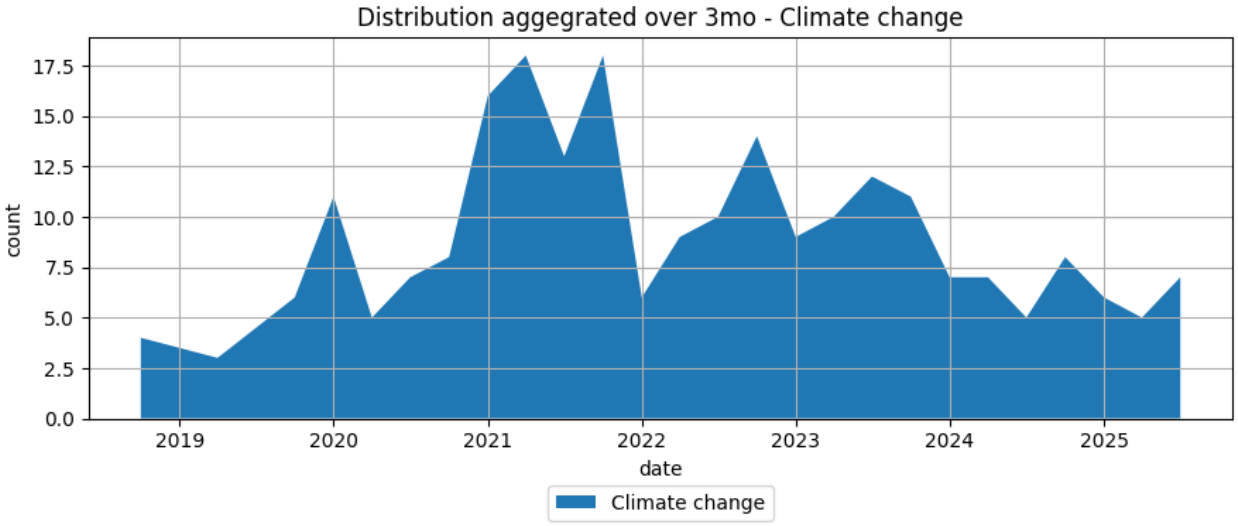


Figure 5: Topic Distribution Aggregated over 3 months for the topic “Climate change”

Policies” and “Policies” are creating a cluster of co-occurance, where “Monetary Policies” and “Policies” are also creating another cluster with “Economic development” and “Inflation”. Other rather logic cooccurances are “Inflation” and “Key ECB interest rates” or “Money” and “Euro”. What is intersting to see ios that there is no such proper cluster formed with the topic of “Climate change” which indicates that it can be a backghriound topic with coocures with the other topics in a similar manner.

Another interesting part to diove deeper is how different authors contributed to different topics, meaning what are the topics authors focused on more and by whihc authors certaion topics are shaped. We can see for example Benoit Coure is the main cotributor to the topic of “Benchmark rates”, while he is the main contributor there overall, he also has a lot of contributions in “Policy” topics or “Euro area” topic. Again, similar to the other findings, we can see that climate change is also a topic driven by multiple authors. Christin Lagard and Frank Elderson seem to be one of the main contributors, while both not limiting their publication to this topic.

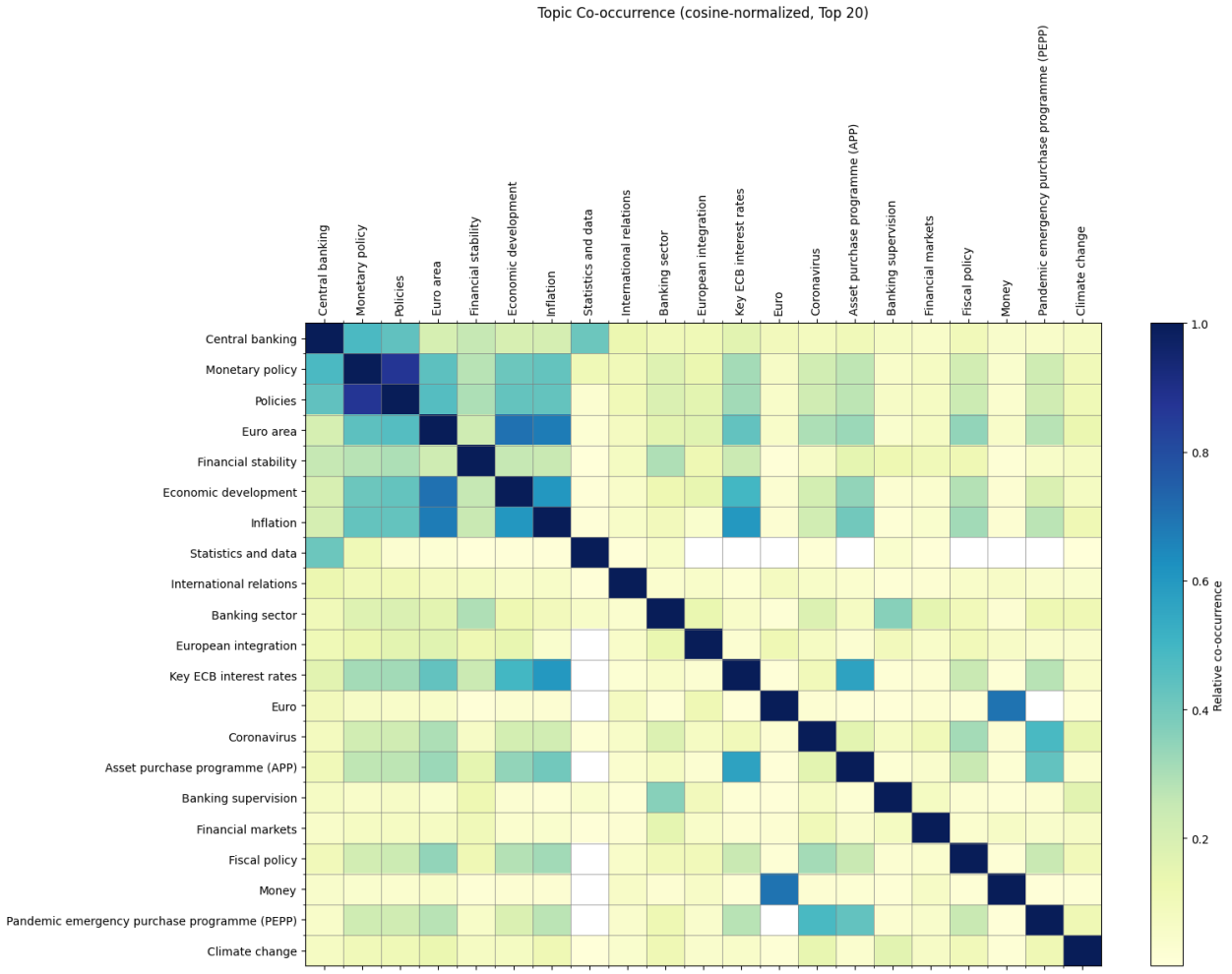
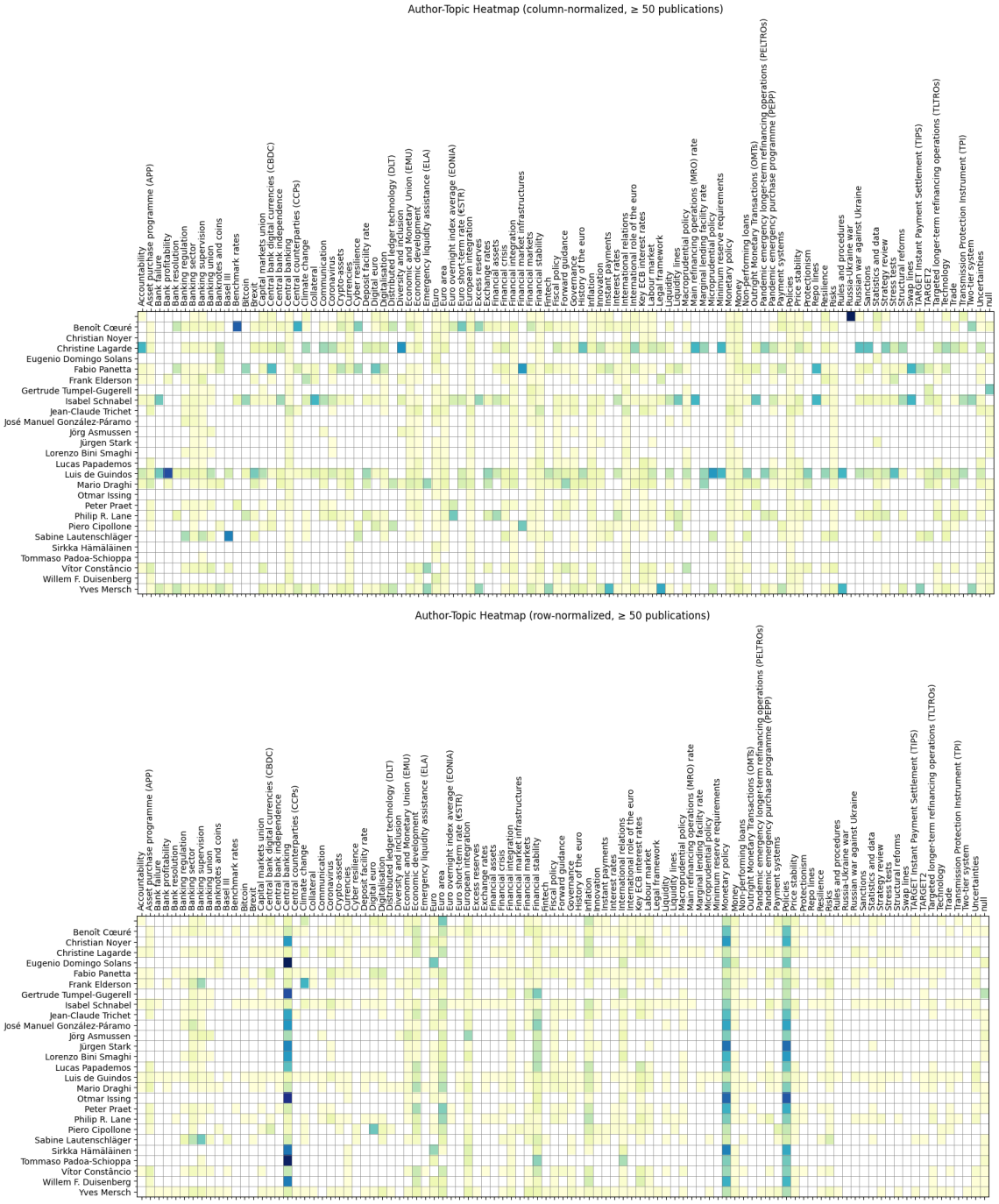


Figure 6: Topic Co-occurrence (cosine-normalized) of the Top 20 topic ("Climate change" is top 22)



Summarizing the results regarding the topic distributions, we can see that topics dynamically change over the time in a certain degree, having some underlying base rate for some topics, while other topics appear and disappear in the pulications along the lines. Climate change being one of them, which started to be apparent since 2018, and since then gaining more importance in general. Furhtermore we can see that there are not a lot of authors who are just specializing in a single topic, but most of the time its a collaboration between authors,

even though they might not work ona single publication together, they shape the topic in as a whole.

Change Point Detection

In this chapter we will now investigate if those findings of the previous chapter regarding the topic distriobution can be proven with statistical methods. More specifially we noe investigate if and at which time can we detect statistical valid change points.

We model the sequence of publication times mentioning the topic "Climate change" as a (piecewise-constant) Poisson process. Under the null hypothesis of a constant rate λ (no changepoint), the inter-arrival times are i.i.d. exponential and, conditional on the total observation window, the ordered arrival times behave like order statistics of a Uniform(0,1) sample. Following Galeano (2007), we test for deviations from this null using a CUSUM-type statistic on the (rescaled) event times.

Let $0 < T_1 < \cdots < T_n = T$ be the arrival times (in days) since the first observed event (first pulication of the topic climate change). Define the rescaled times $U_i = T_i/T$ and the CUSUM deviations

$$D_i \;=\; \sqrt{n}\,(U_i - \frac{i}{n}), \qquad i = 1, \ldots, n.$$

Under the null (homogeneous Poisson), $\{D_i\}$ converges to a Brownian bridge and the test statistic

$$\Lambda_{\max} \;=\; \max_{1 \le i \le n} |D_i|$$

has a known null distribution that we approximate by Monte-Carlo simulation (in code: `galeano_crit_value(n, alpha, nsim)`). We reject "no change" if $\Lambda_{\max} > c_{1-\alpha}(n)$, where $c_{1-\alpha}(n)$ is the simulated critical value at level α.

To locate the change, we take the index $\hat{\tau} = \arg\max_i |D_i|$ as the estimated changepoint. Multiple changepoints are obtained via binary segmentation: split at $\hat{\tau}$, then recurse on the left/right segments, applying the same test and controlling the family-wise error by gradually tightening the per-split level (Bonferroni-style).

Between two detected changepoints we assume a constant rate. The slope of the expected cumulative count within a segment equals the Poisson rate

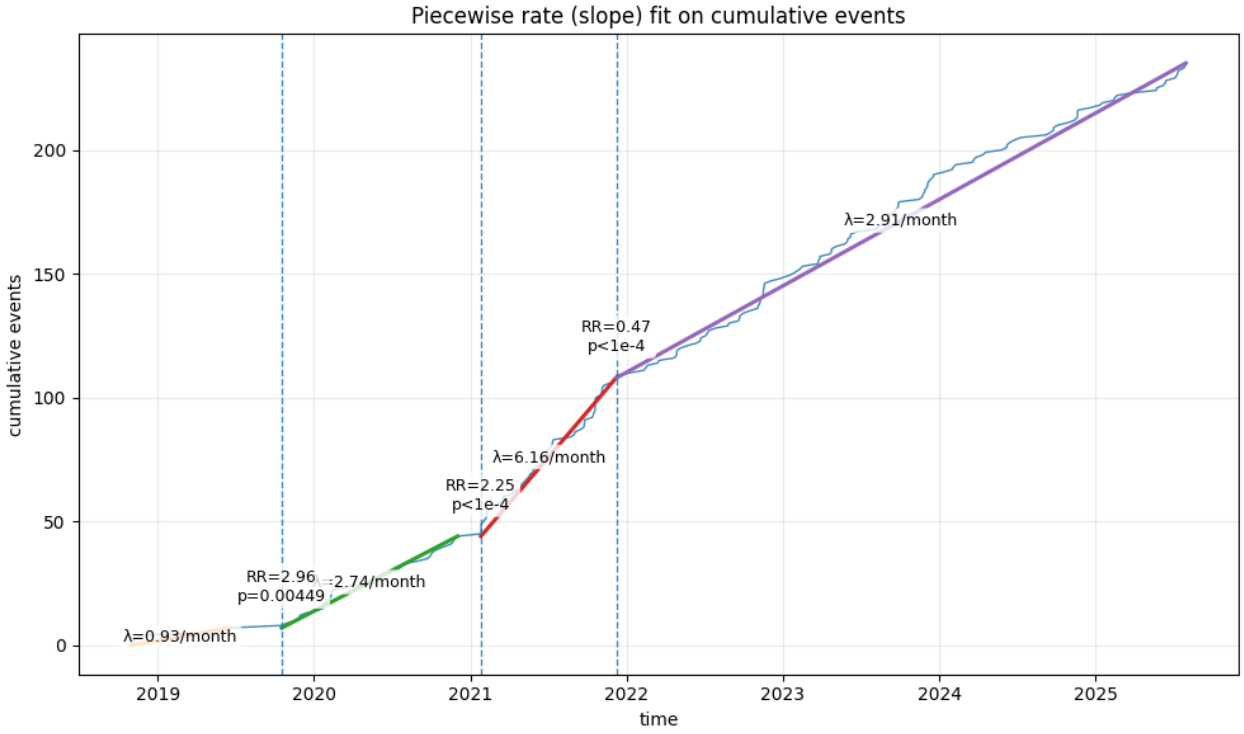
$$\hat{\lambda}_s \;=\; \frac{\#\text{events in segment } s}{\text{exposure time in days}},$$

which we display in the plot as a piecewise-linear fit to the empirical cumulative events. A kink in that line (and a significant $\Lambda_{\max}$) indicates a change in frequency. The effect size at a changepoint is the rate ratio

$$\mathrm{RR} = \frac{\hat{\lambda}_{\text{after}}}{\hat{\lambda}_{\text{before}}},$$

with RR > 1 meaning the topic becomes more frequent. For inference on direction and magnitude, we test the two adjacent segments with an exact Poisson two-sample rate test, reporting RR and a p-value.

The figure shows three statistically significant changepoints that split the series into four regimes: at index 7 (17 Oct 2019), index 44 (25 Jan 2021), and index 108 (08 Dec 2021). In the first three regimes, the estimated arrival rate of "Climate change" publications increases at each break (rate ratios $RR > 1$), indicating progressively higher publication frequency. At the third changepoint (index 108), the rate ratio drops below 1, implying a significant decline in the publication rate relative to the preceding regime.



7. Conclusion and Outlook

This Interdisciplinary project we set out to measure what the ECB communicates with an emphasis on climate, and to identify when its focus shifts by integrating a custom data pipeline, topic labeling, and statistically principled change point detection. Across roughly 9.5

thousand publications that span speeches, blogs, and press material, supervised approaches yielded the most consistent and policy aligned topics. A CatBoost one versus rest classifier trained on text-embedding-3-large embeddings achieved the strongest overall performance across example based and label wise metrics, with a fine tuned XLM-RoBERTa close behind. Unsupervised BERTopic pipelines were informative but lagged on strict multi label criteria, reflecting the difficulty of aligning clusters to a fixed policy taxonomy. Per topic analyses further showed stable recognition of the Climate change label. Co occurrence patterns suggest that climate behaves as a cross cutting theme rather than forming a narrow cluster, which fits its diffusion across policy areas.

The temporal profile of climate related communication is clear. The first appearance in our corpus occurs in October 2018, followed by a rapid expansion that peaks around late Q1 2021 and then moderates. A CUSUM or Galeano style event time analysis isolates three statistically significant structural breaks dated 17 Oct 2019, 25 Jan 2021, and 08 Dec 2021. The first two breaks correspond to increases in the arrival rate of climate related publications, the third to a decrease relative to the immediately preceding regime. These dates should be read as descriptive markers in the communication record rather than as causal attributions.

Methodologically, the project contributes a multi channel ECB corpus with rich metadata and partial native labels, a head to head comparison of unsupervised pipelines with post hoc label mapping versus supervised pipelines on identical splits and metrics, and an event oriented change point workflow with Monte Carlo calibrated critical values and binary segmentation under family wise error control. The pipeline scales to the unlabeled majority of the corpus and preserves reproducibility for downstream time series analysis.

Several limitations remain. Only about 36 percent of items carried native ECB topic tags, so model predictions populate the remainder and may introduce residual mislabeling even after per label F1 threshold calibration. Document level scores aggregate chunk level predictions by a max operator, which may overweight localized mentions and underweight diffuse ones. Post hoc alignment for unsupervised clusters can drift when cluster structure changes over time. Most importantly, the detection of structural breaks provides timestamps, not explanations, and causal interpretation requires external event data, furthermore the what could be done in future research is to test for correlation with other topics which could be an explanation for a lower number of publications in one topic.

The implications are practical. Climate appears as an evolving cross theme within ECB outreach, not a niche confined to a single channel or author. The dated breaks offer anchors for qualitative follow up and for market side event studies that condition on communication regimes. The pipeline can be extended to analyze channel specific or author specific dynamics, and it can incorporate sentiment or framing to capture how climate is discussed, not only how often.

Future work should link the detected breaks to a curated timeline of ECB policy actions, EU initiatives, and salient macro shocks, and should test channel and author heterogeneity explicitly. Human evaluation on a stratified sample would help assess topic faithfulness and

boundary cases, complementing automated metrics. Finally, porting the workflow to other central banks would allow comparative analysis of institutional heterogeneity and the diffusion of climate discourse across the Eurosystem and beyond.

In sum, a reproducible, measurement first approach that pairs supervised topic labeling on strong embeddings with event time change point detection yields interpretable themes and dated structural breaks in ECB communication. This provides a transparent foundation for subsequent causal and comparative work on how central banks speak about climate and how that speech evolves over time.

Arseneau, D. M., Drexler, A., & Osada, M. (2022). *Central bank communication about climate change* (Finance and Economics Discussion Series 2022-031). Board of Governors of the Federal Reserve System. https://doi.org/10.17016/FEDS.2022.031

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in knowledge discovery and data mining* (pp. 160–172). Springer Berlin Heidelberg.

Casiraghi, M., & Perez, L. P. (2022). *Central bank communications* [Technical Assistance Handbook Chapter]. International Monetary Fund, Monetary; Capital Markets Department. https://www.imf.org/-/media/Files/Publications/Miscellaneous/English/2022/mcm-technical-assistance-handbook/central-bank-communications.ashx

European Central Bank. (2025). *Guiding principles for external communication for high-level officials of the European Central Bank* [Policy Document]. European Central Bank. https://www.ecb.europa.eu/ecb/our-values/transparency/html/eb-communications-guidelines.en.html

Fallah, H., Bellot, P., Bruno, E., & Murisasco, E. (2022). Adapting transformers for multi-label text classification. *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE)*, 1–14. https://ceur-ws.org/Vol-3178/CIRCLE_2022_paper_07.pdf

Fortes, R., & Le Guenedal, T. (2021). Tracking ECB's communication: Perspectives and implications for financial markets. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3791244

Galeano, P. (2007). The use of cumulative sums for detection of changepoints in the rate parameter of a poisson process. *Computational Statistics & Data Analysis*, *51*(12), 6151–6165. https://ideas.repec.org/a/eee/csdana/v51y2007i12p6151-6165.html

Hansson, M. (2021). *Evolution of topics in central bank speech communication*. https://arxiv.org/abs/2109.10058

Jurafsky, D., & Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, with language models* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/

Kaminskas, R., & Jurkšas, L. (2024). ECB communication sentiments: How do they relate to the economic environment and financial markets? *Journal of Economics and Business*, *131*, 106198. https://doi.org/https://doi.org/10.1016/j.jeconbus.2024.106198

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590–1598. https://arxiv.org/abs/1101.1438

Kuo, C. (2023). *The handbook of NLP with gensim: Leverage topic modeling to uncover hidden patterns, themes, and valuable insights within textual data*. Packt Publishing. https://books.google.at/books?id=LJPeEAAAQBAJ

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform manifold approximation and projection for dimension reduction*. https://arxiv.org/abs/1802.03426

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). *CatBoost: Unbiased boosting with categorical features*. https://arxiv.org/abs/1706.09516

Shmuel, A., Glickman, O., & Lazebnik, T. (2024). *A comprehensive benchmark of machine and deep learning across diverse tabular datasets*. https://arxiv.org/abs/2408.14817