Investment Institute

WORKING PAPER 173 | MARCH 2025

Topic Modeling with AI Tools

Amundi
Investment Solutions

Trust must be earned

Topic Modeling with AI Tools

Abstract

Sonja TILLY
Amundi Investment Institute
sonja.tilly@amundi.com

Imad AMRI
Amundi Investment Institute
imad.amri@amundi.com

Théo LE GUENEDAL
Amundi Technology
theo.leguenedal-ext@amundi.com

Sofia SAKOUT
Amundi Technology
sofia.sakout@amundi.com

Takaya SEKINE
Amundi Investment Institute
takaya.sekine@amundi.com

We present a robust topic modeling framework that mitigates overfitting while capturing the evolving nature of discourse. Our approach proposes a dynamic, temporal network-based model that adapts to emerging topics while maintaining semantic stability through alignment with predefined themes. To complement this adaptability, we introduce a static novelty detection method, catering to audiences favoring consistent topic structures. By balancing flexibility and stability, our framework enhances the interpretability and reliability of topic modeling in dynamic environments. Our findings significantly improve the capacity to monitor and analyze novel topics across various applications, including news tracking and social media analysis, ultimately providing a more robust framework for understanding the evolution of textual data over time. We apply our analysis framework on the coverage by brokers of the French snap elections of 2024.

Keywords: Topic modeling, novelty detection, NLP, AI

JEL classification: D83, D85, G14.

Acknowledgement

The authors are very grateful to Rami Mery from Amundi Technology, Dorianne Lucius, Amina Cherief, Jiali Xu and Frédéric Lepetit from Amundi and Nick Wade from Northfield for their helpful comments. The opinions expressed in this research are those of the authors and are not meant to represent the opinions or official positions of Amundi Investment Solutions.

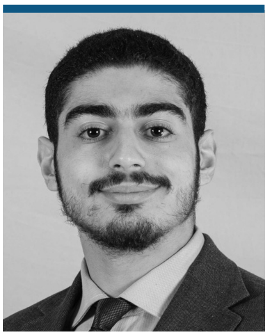
About the authors

Sonja TILLY

Sonja Tilly, PhD, CFA joined Amundi as a Quantitative Analyst in 2024. Commencing her career in 2008 in London, Sonja started out as an Investment Analyst, then transitioning to Quantitative Analyst roles. Her time at Stanhope Capital, Aberdeen Asset Management, and Hiscox honed her skills in investment analysis, financial asset modelling, economic scenario development, and stress-testing portfolios.

While working as Quantitative Researcher at Qoniam Asset Management, Sonja created a filtering method based on deep learning for processing extensive news text, extracting signals that were transformed into a systematic equity trading strategy. Sonja's experience working in traditional Finance is complemented by her insights into the crypto space, gained as Quantitative Researcher at decentralised finance start-up Allora Network, where she led the creation of loan terms for fully automated NFT-backed on-chain loans.

Sonja holds a PhD in Computer Science from University College London. Her research focuses on the impact of news narrative on the economy and financial markets, blending methodologies from data science and econometrics. Further, she is a CFA Charterholder.



Imad AMRI

Imad Amri joined Amundi in July 2024 as an intern in the Quant Portfolio Strategy team of Amundi Investment Institute. His work focuses on applying AI tools to financial markets, particularly in topic modelling and information network analysis. He has worked on supply chain mapping using NLP and novel topic identification, extracting insights from financial reports and news databases.

Imad is an engineering student at CentraleSupélec. He has a strong background in applied mathematics, statistics, and machine learning, complemented by research experience in numerical optimization and risk modeling. He also holds a Bachelor's degree in Applied Economics from Université Paris-Dauphine, PSL.

Théo LE GUENEDAL

Théo Le Guenedal is the Head of Prospective and Quantitative Solutions at the Innovation Lab of Amundi Technology. Prior to this, he worked in the Quantitative Research department of the Amundi Institute since 2018, starting with a project on the performance of ESG investing in the equity market. Since then, he has been involved in an extensive research project on incorporating ESG factors, alternative signals and climate risks into asset allocation strategies. In 2020, he co-authored a paper titled “Credit Risk Sensitivity to Carbon Price,” which was awarded the GRASFI Best Paper Prize for Research on Climate Finance, a prestigious honor sponsored by Imperial College London. He also made significant contributions to the academic field of physical risk assessment by developing the Tropical Cyclone Generation Algorithm. Théo completed his Ph.D. thesis, “Financial Modeling of Climate-related Risks” in Applied Mathematics at the Institut Polytechnique in December 2023, covering both transition and physical risks. Recently, he has focused on integrating advanced climate metrics, stress tests, and analytics into investment tools at Amundi Technology’s Innovation Lab.

Sofia SAKOUT

Sofia Sakout, PhD is Lead Data Scientist in Natural Language Processing and Generative AI within the Innovation Lab of Amundi Technology. Sofia plays a central role in harnessing open source text data and in the development of NLP pipelines, with a particular focus on question-answering (QA) systems. Her focus at Amundi Technology is to channel innovation, data science, artificial intelligence toward practical applications in Amundi.

Sofia holds an engineering degree from École Mohammadia d’Ingénieurs, a Master’s degree from École des Ponts ParisTech, and a PhD from the Sciences, Ingénierie et Environnement Doctoral School in partnership with École Polytechnique.

Takaya SEKINE

Takaya Sekine, CFA is the Deputy Head of Quant Portfolio Strategy within Amundi Investment Institute (formerly known as the Quantitative Research Team of Amundi). In this role, he works on the practical implementation of quant research, artificial intelligence and alternative data for investment strategies.

He joined Amundi in 2000 and is in his current position since July 2018. Prior to that, he was Deputy CIO at Amundi Japan (between 2011 and 2018) with a focus on global quantitative strategies, Head of Index and Multi-Strategies at Amundi Japan (between 2010 and 2011), Fund Manager (between 2007 and 2010) and Financial Engineer (between 2001 and 2007). He has been involved in macro and policy related investment strategies for both retail and institutional clients. Takaya began his career as an IT Manager at Amundi Japan's predecessor company (between 2000 and 2001).

Takaya is a CFA charterholder since 2005. He received the Ingénieur Civil des Mines degree from Ecole des Mines de Nancy in 2000.

Key takeaways

- We introduce a resilient topic modeling framework designed to mitigate overfitting.
- Our approach employs a temporal, network-based model that adapts to emerging topics while aligning them with predefined themes, ensuring both semantic stability and adaptability.
- To accommodate audiences preferring consistent topics, we complement our dynamic model with a static method for novelty detection.

1 Introduction

Understanding how public discourse evolves during elections is crucial, as narratives can quickly shift and influence capital markets and events, such as political outcomes (Keynes, 1937; Shiller, 2017). The detection of novel news narratives has been extensively researched. Topic novelty can be defined as the emergence of new and distinct themes within a corpus of information that deviate from prior observations (Aggarwal & Yu, 2001). The concept is particularly relevant in dynamic systems, such as social media and news streams where identifying new topics is essential for staying updated on emerging trends. To enhance the exploration of novel topics, researchers advocate shifting from analyzing individual news narratives to examining their meta-representations (Newell, 1972). Understanding the meta-structure of a large corpus of news articles can unveil hidden patterns or insights. Traditional methods for identifying the structure of news content often rely on manual categorization, keyword tagging (Westerski *et al.*, 2013), or human similarity judgments (Richie & Bhatia, 2021).

Artificial intelligence [AI] has from its origins the aim to simulate elements of intelligence with machines. We can find a reprint of the 1955 Dartmouth research project in (McCarthy *et al.*, 2006). In asset management, AI is inherently tied to the depth of business intelligence it leverages. The latest advancements are built on augmented knowledge frameworks designed to adapt to an evolving financial landscape. A fundamental step toward true intelligence lies in the ability to autonomously detect trends-differentiating novelty from pre-existing patterns–rather than focusing solely on advanced retrieval or restitution techniques. Consequently, integrating a robust assessment of novelty is critical to the development of AI-assisted tools, ensuring they provide meaningful insights and drive decision-making in asset management.

Traditional news analysis often focuses on static snapshots, overlooking temporal trends. By combining topic modeling, temporal networks, and sentiment analysis, this study compares and contrasts two novel approaches to monitoring election narratives, providing valuable insights for analysts and researchers tracking narrative dynamics over time.

This paper adds to the existing body of literature on novelty detection in narrative by proposing a methodological framework leveraging topic modelling and temporal networks. For interpretability purposes, we relate evolving topics back to overarching themes, akin to themes in the Global Database of Events, Language and Tone (GDELT) ("GDELT Project", 2015). Such themes derived from news narrative can improve financial and macroeconomic forecasts (Thorsrud, 2016; Tilly & Livan, 2021). Our framework can be applied to large volumes of temporal narrative, which allows the tracking and large-scale analysis of novelty and news evolution in real-time. Our work lies at the intersection between dynamic topic modeling and graph theory which is, to the best of our knowledge, a very much understudied area of application.

This paper is structured as follows. Section 2 reviews existing works. We describe the data set used in our research in Section 3. Section 4 provides an overview of the approaches we adopt in this paper. Sections 5 and 6 outline the two approaches we employ for novelty detection. Section 7 compares and contrasts the results from the two novelty detection methods introduced in the prior sections. Finally, section 8 concludes the paper.

2 Related work

There is a broad body of literature on the emergence and identification of new topics in news narratives. More recent works analyze central bank communication and use expert knowledge to categorize groups of words extracted with the Latent Dirichlet Allocation into topics (Blei *et al.*, 2003; Fortes & Guenedal, 2021). Likewise, Blanqué *et al.* (2022) and Cherief *et al.* (2025) use custom narratives on biodiversity and their relation with financial assets rely on a definition of narratives and themes defined with expert knowledge from the aggregation of topical events which rely on taxonomies such as the World Bank Group Topical Taxonomy (World Bank, 2016). Advances in AI-based language models allow the systematic classification of news texts based on their semantic content.

Research analyzing large volumes of news articles utilizes bag-of-words models, representing semantic context using term frequencies or topics (Xu *et al.*, 2019), or text embeddings (Hao *et al.*, 2017; Kenter & De Rijke, 2015; Tang *et al.*, 2015). Text embeddings numerically encode textual content, allowing to compare semantic similarity (Muennighoff *et al.*, 2022).

To identify novelty, earlier research combines text embeddings with distance-based novelty detection methods such as k-nearest neighbors (Ramaswamy *et al.*, 2000), local outlier factor (Breunig *et al.*, 2000), a one-class Support Vector Machine model (SVM) Mũnoz-Marí *et al.* (2010) and Support Vector Data Description Tax and Duin (2004).

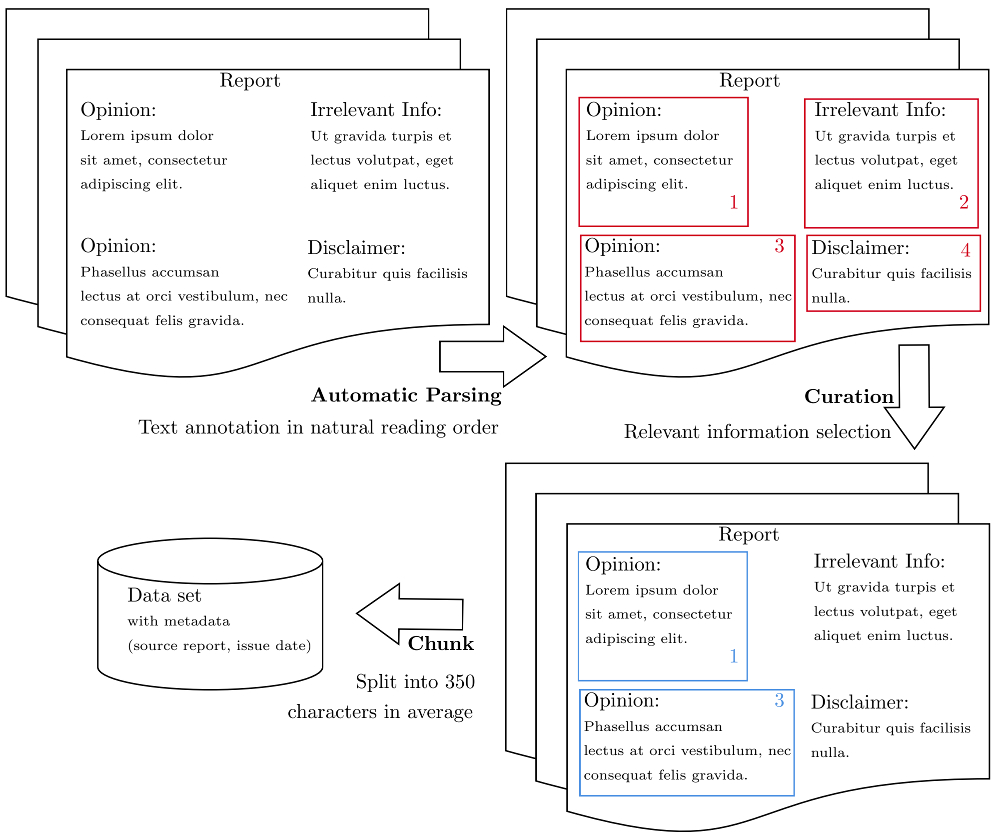
More recent papers incorporate neural networks and in particular transformer models to identify novelty and emerging trends in narrative. For instance, Ahmed and Courville (2020) proposes out-of-distribution detection to identify semantic distribution shifts in a specified context. Barnes *et al.* (2024) look at interconnected topic dynamics between news items, using Latent Dirichlet Allocation and BERTopic to create a set of topics that is then compared with a large number of image with text memes on Reddit. The authors find that topicality features have explanatory power when linking New York Times topics to Reddit. Ghosal *et al.* (2022) propose semantic novelty detection on document level, checking if a text is redundant against a number of given priors. Kumar and Bhatia (2020) introduce a high speed novelty detection mechanism that can be incorporated in a web crawler architecture. It summarizes a text and calculates similarity based on ontology using wordnet 3.0. It then derives a hash value that is matched with others to calculate a similarity index. Novelty text is identified given a set threshold. Ma *et al.* (2022) propose a entity-level novelty detection method. The authors employ an attention model to train it on entity-level background knowledge using the Wikipedia corpus, permitting entity pair relation classification and subsequent novelty scoring [1]. Boutaleb *et al.* (2024) present a mechanism for emerging trend detection based on BERTopic. The authors introduce a new metric that quantifies topic popularity considering the number of documents that belong to a topic over time, resulting in noise, weak and strong signals.

3 The Data Set

Our data set consists of chunks of text from broker reports originating from multiple brokers, concentrated around the period of the legislative snap-elections in France in 2024. As illustrated in Figure 1, we parse the original broker reports. Once the documents are parsed, we curate the proposed blocks of text. If we deem areas are irrelevant to our analysis because they refer for instance to specific disclosures of the broker report, we will not retain it. Then we split the text into chunks of about 350 characters on average, ensuring a granular breakdown that facilitates the extraction of key information. We assign an issue date to each of the 1565 chunks of our data set to preserve the temporal context.

[1] For the definition of named entity, see https://en.wikipedia.org/Named_entity.

Figure 1: From broker reports to our data set



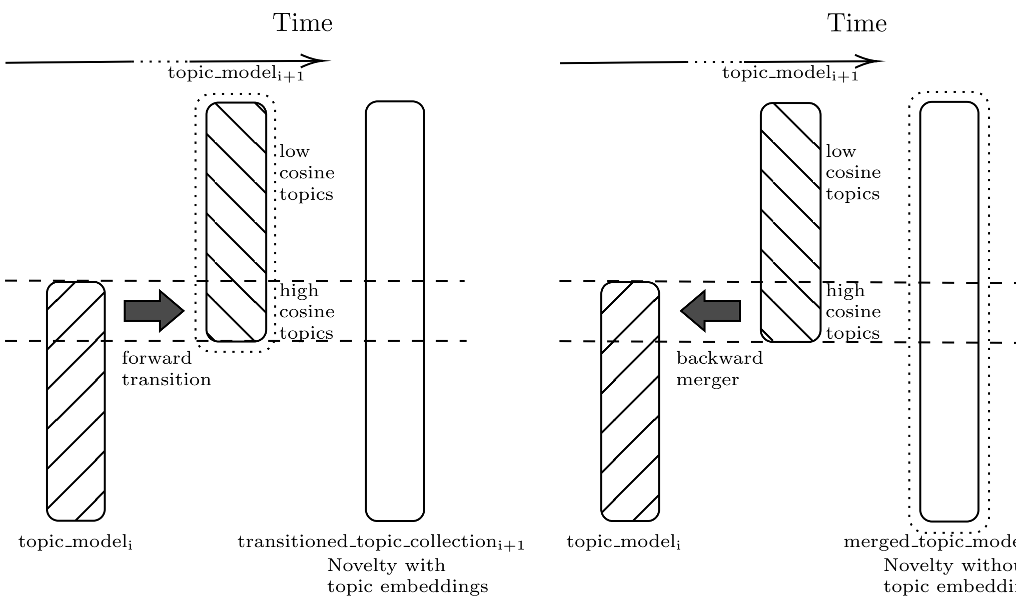
Source: Amundi Investment Institute

4 Methodology

4.1 Approach Overview

Figure 2 illustrates the methodology applied in this paper, and Algorithm 1 provides pseudocode outline of our framework.

Figure 2: Two different approaches to Topic Novelty



(a) the Forward Transition

(b) the Backward Merge

Source: Amundi Investment Institute.

Our novel methodology systematically selects and applies a representative model for topic detection over time, ensuring coherence and adaptability in dynamic datasets, directly nourished by changing news and terminologies. Egger and Yu (2022) compare topic modeling techniques by analyzing the results for different approaches; Latent Dirichlet Allocation (LDA, a generative process approach), Non-negative Matrix Factorization (NMF, a linear-algebraic approach), BERTopic[2] and Top2Vec (the latter two approaches use word embeddings). The authors find that BERTopic performs well on their short and unstructured set of tweets, leveraging the ability to consider contextual information.

In this paper we choose to model topics with BERTopic. The process is as follows. Initially, the most representative model is identified through training and selection (for details, see section 6). For each time point, this model is applied to detect topics, enforcing a minimum threshold of three topics to maintain a meaningful topic extraction. If the detected number of topics meets or exceed this threshold, the model is directly used; otherwise, additional data points are incorporated iteratively until the threshold is satisfied. To handle topic transitions, two strategies are implemented simultaneously: a forward transition approach, which ensures temporal continuity by evaluating embedding similarity between consecutive

[2]BERTopic uses a combination of BERT embeddings and clustering techniques

time points and a backward merging approach, which retains only the top three novel topics based on predefined novelty criteria. The topics labeled -1 by BERTopic, representing outliers, are systematically excluded from the transition process (Grootendorst, 2022).

The detection and tracking of topics over time is a fundamental challenge in topic modeling and natural language processing. Traditional approaches, which maintain a static set of topics and only update this set under strict criteria, offer stability but suffer from the persistence of outdated or irrelevant topics, leading to a bloated and increasingly noisy topic set. Conversely, more dynamic approaches, such as those leveraging embedding-based similarity, allow topics to evolve and adapt as new data emerges. However, this flexibility comes at the cost of continuity: topics may drift, split, or merge, making it difficult to trace their evolution over time.

This trade-off between stability and adaptability highlights the need for an integrated approach that balances topic evolution with continuity. To address the lack of continuity in dynamic methods, we propose leveraging zero-shot classification to map evolving topics back to predefined, interpretable themes. In doing so, we aim to maintain the benefits of dynamic topic modeling flexibility and responsiveness while introducing a layer of semantic stability that allows for better tracking, analysis, and interpretability of topic changes. We apply our frameworks to the example of the French snap election that took place in summer 2024.

Algorithm 1 High Level Approach

```
Train model: select most representative model
for every date do
    Most representative model
    Check minimum topics: min 3
    if number of topics ≥ 3 then
        Use most representative model
    else
        Handle case for topics less than 3: consider current and subsequent data point(s)
        until number of topics is sufficient
    end if
end for
Transition Approach[3]:
    a) Forward transition: Criteria based on embedding similarity.
    b) Backward merge: Criteria on novelty (top 3).
```

5 Dynamic Implementation

5.1 Tracking Topic Evolution

This section outlines a dynamic approach of tracking topic evolution and novelty leveraging narrative and temporal networks. To capture the evolution of topics between two dates, we use the BERTopic framework to generate interpretable topic models. For each date, we train an independent BERTopic model on the corresponding subset of news articles, producing a set of topics with associated representations and document probabilities. This allows us to

[3] Both approaches a) and b) exclude topics that are labeled -1 as they represent outliers and should typically be ignored (Grootendorst, 2022).

capture the nuances of daily narratives, highlighting how new topics emerge and others fade across time.

It is imperative to acknowledge the inherent pseudo-randomness in the BERTopic library, which necessitates the implementation of a robustness procedure. This procedure entails the training of 500 models for each date. Subsequently, Gaussian Mixture Models are employed to cluster the topic embeddings derived from these models (see eq. 3). The optimal number of clusters is determined using eq. 5. For each cluster, the distance of each topic embedding from the centroid is calculated. The model with the topic embeddings that are, on average, closest to their respective cluster centroids is selected as the most representative model. If a retained most representative model contains less than three topics, we consider that date together with the subsequent date. This is a compromise, allowing a minimum level of topic diversity while not relinquishing any information. In Figure 6, we measure the percentage of outlier topics labeled -1 on subsequent dates generating a model with the documents starting from our initial two days of data. As we spike at thirty-two percent of such topics, we consider that applying the above-outlined robustness procedure returns well-fitted models.

We then create a temporal network using the topics generated by BERTopic. Temporal networks are effective structures to study and analyze the evolution and propagation of topics and events extracted from news narrative (Bögel & Gertz, 2015; Li *et al.*, 2018; Si *et al.*, 2020). Temporal networks are networks whose topology evolves over time. Two nodes in a temporal network are linked at a discrete time step only if they have a connection at that time (Holme & Saramäki, 2012).

In our network model nodes represent individual topics for a specific day. Each node v_i represents a topic. Let T_i be the topic represented by node v_i, e_{T_i} be the embedding of topic T_i at a specific time step. Thus, each node v_i can be represented by $v_i = T_i$.

Edges indicate transitions between topics across consecutive days, based on their semantic similarity. An edge e_{ij} between nodes v_i and v_j represents a transition beetween topics T_i and T_j across time steps. Let $e_{ij} \in$ E, where E is the set of all edges, w_{ij} be the weight of the edge, based on the cosine similarity between e_{T_i} and e_{T_j}. Then, the edge e_{ij} is given by $e_{ij} = (v_i, v_j, w_{ij})$.

We use cosine similarity between topic embeddings of two time steps to quantify the relationship between topics across two consecutive days. The cosine similarity is utilized to calculate the cosine of the angle between two embedding vectors e_{T_i} and e_{T_j} in a multidimensional space (Kenter & De Rijke, 2015; Thongtan & Phienthrakul, 2019). It is given by

$$\text{Cosine Similarity} = \frac{e_{T_i} \cdot e_{T_j}}{\|e_{T_i}\|\|e_{T_j}\|} \tag{1}$$

where numerator is the dot product and the denominator stands for the euclidean norm of the embeddings.

A threshold value is applied to the cosine similarity scores to ensure that only meaningful topic transitions are included as network edges. Only topics that are in the 90^{th} percentile of similarities are taken into account for transitions. The transitions are converted into transition probabilities by dividing each element by its respective row sum to standardize edge weights and facilitate analysis.

Further, we utilize VADER sentiment analysis on the corpus of documents at each time step to gauge the emotional tone of each topic (Hutto & Gilbert, 2014). We are aware of the literature on the vulnerabilities of the dictionary-based sentiment models (Leippold, 2023) but we will not address the subject in this paper. In prior research, we have adapted alternative strategies with fine-tuned models (Stagnol *et al.*, 2023). The sentiment scores extracted enable us to monitor sentiment shifts over time, revealing nuanced changes in discussions on focal points like candidates and policies across the election cycle. This method allows for the highlighting of the flow of narratives over time, including merges, splits, or the emergence of new topics and their associated sentiment.

To detect novel topics in a temporal topic network, we apply two conditions. First, we identify nodes with no in-degree. They represent topics that emerge without any prior transition. As a second condition for novelty detection, we cluster all topic embeddings up to a given date. The purpose of the clustering is to determine if a specific topics has been present previously, which would negate novelty. Topic embeddings are aggregated from the time step 1 to time step t. At each time step, these aggregated embeddings are clustered using Gaussian Mixture Models[GMM]. GMM is a soft clustering technique that is used to derive the probability that any given data point belongs to a cluster (Abbas, 2008).

A Gaussian, or Normal distribution, is a continuous probability distribution defined as

$$N(X|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right) \tag{2}$$

where μ represents a D-dimensional mean vector, Σ is a $D \times D$ covariance matrix that describes the structure of the Gaussian distribution, and $|\Sigma|$ stands for the determinant of Σ. A single Gaussian distribution, described by its mean and standard deviation, cannot adequately capture the complexity of multimodal datasets.

Gaussian Mixture Models are a blend of individual Gaussian probability distributions, defined as:

$$p(X) = \sum_{k=1}^{K} \pi_k N(X|\mu_k, \Sigma_k) \tag{3}$$

where K stands for the number of components in the model and π_k represents the mixing coefficient estimating the density of each Gaussian component. The Gaussian density $N(X|\mu_k,\Sigma_k)$ characterizes a component within the mixture model, with each component k defined by a Gaussian distribution featuring mean μ_k, covariance Σ_k, and mixing coefficient π_k.

Considering a set of N independent and identically distributed observations $\{x_1, x_2, ..., x_N\}$, the log likelihood function for a Gaussian mixture is given by

$$\ln p(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\Sigma_k)\right) \tag{4}$$

The Expectation-Maximization (EM) algorithm provides maximum likelihood estimates for the Gaussian Mixture model concerning the mean vector μ, covariance matrix Σ, and mixing coefficients π as defined in Equation 3 (Reynolds *et al.*, 2009).

GMMs, through the Expectation-Maximization algorithm, form clusters with ellipsoidal shapes based on probability density estimations, allowing for the modeling of complex, multimodal distributions (Patel & Kushwaha, 2020). The silhouette score evaluates the quality of clusters in clustering algorithms by measuring how similar an observation is to its assigned cluster compared to other clusters. It is given by

$$\text{Silhouette score} = \frac{1}{N}\sum_{i=1}^{N} \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{5}$$

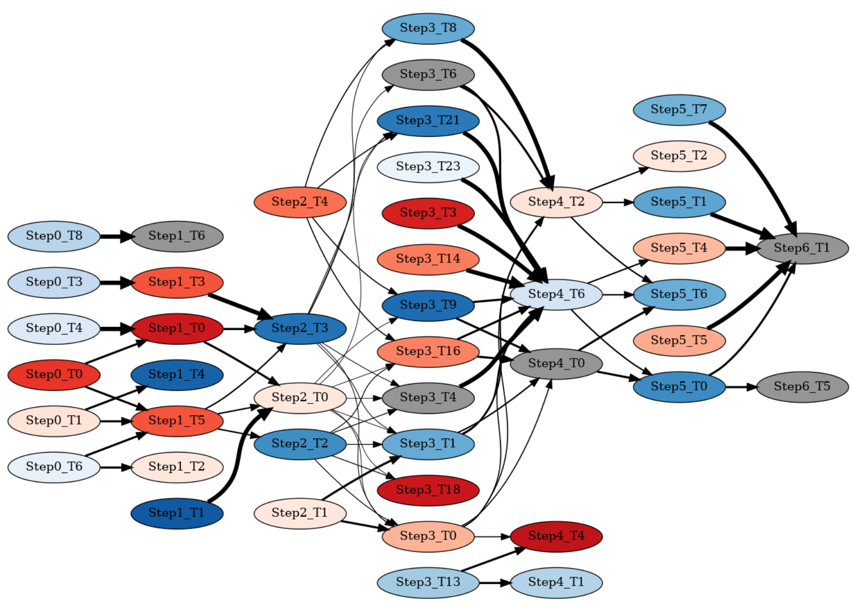
where a_i is the mean distance between a sample and all other data points in the same cluster and b_i is the mean distance between a sample and all other points in the nearest cluster to that sample (Ogbuabor & Ugwoke, 2018; Shahapure & Nicholas, 2020).

For each cluster, topic embeddings with a log-likelihood that falls into the 1st percentile are selected as "novel" topics as they represent the "tail" of the distribution. These "novel" topics signify subjects that have not previously appeared in the corpus, aiding in the identification of unique and potentially significant temporal outliers within the network.

5.2 Illustration of the model

Figure 3 is based on narrative from broker reports issued in the run-up to the French snap elections in 2024. It only shows all topics that transition into other topics on a subsequent time step and omits unconnected topics for illustrative purposes.

Figure 3: Temporal topic graph



Source: Amundi Investment Institute

Topics flow are arranged sequentially, with a chronological progression from the past to the present, that is to say, from left to right. Topics are colored according to sentiment, with blue, red and grey representing positive, negative and neutral sentiment, respectively. The stronger the gradient, the more intense the sentiment. The directed edges represent the transition probabilities between topics of different time steps; the wider the edge, the higher the transition probability.

5.3 Adding interpretability to topic evolution

While the temporal topic network set out in section 5 allows to track the changes in topics and narrative over time, the approach lacks stability in topic labels. To address this lack of continuity, we prompt a LLM (Kocoń *et al.*, 2023) to create the following distinct and well-separated themes based on the topic representations from the BERTopic model incorporating half of the dates considered (see appendix A.1 for prompt details): (*i*) Financial Markets and Instruments; (*ii*) Investment Strategies and Trends; (*iii*) Sector Performance and Comparisons and (*iv*) Banking and Financial Institutions. These themes are then applied to all data points considered, demonstrating that the themes are sufficiently universal and general to maintain relevance in relation to future narrative, akin to themes contained in the Global Knowledge Graph of GDELT (Leetaru & Schrodt, 2013).

To align evolving topics with these predefined themes, we apply zero-shot classification to the topic representations, mapping each topic at each time step to one of the themes. Zero-shot learning is a machine learning paradigm where a model can make predictions on new, unseen classes without having access to any specific training data for them. This is achieved by leveraging knowledge from related tasks or classes that the model has already been trained on. Typically, zero-shot learning relies on a shared semantic representation such as embeddings to link seen and unseen classes (Pushp & Srivastava, 2017).

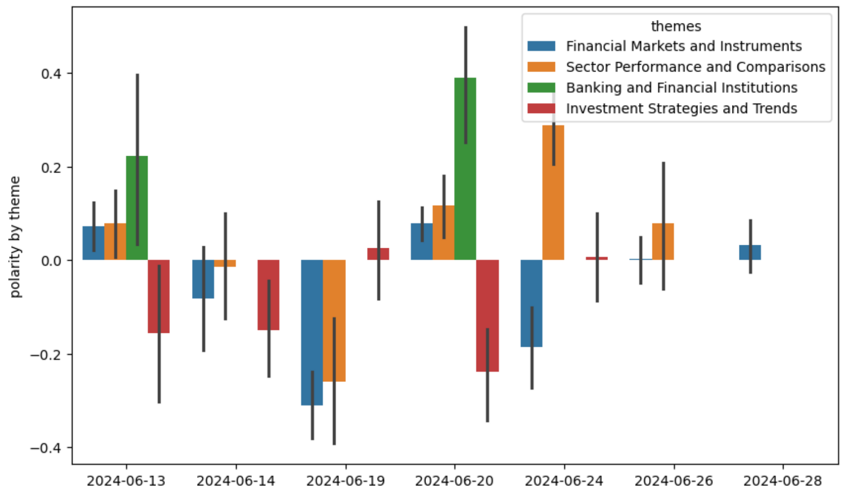
Applying the topic novelty detection method described in Section 5, we identify nine new topics within the corpus (see appendix A.2 for a list of all novel topics).

French president Emmanuel Macron's decision to call a snap election in summer 2024 stirred fears that the political turmoil may spill into the stock market and cause a so-called "contagion" effect. We track the notion of "contagion" in our word corpus. Interestingly, topic representations containing "contagion" occur on step 2 (19/06), 3 (20/06 - 23/06) and 4 (25/06 and 26/06), albeit not as novelty. The last two occurrences of the "contagion" topic are unconnected nodes [4] (i.e. they do not show in Figure 3 as it only shows connected nodes). With the benefit of hindsight, this manifestation of the "contagion" theme is consistent as contagion fears appeared, lingered for a short period of time and then subsided as the French snap elections approached.

Linking evolving topics back to themes provides some continuity over time. Figure 4 illustrates the count of each theme for each time step. Research shows that themes from narrative can improve financial and macroeconomic forecasts (Thorsrud, 2016; Tilly & Livan, 2021).

[4] unconnected nodes do not have incoming or outgoing edges

Figure 4: Aggregate Polarity By Date And Theme



Source: Amundi Investment Institute

6 Identifying Novel N-grams in BERTopic Models

This section describes a "classical" approach to novelty detection. This approach aligns with the investment industry's preference for stable models, as exemplified by the evolution of the Barra risk models. These models, detailed by Kahn (2019) and Menchero *et al.* (2011), are multi-factor models designed to assess the risk and returns of equity portfolios. They provide a structured framework that evolves gradually, mirroring the principles of novelty detection, which relies on identifying deviations within well-defined and stable environments. Just as the Barra risk models aim to capture systematic risks without frequent changes, classical novelty detection approaches emphasize robustness and minimal variation, making them suitable for identifying meaningful departures from established norms. To readers who are not in the investment management industry it might seem un-intuitive to refer to a risk model structure. However agents in the investment industry have a strong anchoring to the Barra risk models, especially if they are exposed to equity universes. In our previous work on economic narratives (Blanqué *et al.*, 2022) and custom narratives (Cherief *et al.*, 2025), we parallel risk models which have risk indices composed of descriptors with narratives which have themes which are themselves composed by GDELT indicators. In this section, we answer to the habit of investment managers to have continuity in the Barra risk models.

We add new topics to our initial model based on the methodology outlined below. This approach is similar to a model that is periodically updated based on a set of criteria. In dynamic text datasets, identifying novel terms that distinguish emerging topics is crucial for understanding the evolution of discourse. BERTopic, provides both topic embeddings and representative documents for each topic (Grootendorst, 2022). Representative documents summarize the core content of topics, offering a rich basis for extracting meaningful n-grams. This section introduces a method for identifying and ranking n-grams based on novelty and importance within new BERTopic models.

Representative Documents and N-gram Extraction Let $\mathcal{T}_i$ represent a topic in a BERTopic model. Each topic $\mathcal{T}_i$ is associated with a set of representative documents $\mathbf{D}_i = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k\}$, where $\mathbf{e}_j$ is a document embedding derived from transformer-based models. Representative documents are selected based on their proximity to the topic centroid in embedding space.

To identify representative terms for each topic, BERTopic employs a class-based TF-IDF approach. For a topic $\mathcal{T}_j$, the c-TF-IDF score for a term w is given by:

$$\text{c-TF-IDF}_{j,w} = \text{TF}_{j,w} \cdot \log\left(\frac{N}{\sum_{k=1}^{K} \nVdash_{\{w \in \mathcal{T}_k\}}}\right) \tag{6}$$

where $\text{TF}_{j,w}$ is the term frequency of w in documents assigned to $\mathcal{T}_j$; N is the total number of documents and $\nVdash_{\{w \in \mathcal{T}_k\}}$: Indicator function that equals 1 if w appears in topic $\mathcal{T}_k$ and 0 otherwise. This formulation ensures that terms frequently occurring in a specific topic, but rarely elsewhere, are given higher importance. The top n terms with the highest c-TF-IDF scores are selected to represent each topic $\mathcal{T}_j$, providing an interpretable summary of the topic.

The top n terms with the highest c-TF-IDF scores are selected to represent each topic $\mathcal{T}_j$, providing an interpretable summary of the topic.

To capture the linguistic essence of each topic, we generate n-grams ($n = 3, 4$) from these representative documents. For a document $\mathbf{d} \in \mathbf{D}_i$, the set of n-grams is defined as:

$$\mathcal{N}^{(i)} = \{ng_1, ng_2, \ldots, ng_m\}, \tag{7}$$

where ng_j represents an individual n-gram.

Defining Novelty in N-grams Given two BERTopic models, $\mathcal{M}_0$ and $\mathcal{M}_1$, trained on corpora $\mathcal{C}_0$ and $\mathcal{C}_1$ respectively, our goal is to identify novel n-grams in the topics $\mathcal{T}_i^{(1)}$ from $\mathcal{M}_1$ that do not appear in any topics $\mathcal{T}_j^{(0)}$ from $\mathcal{M}_0$.

Let $\mathcal{N}^{(i)}$ denote the set of n-grams extracted from representative documents of topic $\mathcal{T}_i$. An n-gram ng is considered novel if:

$$ng \notin \bigcup_{j=1}^{k_0} \mathcal{N}^{(j)}, \tag{8}$$

where $\mathcal{N}^{(j)}$ is the set of n-grams in topic $\mathcal{T}_j^{(0)}$ from $\mathcal{M}_0$. The novelty condition ensures that ng is unique to the topics in $\mathcal{M}_1$.

To quantify the significance of novel n-grams, we compute an importance score that combines the term frequency (TF) with sentiment analysis. For an n-gram $ng \in \mathcal{N}^{(i)}$, the importance score is defined as:

$$\text{Importance}(ng) = \text{TF}_{ng} \cdot |\text{Sentiment}_{ng}|, \tag{9}$$

where TF_{ng} is the term frequency of ng within $\mathcal{D}_i$, the representative documents of $\mathcal{T}_i^{(1)}$ and Sentiment_{ng} is the compounded sentiment score of ng, derived using the VADER sentiment analyzer. The absolute value ensures that both strongly positive and strongly negative terms are emphasized.

Model Update In Figure 5, we measure the novelty of the ngrams contained in the representative documents of $\mathcal{M}_1$ for 19/06/2024. We consider the ngram based novelty measure for a topic to be the minimum of the top 3 TF $\cdot$ |Sentiment| of the ngrams of that topic. The similarity threshold which is required for the merge operation between $\mathcal{M}_0$ and $\mathcal{M}_1$ in BERTopic is dynamically determined so that the top 3 ngram based novelty topics are retained in the merged model. As identified in Table 1, on 19/06/2024, from our novelty analysis, we identify a topic that discusses a possible regional crisis "contagion". The possible adverse effects constitute again a novel topic on 23/06/2024. Finally, the "contagion" to the region is considered overblown on 25/06/2024. On 19/06/2024 and 23/06/2024, the 10Y OAT Bunds spread displayed in Figure 7 does not jump upwards, indicating that the market participants did not price a "contagion" scenario. We display in Figure 6 the percentage of daily items which fall into outlier topics labeled -1. Between a simple BERTopic model calibrated on the text content of 13/06 and 14/06 together and our backward merged model approach. As the backward merged model is partially refit on a regular base, the percentage of outliers is lower during our analysis period.

Figure 5: Top 3 novel n-grams by TF $\cdot$ |Sentiment| score for each topic in $\mathcal{M}_1$ (19/06/2024)

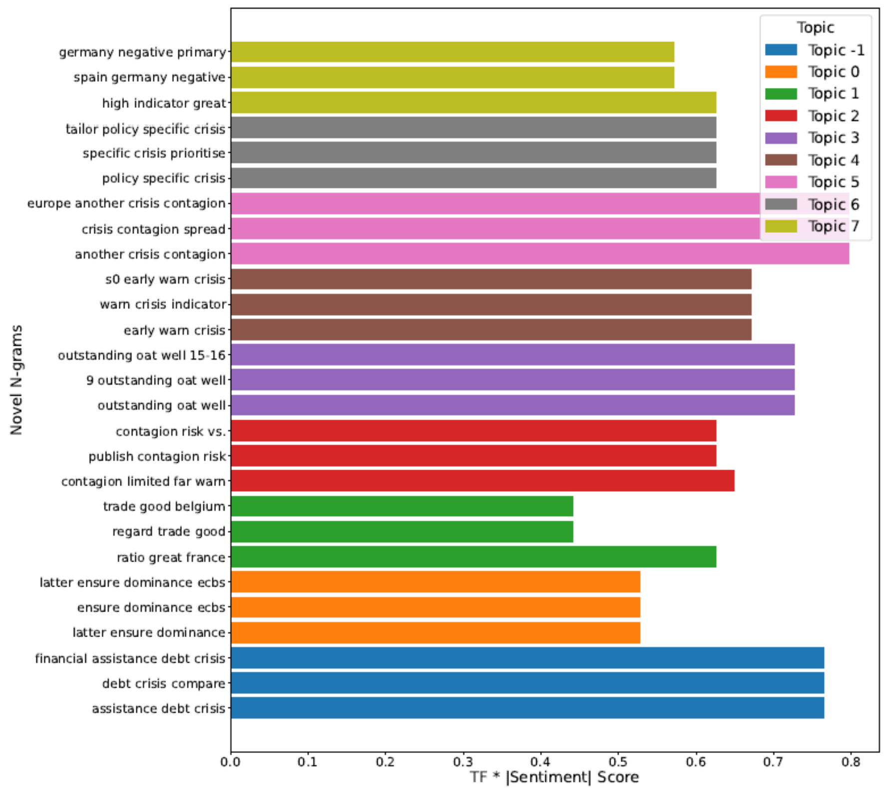


Table 1: Top3 Ngrams for a regional "contagion" topic within the identified novel topics

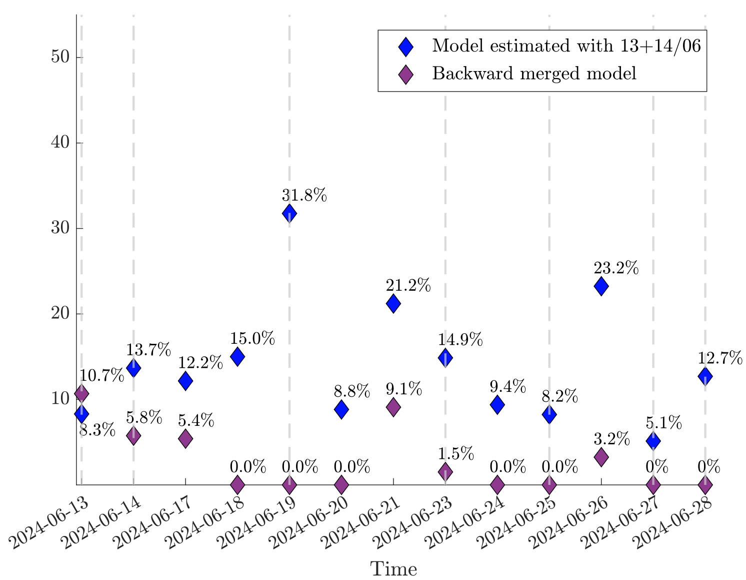
	2024-06-19
Top Ngram 1	another crisis contagion
Top Ngram 2	crisis contagion spread
Top Ngram 3	europe another crisis contagion

	2024-06-23
Top Ngram 1	contagion risk adverse
Top Ngram 2	european contagion risk adverse
Top Ngram 3	contagion risk adverse french

	2024-06-25
Top Ngram 1	regional stress low
Top Ngram 2	stress low last
Top Ngram 3	lead regional stress low

Source: Amundi Investment Institute.

Figure 6: Percentage of outliers



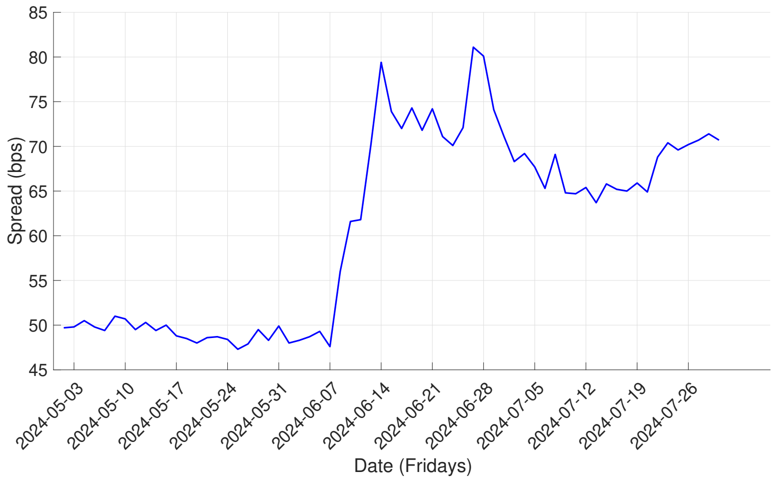
Source: Amundi Investment Institute.

7 Discussion

In this study, we compare a static topic modeling approach, where topics are only updated under strict criteria, with a dynamic approach using topic embedding similarity that allows topics to evolve over time.

Figure 7 shows the 10Y OAT-Bund spread during the period of the French snap elections. The spread increases notably after the announcement of the snap election on 14/06/2024, which is reflected in topics extracted from broker reports containing "uncertainty" and

Figure 7: 10Y OAT Bunds Spread



Source: Amundi Investment Institute

"contagion".

The static approach maintains topic stability, which is useful for tasks requiring consistency over long periods. However, its inability to remove outdated topics results in lower quality as new data emerges. This can lead to topic inflation, where the overall set of topics grows larger over time as new topics are added but become increasingly disconnected from the current content. The difference of the probability distributions for the two approaches measured on the 28/06/2024 by the end of our analysis period is not significant. However we endeavor to apply to Akaike Information Criterion [AIC] to the BERTopic models (Akaike, 1974). The AIC confronts the complexity of a given model against its goodness of fit. It is defined as:

$$\mathcal{AIC} = 2k - 2\ln(\mathcal{L}) \tag{10}$$

where k is the number of topics, the number of words per topic, the assignments of documents to topics and the number of hyperparameters for the vectorisation of the texts and the clustering steps and $\mathcal{L}$ is the likelihood.

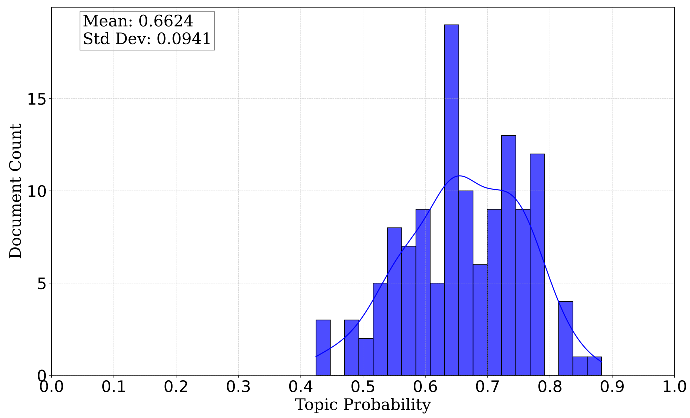
Given that the $\mathcal{AIC}_{\text{Backwards merge model}}$ is 1752.5 vs. 502.9 for the $\mathcal{AIC}_{\text{Unmerged model}}$ on the 28/06/2024, we determine that the models generated by our backwards merger approach are inferior in quality[5] relative to the models used in our forward transition approach.

The dynamic approach utilizing topic embedding similarity effectively captures emerging trends and adapts to shifts in content. However, the lack of topic continuity can make it challenging to track long-term topic evolution or perform consistent comparisons. To address the lack of continuity in the dynamic approach, we apply zero-shot classification to align evolving topics with predefined themes. Zero-shot classification successfully mitigates the continuity issue by mapping evolving topics to predefined, interpretable themes. This approach preserves semantic stability without constraining the flexibility of dynamic approaches and demonstrates the potential to balance adaptability and interpretability in dynamic topic modeling.

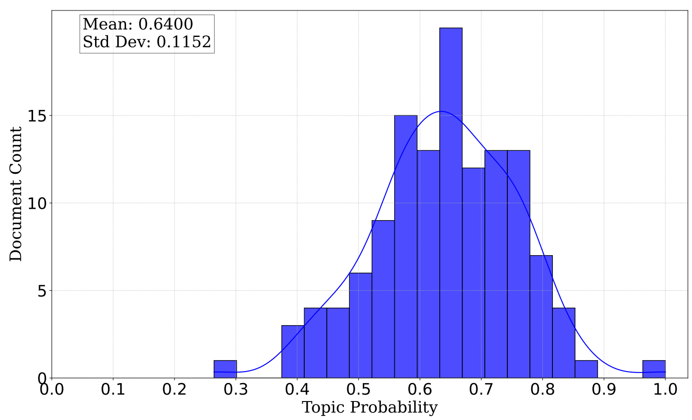
This study has several limitations to be addressed in future work. While zero-shot classification helps maintain continuity, it depends on the predefined themes being well-suited to

[5]as determined by the $\mathcal{AIC}$

Figure 8: Distribution of topic probabilities on 2024-06-28



(a) Backwards merge model



(b) Unmerged model

the domain (Stagnol *et al.*, 2023). If the themes are too coarse or mismatched, the mapping may lose granularity. The performance of BERTopic embeddings and zero-shot classification may also vary depending on the quality and diversity of the input data. Future work could explore alternative methods for topic alignment, such as semi-supervised learning or more adaptive zero-shot classifiers. The choice and construction of the underlying data set is an important decision on which all subsequent analysis is based. The broker data set is relatively small and homogeneous, which is reflected in the results of dynamic topic modeling. Future work could leverage larger and more heterogeneous data volumes, which may generate richer insights. Temporal networks are a powerful tool to track the propagation of topics or events over time. This analysis could be deepened by investigating temporal measures such as propagation length or by examining network topology measures such as depth and breadth of themes.

8 Conclusion

Our findings have practical implications for novelty detection in narratives with a temporal dimension, particularly in applications such as news tracking, social media analysis, and the early identification of economic and geopolitical developments. By balancing adaptability and continuity, the proposed approaches enhance the ability to monitor topics as they emerge, evolve, and, in the dynamic case, align with known themes.

The dynamic, temporal network-based approach allows for the real-time detection of emerging topics, making it well suited for rapidly evolving information landscapes such as breaking news and social media discourse. However, to mitigate the inherent continuity challenges of this method, we employ zero-shot classification to align new topics with predefined themes, ensuring that the evolving discourse remains interpretable and semantically stable over time. Meanwhile, the static topic modeling approach offers greater stability by preserving established themes, although it may lead to increasingly complex models as new topics accumulate over time.

Our dynamic framework provides a more robust methodology to analyze textual data over time. It enables researchers, analysts, and decision-makers not only to detect new topics but also to contextualize them within existing thematic structures, improving the interpretability and utility of automated topic modeling. These insights are particularly valuable in high-stakes domains such as financial market analysis, political risk assessment, and crisis monitoring, where the ability to track and anticipate narrative shifts can inform strategic decision-making.

Ultimately, our study contributes to the broader field of AI-driven text analysis by offering a nuanced approach to novelty detection, one that balances the need for adaptability with semantic continuity, ensuring that evolving narratives can be effectively monitored and understood.

REFERENCES

Abbas, O. A. (2008). *Comparisons Between Data Clustering Algorithms. International Arab Journal of Information Technology (IAJIT)*, *5*(3).

Aggarwal, C. C., & Yu, P. S. (2001). *Outlier Detection for High Dimensional Data. Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pp. 37–46.

Ahmed, F., & Courville, A. (2020). *Detecting Semantic Anomalies. Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), pp. 3154–3162.

Akaike, H. (1974). *A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control*, *19*(6), pp. 716–723.

Barnes, K., Juhász, P., Nagy, M., & Molontay, R. (2024). *Topicality Boosts Popularity: a Comparative Analysis of NYT Articles and Reddit Memes. Social Network Analysis and Mining*, *14*(1), pp. 119–137.

Blanqué, P., Ben Slimane, M., Cherief, A., Le Guenedal, T., Sekine, T., & Stagnol, L. (2022). *The Benefit of Narratives for Prediction of the S&P 500 Index. The Journal of Financial Data Science*, *4*(4), pp. 72–94. https://doi.org/10.3905/jfds.2022.1.107

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation. Journal of Machine Learning Research*, *3*(Jan), pp. 993–1022.

Bögel, T., & Gertz, M. (2015). *Time Will Tell: Temporal Linking of News Stories. Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 195–204.

Boutaleb, A., Picault, J., & Grosjean, G. (2024). *BERTrend: Neural Topic Modeling for Emerging Trends Detection. arXiv preprint arXiv:2411.05930*.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). *LOF: Identifying Density-based Local Outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104.

Cherief, A., Sekine, T., & Stagnol, L. (2025). *A Novel Nature-based Risk Index: Application to Acute Risks and their Financial Materiality on Corporate Bonds. Ecological Economics*, *228*, 108427. https://doi.org/https://doi.org/10.1016/j.ecolecon.2024.108427

Egger, R., & Yu, J. (2022). *A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in Sociology*, *7*, 886498.

Fortes, R., & Guenedal, T. L. (2021). *Tracking ECB's Communication: Perspectives and Implications for Financial Markets. SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3791244

GDELT Project. (2015). https://www.gdeltproject.org/

Ghosal, T., Saikh, T., Biswas, T., Ekbal, A., & Bhattacharyya, P. (2022). *Novelty Detection: a Perspective from Natural Language Processing. Computational Linguistics*, *48*(1), pp. 77–117.

Grootendorst, M. (2022). *BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. arXiv preprint arXiv:2203.05794*.

Hao, Y., Zhang, Y., Liu, K., He, S., Liu, Z., Wu, H., & Zhao, J. (2017). *An End-to-end Model for Question Answering over Knowledge Base with Cross-attention Combining Global Knowledge. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 221–231.

Holme, P., & Saramäki, J. (2012). *Temporal networks. Physics reports*, *519*(3), 97–125.

Hutto, C., & Gilbert, E. (2014). *Vader: a Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), pp. 216–225.

Kahn, R. N. (2019). *The Future of Investment Management*. https://www.cfainstitute.org/research/foundation/2019/the-future-of-investment-management

Kenter, T., & De Rijke, M. (2015). *Short Text Similarity with Word Embeddings. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1411–1420.

Keynes, J. M. (1937). *The General Theory of Employment. The Quarterly Journal of Economics*, *51*(2), pp. 209–223.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., et al. (2023). *ChatGPT: Jack of All Trades, Master of None. Information Fusion*, *99*, 101861.

Kumar, S., & Bhatia, K. K. (2020). *Semantic Similarity and Text Summarization Based Novelty Detection. SN Applied Sciences*, *2*(3), pp. 332–347.

Leetaru, K., & Schrodt, P. A. (2013). *Gdelt: Global Data on Events, Location, and Tone, 1979–2012. ISA Annual Convention*, *2*(4), pp. 1–49.

Leippold, M. (2023). *Sentiment Spin: Attacking Financial Sentiment with GPT-3. Available at SSRN 4337182*.

Li, L., Zhang, Q., Tian, J., & Wang, H. (2018). *Characterizing Information Propagation Patterns in Emergencies: a Case Study with Yiliang Earthquake. International Journal of Information Management*, *38*(1), pp. 34–41.

Ma, N., Mazumder, S., Politowicz, A., Liu, B., Robertson, E., & Grigsby, S. (2022). *Semantic Novelty Detection and Characterization in Factual Text Involving Named Entities. arXiv preprint arXiv:2210.17440*.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Magazine*, *27*(4), 12–12.

Menchero, J., Orr, D., & Wang, J. (2011, August). *The Barra US Equity Model (USE4): Methodology Notes.* MSCI. https://www.msci.com/documents/10199/242721/Barra_US_Equity_Model_USE4.pdf

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). *MTEB: Massive Text Embedding Benchmark. arXiv preprint arXiv:2210.07316.*

Mũnoz-Marí, J., Bovolo, F., Gómez-Chova, L., Bruzzone, L., & Camp-Valls, G. (2010). *Semisupervised One-class Support Vector Machines for Classification of Remote Sensing Data. IEEE Transactions on Geoscience and Remote Sensing*, *48*(8), pp. 3188–3197.

Newell, A. (1972). *Human Problem Solving. Upper Saddle River/Prentive Hall.*

Ogbuabor, G., & Ugwoke, F. (2018). *Clustering Algorithm for a Healthcare Dataset using Silhouette Score Value. Int. J. Comput. Sci. Inf. Technol*, *10*(2), pp. 27–37.

Patel, E., & Kushwaha, D. S. (2020). *Clustering Cloud Workloads: K-means vs Gaussian Mixture Model. Procedia Computer Science*, *171*, pp. 158–167.

Pushp, P. K., & Srivastava, M. M. (2017). *Train Once, Test Anywhere: Zero-shot Learning for Text Classification. arXiv preprint arXiv:1712.05972.*

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). *Efficient Algorithms for Mining Outliers from Large Data Sets. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 427–438.

Reynolds, D. A., et al. (2009). *Gaussian Mixture Models. Encyclopedia of Biometrics*, *741*(pp. 659-663).

Richie, R., & Bhatia, S. (2021). *Similarity Judgment Within and Across Categories: a Comprehensive Model Comparison. Cognitive Science*, *45*(8).

Shahapure, K. R., & Nicholas, C. (2020). *Cluster quality analysis using silhouette score. 2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, 747–748.

Shiller, R. J. (2017). *Narrative Economics. American Economic Review*, *107*(4), pp. 967–1004.

Si, M., Cui, L., Guo, W., Li, Q., Liu, L., Lu, X., & Lu, X. (2020). *A Comparative Analysis for Spatio-temporal Spreading Patterns of Emergency News. Scientific Reports*, *10*(1), 19472.

STAGNOL, L., CHERIEF, A., FARAH, Z., LE GUENEDAL, T., SAKOUT, S., & SEKINE, T. (2023). *Answering Clean Tech Questions with Large Language Models. Available at SSRN 4663447*.

TANG, J., QU, M., & MEI, Q. (2015). *Pte: Predictive Text Embedding Through Large-scale Heterogeneous Text Networks. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1165–1174.

TAX, D. M., & DUIN, R. P. (2004). *Support Vector Data Description. Machine Learning*, *54*, pp. 45–66.

THONGTAN, T., & PHIENTHRAKUL, T. (2019). *Sentiment Classification Using Document Embeddings Trained with Cosine Similarity. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 407–414.

THORSRUD, L. A. (2016). *Nowcasting Using News Topics. Big Data Versus Big Bank*.

TILLY, S., & LIVAN, G. (2021). *Macroeconomic Forecasting with Statistically Validated Knowledge Graphs. Expert Systems with Applications*, *186*, 115765.

WESTERSKI, A., DALAMAGAS, T., & IGLESIAS, C. A. (2013). *Classifying and Comparing Community Innovation in Idea Management Systems. Decision Support Systems*, *54*(3), pp. 1316–1326.

WORLD BANK. (2016). *Theme Taxonomy and Definitions: Revised July 1, 2016*. https://thedocs.worldbank.org/en/doc/275841490966525495-0290022017/New-Theme-Taxonomy-and-definitions-revised-July-012016

XU, G., MENG, Y., CHEN, Z., QIU, X., WANG, C., & YAO, H. (2019). *Research on Topic Detection and Tracking for Online News Texts. IEEE access*, *7*, pp. 58407–58418.

A Appendix

A.1 Prompt for generating themes

"Given these topics, generate distinct and well-separated overarching themes that encompass all of these topics."

A.2 Dynamic approach: details on novelty topics

As indicated by Egger and Yu (2022), the fact that we do not preprocess the text for the embedding approach of BERTopic, the representation displays a majority of stop words. We prompt the representative documents with "summarize in a short topic sentence" with a LLM.

Step1_T1 (14/06/2024), Representation: 'the', 'banks', 'and', 'banking', 'to', 'rn', 'of', 'tax', 'in', 'we'
Short description: The potential implementation of a financial transaction tax by the RN to fund lower VAT on essential goods raises concerns about increased budget deficits and risks to the banking sector, particularly if a RN-led government is formed, which may lead to higher costs of equity and reduced profitability for French banks.
Theme: Sector Performance and Comparisons
Sentiment: 0.836

Step2_T1 (19/06/2024), Representation: 'figure', 'france', 'and', 'the', 'of', 'to', 'deficit', 'in', 'ratio', 'exposure'
Short description: In 2024, France is projected to have a deficit ratio exceeding 5% and a debt ratio of 112.4% of GDP, both significantly higher than the euro area averages, with only Slovakia, Italy, and Greece having worse ratios, and Belgium, Spain, the Netherlands, and Luxembourg being notably exposed to a slowdown in the French economy.
Theme: Financial Markets and Instruments
Sentiment: -0.0772

Step2_T4 (19/06/2024), Representation: 'indicator', 'fiscal', 'the', 'to', 'indicators', 'component', 'crisis', 'vulnerability', 'needs', 'with'
Short description: The S1 and S2 indicators measure the fiscal effort needed to manage debt and fiscal vulnerability in the long term, considering the costs of population aging, with S1 focusing on reducing the debt ratio to 60% by 2070 and S2 on stabilizing debt without additional fiscal consolidations.
Sentiment: -0.4767

Step3_T3 (20 - 23/06/2024), Representation: 'banks', 'bnp', 'and', 'we', 'down', 'more', 'in', 'diversified', 'uk', 'political'
Short description: BNP is the most diversified and stable among the three listed French banks, with a reliable earnings profile and lower direct exposure to France, but its stock is currently down due to political uncertainty.
Sentiment: -0.7478

Step3_T13 (20 - 23/06/2024), Representation: 'investors', 'asset', 'survey', 'likely', 'ex', 'skew', 'investor', 'reduce', 'france', 'more'
Short description: Investors already underweight in Europe, particularly France, are more inclined to reduce their exposure further, while those overweight are more likely to make slight reductions, with France-based investors showing a higher tendency to reduce Europe ex-France exposures.
Sentiment: 0.3612

Step3_T14 (20 - 23/06/2024), Representation: 'bund', 'last', 'rally', 'yield', 'widening', 'move', 'weeks', 'figure', 'egb', 'episodes'
Short description: The document analyzes the impact of significant spread widening episodes on oat and GDP-weighted EGB yields, highlighting the recent movements in oat yields, bund rallies, and the decomposition of GDP-weighted EGB spreads.
Theme: Sector Performance and Comparisons
Sentiment: -0.4404

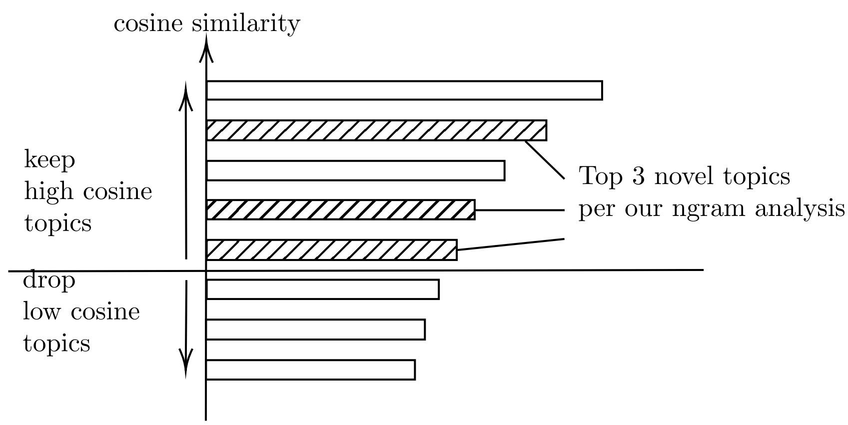
Step3_T23 (20 - 23/06/2024), Representation: 'before', 'seats', 'president', 'elections', 'won', '2022', 'presidential', 'the', '2017', 'conducted'
Short description: The French Constitution restricts the president from calling new parliamentary elections within one year of the last election, with historical precedents in 1988 and 2022 where the presidential majority failed to secure a majority in the legislative elections.
Theme: Financial Markets and Instruments
Sentiment: 0.0516

Step5_T5 (26 - 27/06/2024), Representation: 'the', 'week', 'weekly', 'and', 'yields', '10yr', 'spread', 'french', 'since', 'in'
Short description: The announcement of a snap election in France led to a significant increase in the risk premium on French assets, causing the CAC 40 to experience its worst weekly performance in over two years, while the spread between French and German 10-year bonds saw its largest weekly increase since the eurozone debt crisis.
Theme: Financial Markets and Instruments
Sentiment: -0.2975

Step5_T7 (26 - 27/06/2024), Representation: 'average', 'oecd', 'of', 'labour', 'employment', 'gdp', 'the', 'has', 'rate', 'productivity'
Short description: France faces economic challenges including high government expenditure, low productivity growth, high tax burdens, and structural labor issues, which impact its competitiveness and investment.
Theme: Financial Markets and Instruments
Sentiment: 0.4767

A.3 top 3 novel n-grams as per our ngram analysis

Figure 9: Top 3 novel n-grams as per our ngram analysis



Investment Institute

WORKING PAPER 173 | **MARCH 2025**