From Noise to Signal: When Outliers Seed New Topics

Evangelia Zve[1,2], Gauvain Bourgne[1], Benjamin Icard[1], and Jean-Gabriel Ganascia[1]

[1] Sorbonne Université, CNRS, LIP6, France
[2] Infopro Digital, France

{evangelia.zve, gauvain.bourgne, benjamin.icard, jean-gabriel.ganascia}@lip6.fr
evangelia.zve@infopro-digital.com

Abstract

Outliers in dynamic topic modeling are typically treated as noise, yet we show that some can serve as early signals of emerging topics. We introduce a *temporal taxonomy* of news-document trajectories that defines how documents relate to topic formation over time. It distinguishes *anticipatory outliers*, which precede the topics they later join, from documents that either reinforce existing topics or remain isolated. By capturing these trajectories, the taxonomy links weak-signal detection with temporal topic modeling and clarifies how individual articles anticipate, initiate, or drift within evolving clusters. We implement it in a cumulative clustering setting using document embeddings from eleven state-of-the-art language models and evaluate it retrospectively on HydroNewsFr, a French news corpus on the hydrogen economy. Inter-model agreement reveals a small, high-consensus subset of anticipatory outliers, increasing confidence in these labels. Qualitative case studies further illustrate these trajectories through concrete topic developments.

Keywords: weak signals, emerging topics, dynamic topic modeling, outliers, density-based clustering

1. Introduction

Anticipating emerging topics in fast-evolving news streams is essential for tracking public debate, identifying opportunities, and monitoring risks (Ansoff, 1980). However, most topic-modeling methods identify topics only after coherent clusters have formed, which limits their ability to capture early signals (Churchill and Singh, 2022).

This limitation stems in part from how these methods handle atypical documents (Zve et al., 2025). Such documents often appear before a topic stabilizes, yet mainstream approaches typically assign them to diffuse clusters or discard them as noise (Hiltunen, 2008). We argue that topic modeling should account for the temporal relationship between documents and evolving clusters, while also distinguishing *outliers* that may signal emerging topics. In practice, this requires clustering-based topic-modeling methods (Grootendorst, 2022) that can infer the number of topics, align topic clusters over time, and explicitly account for anomalies.

In this paper, we propose a document-level perspective on topic emergence. We introduce a *temporal taxonomy* of news-document trajectories that characterizes how documents relate to topic formation over time through three key events: a document's first appearance, the emergence of a topic cluster, and, when relevant, the document's first integration into that cluster. Within this framework, we define *anticipatory outliers* as documents that precede the topic clusters they later join.

To validate the taxonomy, we apply it retrospectively to HydroNewsFr, a curated French news corpus on the hydrogen economy. Articles are embedded with eleven text-embedding models and clustered cumulatively in daily windows, after which the resulting topic clusters are aligned over time. Documents are then labeled according to the taxonomy based on their observed trajectories. We evaluate the robustness of these labels through inter-model agreement and complement this analysis with qualitative case studies.

Section 2 reviews prior work on topic modeling and weak-signal detection. Section 3 introduces the taxonomy. Section 4 describes the corpus, and Section 5 details the modeling framework. Sections 6 and 7 present the quantitative results and representative case studies. Section 8 concludes and outlines directions for future work.

2. Related Work

Dynamic topic modeling has been applied in diverse domains, from policy discourse (Wang and McCallum, 2006) and corporate strategy (Kaplan and Vakili, 2015) to scientific research (Hall et al., 2008). It builds on probabilistic methods such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which infers topics from word co-occurrence patterns. Temporal extensions, notably Dynamic Topic Models (DTM) (Blei and Lafferty, 2006) and the Dynamic Embedded Topic Model (DETM) (Dieng et al., 2019), a neural variant, incorporate time-awareness into the inference process. Unlike static topic models, these methods capture how topics evolve over time. A key limitation is that they require the number of topics to be specified in advance (Chuang et al., 2013). They also assign every document to a topic, which obscures outliers.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7523–7533
11–16 May 2026.

More recent embedding-based methods such as `BERTopic` (Grootendorst, 2022) use document representations from pretrained language models and identify topics by clustering in the resulting semantic space. Their temporal extensions (Boutaleb et al., 2024) approximate topic evolution through post-hoc alignment of static clusters across time windows, rather than by modeling temporal dependencies directly. Our framework follows this strategy by aligning cumulative windows (Section 5.3). The results of these methods are sensitive to the choice of clustering algorithm. Partition-based algorithms such as `KMeans` (Hartigan and Wong, 1979) require a preset number of clusters and force all documents into topics, whereas density-based methods such as `HDBSCAN` (McInnes et al., 2017) and `OPTICS` (Ankerst et al., 1999) can infer the number of clusters and label low-density documents as outliers. Typically discarded as noise, these outliers may instead indicate emerging topics, an aspect that remains largely overlooked in dynamic topic modeling (Zve et al., 2025).

Research on computational early detection in text streams spans multiple traditions. Event and first-story detection (Allan, 2002) focus on short-term novelty by identifying events as they appear. While our approach also targets early signals at the document level, it is informed by weak-signal analysis (Hiltunen, 2008), which seeks subtle, fragmented cues that may foreshadow broader thematic shifts. Subsequent work (Christophe et al., 2021) extends this approach to large corpora, but does not connect these signals to topic formation.

To address this gap, we introduce a taxonomy of document trajectories that characterizes anticipatory outliers as weak signals of emerging topics.

3. Taxonomy of News Trajectories

We introduce a *temporal taxonomy* to characterize how individual documents evolve with respect to topic formation. Inspired by Allen's interval algebra for temporal reasoning (Allen, 1983), we represent each document through three key events: its first appearance (T_A), the creation of its eventual topic (T_T), and its first integration into that topic (T_I), summarized in Table 1. Comparing their relative order yields distinct temporal behaviors, including *anticipatory outliers* ($T_A < T_T \leq T_I$), in which a document appears before the topic it later joins.

Symbol	Definition
T_A	Document appearance time in the stream
T_T	Topic creation time
T_I	Document integration time
T_{final}	Final time window in the corpus
θ_{delay}	Delay cutoff separating persistent and recent outliers

Table 1: Temporal notation used in the taxonomy

The taxonomy summarized in Table 2 partitions documents into *seven* mutually exclusive cases determined by the order and presence of these temporal events, as shown in the decision tree in Figure 1. The symbolic notation ($\mathcal{T}$, $\mathcal{O}$, $\mathcal{A}$, $\mathcal{D}$, $\mathcal{NO}$), derived from the initials of *Topic*, *Outlier*, *Anticipatory*, *Drift*, and *Non-Outlier*, provides an intuitive shorthand for each document's trajectory type.

Symbol	Description	Condition
$\mathcal{T}$ (integrated into Topic)		
$\mathcal{TO}$ (Outlier integrated into Topic)		
$\mathcal{TOA}$ (Anticipatory Outlier integrated into Topic)		
$\mathcal{TOA}_{\text{first}}$	at topic creation	$T_A < T_I = T_T$
$\mathcal{TOA}_{\text{late}}$	after topic creation	$T_A < T_T < T_I$
$\mathcal{TOD}_{\text{late}}$ *(Drift Outlier integrated into Topic)*		$T_T < T_A < T_I$
$\mathcal{TNO}$ (Non-Outlier integrated into Topic)		
$\mathcal{T}_{\text{first}}$	at topic creation	$T_A = T_I = T_T$
$\mathcal{T}_{\text{late}}$	after topic creation	$T_T < T_A = T_I$
$\mathcal{O}$ (Outlier, not integrated into a topic)		
$\mathcal{O}_{\text{recent}}$	appeared recently	$T_{\text{final}} - T_A < \theta_{\text{delay}}$
$\mathcal{O}_{\text{old}}$	long-standing outlier	$T_{\text{final}} - T_A \geq \theta_{\text{delay}}$

Table 2: Formal conditions for the taxonomy cases

Two cases correspond to documents that never integrate into a topic during the study period. They are distinguished by their recency with respect to the corpus end time (T_{final}): recently isolated documents are labeled $\mathcal{O}_{\text{recent}}$ and their eventual trajectory may extend beyond T_{final}, while long-standing outliers are labeled $\mathcal{O}_{\text{old}}$. The integration delay $\Delta T = T_I - T_A$ is the time between a document's first appearance and its first integration into a topic. The cutoff θ_{delay}, which separates these two cases, is set to the empirical 90th percentile of delays among outlier documents that later joined a topic.

A further distinction separates documents that integrate directly into a topic from those that spend time as outliers before integration. $\mathcal{T}_{\text{first}}$ and $\mathcal{T}_{\text{late}}$, grouped as $\mathcal{TNO}$, join a topic without passing through an outlier stage. $\mathcal{T}_{\text{first}}$ joins exactly when the topic forms ($T_A = T_I = T_T$), whereas $\mathcal{T}_{\text{late}}$ joins after topic creation ($T_T < T_A = T_I$).

The remaining classes concern documents with an explicit outlier phase. Among them, $\mathcal{TOA}$ corresponds to the *anticipatory-outlier* pattern ($T_A < T_T \leq T_I$). These are divided into $\mathcal{TOA}_{\text{first}}$, where integration coincides with topic creation ($T_I = T_T$), and $\mathcal{TOA}_{\text{late}}$, where it occurs only after the topic is established ($T_T < T_I$). In contrast, $\mathcal{TOD}_{\text{late}}$ represents *topic drift*: an outlier that appears after its future topic is created ($T_T < T_A$), reinforcing an existing topic rather than introducing a new one.

Each document in the corpus is labeled according to this taxonomy. Our analysis focuses on anticipatory cases ($\mathcal{TOA}$) to examine their role as early signals. More broadly, the taxonomy provides a structured view of how news trajectories unfold.

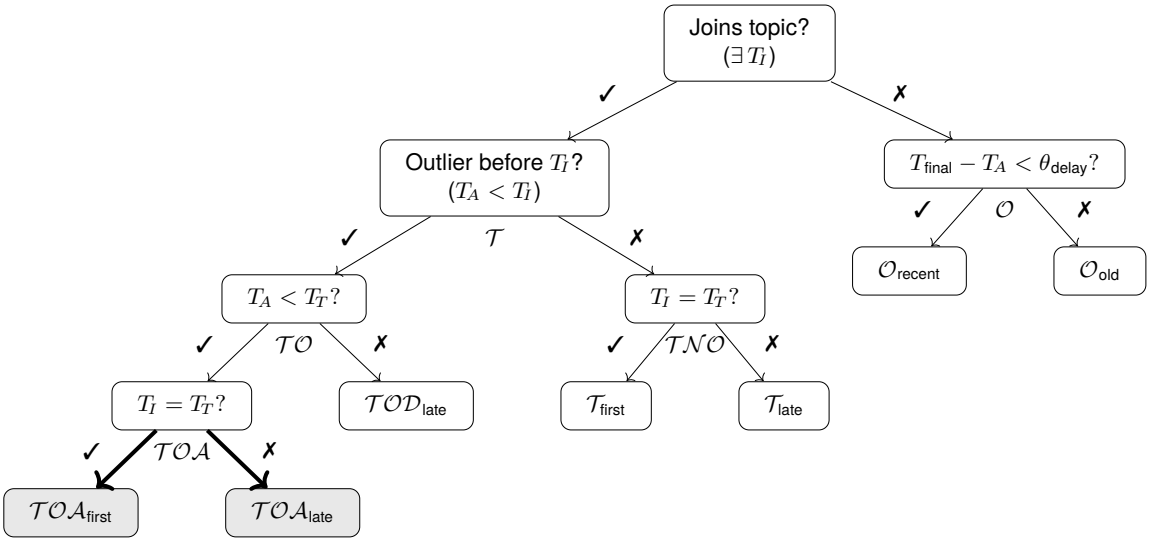


Figure 1: Decision tree mapping conditions on $(T_T, T_A, T_I, \theta_{\text{delay}})$ to the seven cases. *Left* branch splits outlier-integrations $(\mathcal{TOA}_{\text{first}}, \mathcal{TOA}_{\text{late}}, \mathcal{TOD}_{\text{late}})$ from non-outlier integrations $(\mathcal{T}_{\text{first}}, \mathcal{T}_{\text{late}})$; *right* branch yields outliers $(\mathcal{O}_{\text{recent}}, \mathcal{O}_{\text{old}})$ persisting until T_{final}. ✓ / ✗ indicate branch outcomes.

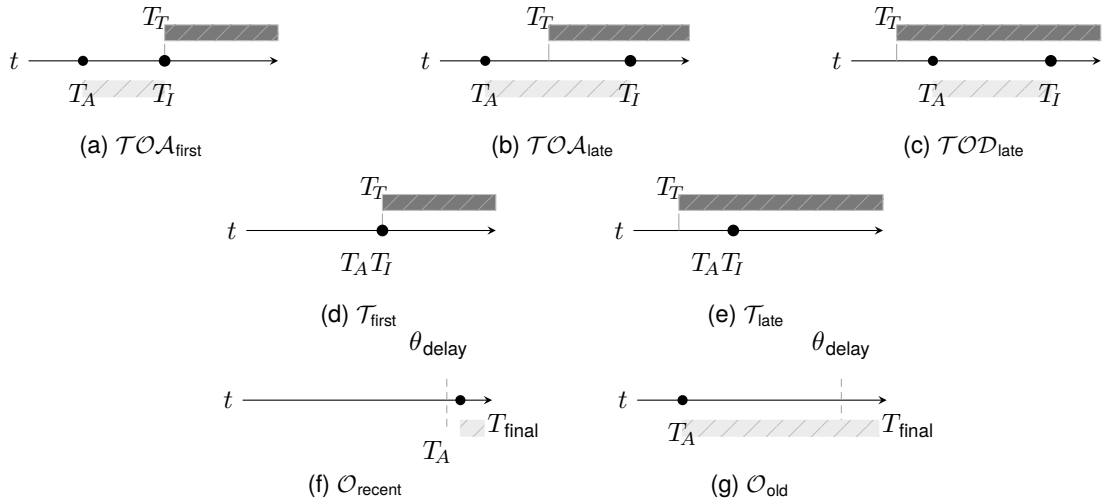


Figure 2: Overview of the seven taxonomy cases. Panels (a–g) show temporal relations between T_A, T_T, T_I, θ_{delay}, T_{final}. Colored belts indicate topic activity and outlier phases.

4. Dataset

To study the document *trajectories* discussed in the previous section, we required a corpus that is both temporally consistent and topically continuous. Existing news collections often rely on a single source or include irregular sampling, creating gaps in coverage. To address this, we built HydroNewsFr, a French-language corpus on hydrogen energy that provides uninterrupted daily reporting across 81 consecutive days (20 March–8 June 2025).

The dataset captures a period of industrial and policy activity related to hydrogen, including new funding schemes, infrastructure projects, financial developments, and vehicle launches. To ensure source diversity and avoid sparsity, it combines material from official communications, mainstream media, local, and specialized press.

Data were collected daily from two complementary sources to ensure broad and diverse coverage. We retrieved social media posts from `X (formerly Twitter)`[1] containing the French keyword *hydrogène* (*hydrogen*) and linking to news articles, and articles gathered using the `Google News`[2] Python library with the same keyword. From `X`, we collected 1,533 posts corresponding to 726 unique articles, with headline and description fields available directly in the API response. `Google News` yielded an additional 891 unique articles, from which the same fields were extracted. For

[1] https://developer.twitter.com/en/docs/twitter-api

[2] https://github.com/ranahaani/GNews

each item, we recorded metadata such as publication date, URL, and, when available, the country. We applied standard preprocessing, removing boilerplate text (e.g., irrelevant content, repeated templates) and normalizing metadata fields (e.g., date formats). As a result, each document consists of a headline and short description, averaging 280 characters. Finally, we performed cross-source deduplication, prioritizing records with more complete content when overlaps occurred. The final dataset comprises 1,616 articles.

Figure 3 shows the dataset's temporal distribution. Daily counts reveal coverage peaks and clear regularity with minimal timeline gaps. Cumulative counts confirm uninterrupted coverage.

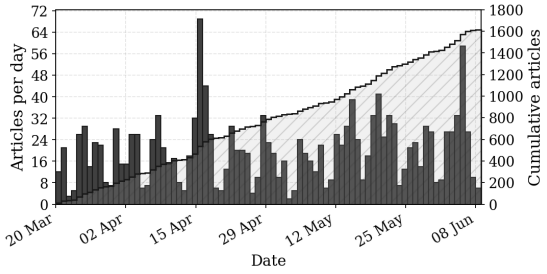


Figure 3: Temporal distribution of HydroNewsFr over the collection period (20 March–8 June 2025). Bars show daily publication counts; the stepped line and hatched area show cumulative totals.

5. Modeling Framework

To assign each article to a taxonomy case according to whether and when it becomes integrated into a topic, we propose a cumulative clustering framework. This approach first maps articles to semantic embeddings and then clusters them over successive time windows to capture topic continuity and emergence.

5.1. Text Representation

Each news article is represented by embeddings produced by eleven pre-trained language models. For each article, the title and short description are concatenated and encoded as a single text sequence. In the resulting semantic vector space, thematically similar articles are expected to lie close to one another.

To mitigate the *curse of dimensionality* and improve clustering efficiency, we employ Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to obtain lower-dimensional representations of the embeddings.[3] We evaluate output dimensionalities of 2, 3, 5, 10, 20, 30, and 40, keeping the default hyperparameters and fixing the random seed for reproducibility.

We selected eleven embedding models representing diverse architectures, based on their performance on the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) as of June 2025. For our French-language corpus, we prioritized multilingual and French-specialized models. The final set includes both open-source models available on Hugging Face and proprietary models.

Model	Dimensions	Language
sentence-camembert-base	768	French
Solon-embeddings-large-0.1	1024	French
paraphrase-..-MiniLM-L12-v2	384	Multilingual
paraphrase-..-mpnet-base-v2	768	Multilingual
LaBSE	768	Multilingual
multilingual-e5-large	1024	Multilingual
snowflake-arctic-embed-l-v2.0	1024	Multilingual
bge-m3	1024	Multilingual
text-embedding-3-small	1536	Multilingual
gemini-embedding-001	3072	Multilingual
mistral-embed	1024	Multilingual

Table 3: The eleven embedding models.

5.2. Cumulative Clustering

Our clustering procedure adapts the BERTopic (Grootendorst, 2022) approach with a cumulative design that incrementally incorporates prior documents to model topic evolution. We cluster reduced embeddings from the beginning of the corpus up to each day, producing a sequence of expanding windows that captures its evolving topical structure. The study period (20 Mar–8 Jun 2025) yields 81 daily windows. This daily granularity aligns with the pace of the news cycle.

We use two density-based clustering algorithms, HDBSCAN[4] and OPTICS[5]. Both algorithms were run with their default parameters, and a fixed random seed for reproducibility, each assigning documents a cluster ID or outlier label (-1). In HDBSCAN, outlier detection relies on the GLOSH algorithm, which estimates relative local density and designates documents in sparse regions as outliers. In OPTICS, outliers are inferred from the reachability plot when a document's reachability distance exceeds the cluster-boundary threshold.

Clustering quality is assessed using the silhouette score (Shahapure and Nicholas, 2020), which measures intra-cluster cohesion versus inter-cluster separation. Scores above 0.7 indicate strong structure, 0.5–0.7 moderate structure, and below 0.25 weak structure. We report mean and median silhouette scores across all time windows for all combinations of embedding models, clustering algorithms, and UMAP dimensionalities.

[3] https://umap-learn.readthedocs.io

[4] https://hdbscan.readthedocs.io

[5] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS

5.3. Topic Alignment

Since the cumulative clustering pipeline produces clusters interpreted as topics independently at each time window, a linking method is required to trace their trajectories. Unlike more flexible approaches such as ANTM (Rahimi et al., 2024), which allow many-to-many mappings, we enforce one-to-one correspondences to preserve interpretability. Each previous topic either continues or disappears, and new topics can emerge (Vaca et al., 2014). Splits and merges are treated as discontinuities.

Formally, each topic cluster is represented by the centroid of its UMAP-reduced embeddings. Let

$$T^{(t)} = \{\tau_1^{(t)}, \ldots, \tau_{n_t}^{(t)}\}$$

be the set of topics (i.e., clusters) at time t, with centroid

$$\mathbf{c}_i^{(t)} = \frac{1}{|\mathcal{D}_i^{(t)}|} \sum_{\mathbf{d} \in \mathcal{D}_i^{(t)}} \mathbf{d},$$

where $\mathcal{D}_i^{(t)}$ is the set of document embeddings assigned to topic $\tau_i^{(t)}$, excluding outliers.

To align topics across consecutive windows, we compute the cosine distance matrix:

$$D_{ij} = 1 - \frac{\mathbf{c}_i^{(t-1)} \cdot \mathbf{c}_j^{(t)}}{\|\mathbf{c}_i^{(t-1)}\| \, \|\mathbf{c}_j^{(t)}\|},$$

and apply the Hungarian algorithm (Kuhn, 1955) to obtain the optimal one-to-one matching. A distance threshold θ_{align} determines whether a topic continues over time: if $D_{ij} \leq \theta_{\text{align}}$, topic $\tau_j^{(t)}$ is matched to $\tau_i^{(t-1)}$; otherwise, it is treated as a new topic. We vary θ_{align} from 0.20 to 0.70 in increments of 0.10 to assess sensitivity. Lower values yield stricter matching and more splits, whereas higher values allow looser matching and fewer new topics.

5.4. Case Assignment

For each configuration, defined by the UMAP dimensionality, clustering algorithm, and θ_{align}, we track the trajectory of every document across daily windows. Each of the eleven embedding models assigns one taxonomy case to every document, producing a document–model label matrix.

6. Experiments

6.1. Topic-based Clustering

Following the methodology described in Section 5, we performed cumulative clustering over 81 daily windows in the HydroNewsFr corpus. We tested eleven embedding models, seven UMAP dimensions (2–40D), two clustering algorithms (HDBSCAN and OPTICS), and six topic-alignment thresholds (θ_{align}) ranging from 0.20 to 0.70.

Clustering quality was evaluated for each configuration using the mean silhouette score across all cumulative time windows. The results in Table 4 indicate moderate to strong cluster separation, with scores ranging from 0.419 to 0.653 for HDBSCAN and from 0.551 to 0.643 for OPTICS. Across embedding models, both algorithms reach their highest mean and median values at 2D UMAP reduction. mistral-embed yields the highest silhouette scores at all dimensionalities for HDBSCAN, whereas for OPTICS the top score varies across dimensionalities. Overall, the best configuration is UMAP 5D with HDBSCAN on mistral-embed, achieving a silhouette score of 0.653.

Figure 4 illustrates typical trajectory patterns observed in cumulative clustering. Some articles first appear as outliers and later integrate into a topic, anticipating its emergence ($\mathcal{TOA}$). Others appear only after a topic has formed, reflecting drift within established topics ($\mathcal{TOD}_{\text{late}}$). A third group integrates directly without an outlier phase ($\mathcal{T}$).

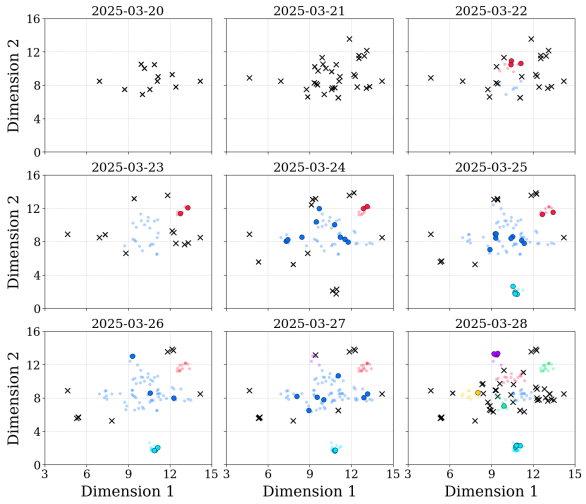


Figure 4: Cumulative clustering over the first 9 time windows with mistral-embed and 2D UMAP. Colors indicate topics; newly assigned documents are larger and more opaque; black $\times$ denote outliers.

6.2. Topic Integration Delays

The trajectory analysis reveals when documents transition from outliers to integrated topics. To better understand this temporal dynamic, we quantify how long such transitions take using the notion of *integration delay*, defined as $\Delta T = T_I - T_A$ (in days) for an article that first appears as an outlier and later joins a topic. The empirical survival function is given by $S(t) = P(\Delta T > t)$, pooled across models and configurations. Only outlier articles that eventually integrated are included in this analysis ($\mathcal{TOA}_{\text{first}}$, $\mathcal{TOA}_{\text{late}}$, $\mathcal{TOD}_{\text{late}}$).

Model	HDBSCAN							OPTICS						
	2D	3D	5D	10D	20D	30D	40D	2D	3D	5D	10D	20D	30D	40D
bge-m3	0.621	0.611	0.589	0.618	0.583	0.602	0.615	0.634	0.615	0.610	0.592	**0.637**	0.602	**0.638**
sentence-camembert-base	0.582	0.579	0.568	0.562	0.574	0.559	0.561	0.612	0.565	0.585	0.551	0.600	0.571	0.578
gemini	0.601	0.581	0.575	0.576	0.574	0.578	0.565	0.633	0.605	0.619	0.598	0.626	0.621	0.602
multilingual-e5-large	0.624	0.604	0.616	0.603	0.617	0.601	0.608	0.635	0.624	0.612	0.597	0.620	0.620	0.611
mistral-embed	**0.638**	**0.628**	**0.653**	**0.641**	**0.636**	**0.636**	**0.630**	0.638	**0.643**	**0.636**	**0.606**	0.620	**0.628**	0.617
text-embedding-3-small	0.606	0.611	0.608	0.602	0.597	0.600	0.596	**0.639**	0.612	0.612	0.603	0.618	0.609	0.614
Solon-embeddings-large-0.1	0.606	0.612	0.601	0.583	0.589	0.594	0.590	0.628	0.625	0.601	0.567	0.610	0.600	0.592
LaBSE	0.596	0.603	0.577	0.559	0.553	0.552	0.568	0.599	0.607	0.611	0.573	0.592	0.593	0.579
paraphrase..-MiniLM-L12-v2	0.519	0.496	0.565	0.551	0.547	0.483	0.419	0.603	0.571	0.566	0.545	0.589	0.565	0.564
paraphrase..-mpnet-base-v2	0.514	0.545	0.567	0.544	0.564	0.565	0.553	0.607	0.580	0.574	0.577	0.570	0.577	0.573
snowflake-arctic..-l-v2.0	0.612	0.621	0.594	0.619	0.613	0.602	0.575	0.631	0.610	0.617	0.600	0.595	0.579	0.594
Mean	**0.593**	0.590	0.592	0.587	0.586	0.579	0.571	**0.624**	0.605	0.604	0.583	0.607	0.597	0.597
Median	**0.606**	0.604	0.589	0.583	0.583	0.594	0.575	**0.631**	0.611	0.610	0.585	0.614	0.601	0.597

Table 4: Average silhouette scores per model and UMAP dimension. Best score per dimension is bolded.

Figure 5 shows that the distribution decays rapidly but exhibits a long tail. The median is $p_{50} = 5$ days and the upper quartile $p_{75} = 14$ days; the tail extends to about a month ($p_{90} = 26$ days; $p_{95} = 34$ days).

Based on this empirical profile, we set the delay cutoff to $\theta_{\text{delay}} = 26$ days, at which at least 90% of integrating documents have joined a topic. Consistent with our taxonomy, we use θ_{delay} to partition non-integrated items ($\mathcal{O}$) into $\mathcal{O}_{\text{recent}}$ (recent outliers) and $\mathcal{O}_{\text{old}}$ (persistent outliers).

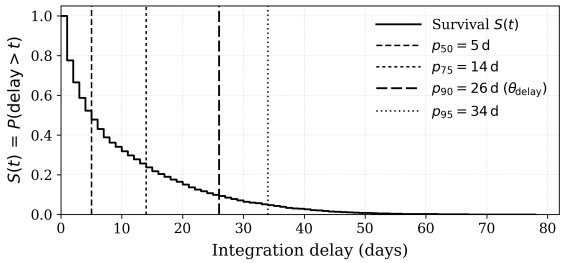


Figure 5: Empirical survival curve $S(t) = P(\Delta T > t)$ for integration delays, pooled across models and configurations. Dashed lines mark quantiles; the 90th percentile ($p_{90} = 26$ days) defines θ_{delay}.

6.3. Inter-Model Agreement

This section examines the sensitivity of $\mathcal{TOA}$ labels and delay assignments to the choice of embedding model. The $\mathcal{TOA}$ task is binary ($\mathcal{TOA}$ vs. all other cases), whereas delay is multiclass, representing the temporal offset until integration. Although $\mathcal{TOA}$ labels vary across embedding models, agreement declines gradually as the number of agreeing models increases, leaving a small high-agreement core of articles. This pattern suggests that inter-model agreement can serve as a proxy for label robustness and help rank anticipatory signals by confidence.

As a first step, we verified that varying the UMAP dimensionality (2–40D) has limited impact on $\mathcal{TOA}$ assignments. As shown in Table 5, Fleiss' κ for both case and delay remains in the *fair-to-moderate* range (0.34–0.51) across models and is stable across alignment thresholds (variation < 0.01). These results suggest that inter-model agreement is largely invariant to UMAP dimensionality, with most variation arising from differences among embedding models. Detailed per-model results are reported in Appendix Table 9.

Algorithm	Task	Mean Fleiss' κ	Range θ_{align}	Model Range
HDBSCAN	Case	0.46	0.46–0.47	0.36–0.51
HDBSCAN	Delay	0.36	0.36–0.37	0.20–0.43
OPTICS	Case	0.37	0.37–0.37	0.34–0.41
OPTICS	Delay	0.30	0.30–0.31	0.23–0.33

Table 5: Mean Fleiss' κ on anticipatory outliers ($\mathcal{TOA}$) across all models and alignment thresholds $\theta_{\text{align}} \in [0.2, 0.7]$. "Range θ_{align}" shows variability across θ_{align} for fixed models, while "Model Range" shows variability across models for fixed θ_{align}.

With dimensionality effects shown to be minimal, we focus on variation across embedding models. While Fleiss' κ provides an overall measure of agreement corrected for chance and class imbalance, it does not indicate how consistently individual documents are labeled across embedding models. To capture this document-level consistency, we introduce the *majority agreement* (MA), treating the eleven models as independent raters that each assign a label to every article. For each article, we calculate the proportion of models agreeing on the most frequent label, and then average this proportion across all articles:

$$\text{MA} = \frac{1}{D}\sum_{i=1}^{D}\text{MA}_i, \quad \text{with } \text{MA}_i = \max_k \frac{n_{ik}}{M}$$

where $M = 11$ is the number of models, D the number of documents, and n_{ik} the number of models assigning label k, with k denoting a taxonomy case for the *case* task or a delay value for the *delay* task. For the anticipatory-outlier (*case*) task, $k \in \{0, 1\}$ represents binary decisions, while for *delay*, k takes multiple discrete values, with $\Delta T = 0$ for immediate integration and $\Delta T = \infty$ for never integrating.

Algorithm	Task	Mean MA	Range	Best UMAP/θ_{align}
HDBSCAN	Case	0.94	0.94–**0.95**	20D / 0.30
	Delay	0.74	0.72–**0.76**	40D≃20D / 0.30
OPTICS	Case	0.92	0.91–**0.93**	2D / 0.20
	Delay	0.66	0.65–**0.68**	10D / 0.60

Table 6: Mean majority agreement (MA) for $\mathcal{TOA}$ across $\theta_{\text{align}} \in [0.2, 0.7]$ and UMAP dimensions (2–40D). The range shows min–max values, and "Best UMAP/θ_{align}" indicates where the maximum occurs.

Algorithm	Task	Mean Fleiss' κ	Range	Best UMAP/θ_{align}
HDBSCAN	Case	0.30	0.24–**0.36**	30D / 0.30
	Delay	0.18	0.13–**0.20**	20D / 0.40
OPTICS	Case	0.20	0.13–**0.26**	10D / 0.40
	Delay	0.16	0.09–**0.24**	10D / 0.30

Table 7: Mean Fleiss' κ for $\mathcal{TOA}$ across $\theta_{\text{align}} \in [0.2, 0.7]$ and UMAP dimensions (2–40D). The range shows min–max values, and "Best UMAP/θ_{align}" indicates where the maximum occurs.

Tables 6 and 7 summarize the highest majority agreement (MA) across all UMAP dimensionalities (2–40D) and alignment thresholds (θ_{align}), along with the corresponding ranges, grouped by clustering algorithm. HDBSCAN systematically outperforms OPTICS on both MA and Fleiss' κ for the case and delay tasks. The top configuration is HDBSCAN with UMAP 20D and $\theta_{\text{align}} = 0.30$.

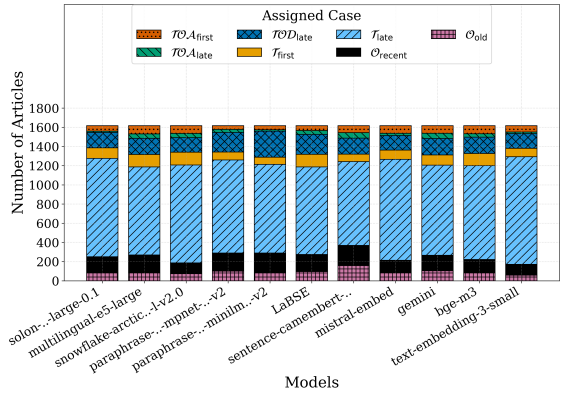


Figure 6: Distribution of taxonomy cases per model for HDBSCAN, UMAP 20D, and $\theta_{\text{align}} = 0.30$.

Under this configuration, Figure 6 shows that most articles are assigned to established topics ($\mathcal{T}_{\text{late}}$), followed by drift outliers ($\mathcal{TOD}_{\text{late}}$), long-standing or recent non-integrated outliers ($\mathcal{O}_{\text{old}}$, $\mathcal{O}_{\text{recent}}$), and at-creation integrations ($\mathcal{T}_{\text{first}}$). Anticipatory outliers ($\mathcal{TOA}_{\text{first}}$, $\mathcal{TOA}_{\text{late}}$) are also observed across all eleven models, with a consistent distribution.

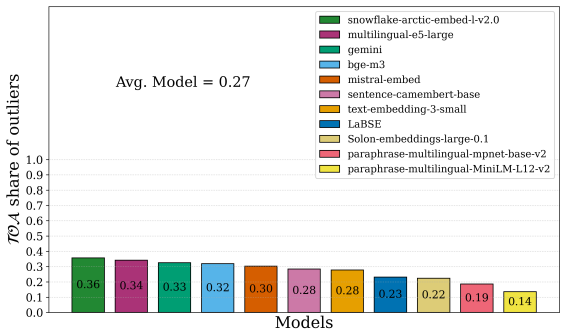
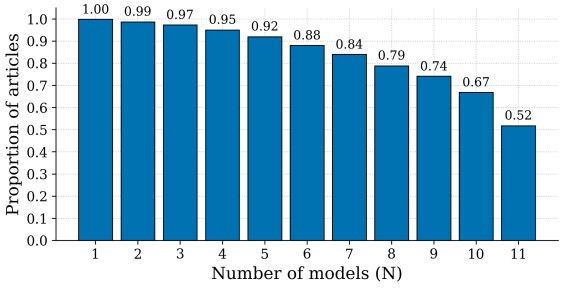
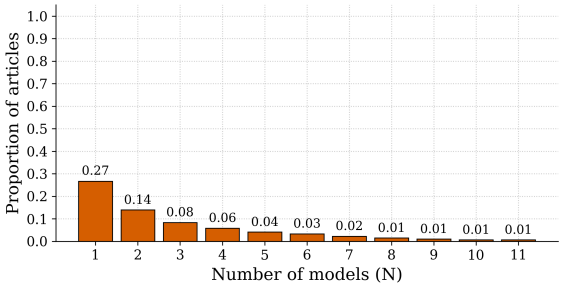


Figure 7: Per-model share of anticipatory outliers ($\mathcal{TOA}$) among all outlier cases ($\mathcal{TO}$, $\mathcal{O}$)

Figure 7 shows that, on average, 27% of outliers are anticipatory outliers ($\mathcal{TOA}$). The highest $\mathcal{TOA}$ proportions are observed for snowflake (36%), e5-large (34%), gemini (33%), and bge-m3 (32%). In contrast, minilm-l12 and mpnet show the smallest shares, 14% and 19%, respectively. For fixed $\theta_{\text{align}} = 0.30$, the average $\mathcal{TOA}$ proportion is stable across UMAP dimensionalities, ranging from 0.27 to 0.32.



(a) Integrated to Topic ($\mathcal{T}$).



(b) Anticipatory Outliers ($\mathcal{TOA}$).

Figure 8: Proportion of articles for which at least N models agree on (a) $\mathcal{T}$ and (b) $\mathcal{TOA}$ cases.

Figures 8a and 8b show cumulative agreement by the number of agreeing models (N). $\mathcal{T}$ assignments are highly consistent: at least $N{=}4$ models agree for 95% of articles, $N{=}7$ for 84%, and all eleven for 52%. In contrast, $\mathcal{TOA}$ unanimity is much lower: at least one model flags 27% of articles, three models flag 8%, six flag 3%, and all eleven flag about 1%. Although $\mathcal{TOA}$ agreement is far lower than for $\mathcal{T}$, it remains a practical measure of robustness. Using thresholds such as $N \geq 6$, or unanimity, isolates a small, high-confidence core of anticipatory outliers.

7. Case Studies

To illustrate the taxonomy's interpretability, we retrospectively analyze three HydroNewsFr topics that include *anticipatory outlier* ($\mathcal{TOA}$) documents. For each topic, we report majority agreement (MA) for both the taxonomy label and the delay under the best-performing configuration reported in Section 6.3 (HDBSCAN, UMAP 20D, θ_{align}=0.3). For anticipatory outliers, anticipation is defined as the time difference $T_T - T_A$ (in days).

Hyundai NEXO (2025) release. This topic captures the launch of Hyundai's second-generation NEXO hydrogen vehicle, officially revealed on 3 April 2025. The topic forms at $T_T =$ 3 April 2025. The earliest signal appears on 21 March at *NewAutoPost* (newautopost.co.kr/fr/), reporting prototype sightings (FR: *"Le véhicule d'essai du NEXO de deuxième génération… repéré."*, EN: *"Second-generation NEXO test vehicle spotted"*), anticipating the topic by 13 days and labeled $\mathcal{TOA}_{\text{first}}$ by a majority of models ($T_A < T_T = T_I$; $MA_{\text{case}} = 11/11$; $MA_{\text{delay}} = 9/11$). On 2 April, the Canadian outlet *RPM* (rpmweb.ca) published prelaunch specifications (FR: *"NEXO à l'hydrogène est plus puissante, plus spacieuse et propose une autonomie d'environ 700 kilomètres"*, EN: *"The hydrogen NEXO is more powerful, more spacious, and offers about 700 km of range"*), anticipating the topic formation by 1 day and labeled $\mathcal{TOA}_{\text{first}}$ ($T_A < T_T = T_I$; $MA_{\text{case}} = 8/11,\ MA_{\text{delay}} = 11/11$). On the release day (3 April), coinciding with topic creation (T_T) for the majority of models, multiple French media such as *H2-Mobile* (h2-mobile.fr; FR: *"Les spécifications clés sont confirmées ce matin."*, EN: *"Key specifications are confirmed this morning"*) and the official Hyundai press release (www.hyundai.news; FR: *"Hyundai dévoile NEXO Nouvelle Génération"*, EN: *"Hyundai unveils new-generation NEXO"*) form and stabilize the cluster, typically as $\mathcal{T}_{\text{first}}$ ($T_A = T_T = T_I$; $MA_{\text{case}} = 5/11,\ MA_{\text{delay}} = 9/11$). Subsequent coverage following topic formation, for example by *Rouler Electrique* (www.rouler-electrique.fr; FR: *"Hyundai mise encore sur l'hydrogène malgré les doutes scientifiques"*, EN: *"Hyundai is still betting on hydrogen despite scientific doubts"*), reinforces the topic on 16 April and is predominantly labeled $\mathcal{T}_{\text{late}}$ ($T_T < T_A = T_I$; $MA_{\text{case}} = 10/11,\ MA_{\text{delay}} = 10/11$).

Safra's Financial Crisis. A second case concerns Safra, a French manufacturer of hydrogen buses, which faced severe financial difficulties before its acquisition by the Chinese group Wanrun. The topic forms at $T_T =$ 28 April 2025. Early signals in local and national media anticipated the crisis: on 8 April, *France Bleu* (francebleu.fr; FR: *"170 emplois menacés à Safra Albi"*, EN: *"170 jobs threatened at Safra Albi"*) is labeled $\mathcal{TOA}_{\text{first}}$ ($T_A < T_T = T_I$), anticipating by 20 days with fair agreement ($MA_{\text{case}} = 7/11,\ MA_{\text{delay}} = 5/11$). On 9 April, *France 3 Occitanie* (france3-regions.francetvinfo.fr; FR: *"avenir incertain des 174 salariés de Safra"*, EN: *"uncertain future for Safra's 174 employees"*) reports an uncertain financial future for Safra and is labeled $\mathcal{TOA}_{\text{first}}$ ($T_A < T_T = T_I$) with 19 days of anticipation and fair delay agreement ($MA_{\text{case}} = 8/11,\ MA_{\text{delay}} = 5/11$). On 16 April, *La Tribune* (latribune.fr; FR: *"l'unique constructeur français de bus à hydrogène au bord de la faillite"*, EN: *"the only French hydrogen-bus manufacturer on the brink of bankruptcy"*) accentuates the importance of the risk and is labeled $\mathcal{TOA}_{\text{late}}$ ($T_A < T_T < T_I$), with 12 days of anticipation ($MA_{\text{case}} = 7/11,\ MA_{\text{delay}} = 5/11$). On the day of topic formation (28 April), *HydrogenToday* (FR: *"Bus à hydrogène : quel avenir pour Safra ?"*, EN: *"Hydrogen buses: what future for Safra?"*) is classified $\mathcal{T}_{\text{first}}$ ($T_A = T_T = T_I$), with low to moderate agreement ($MA_{\text{case}} = 4/11,\ MA_{\text{delay}} = 6/11$), marking the topic's formation. Later coverage, including the 29 April takeover proposal in *La Dépêche* (ladepeche.fr; FR: *"Un groupe asiatique propose de racheter Safra"*, EN: *"An Asian group proposes to buy Safra"*) and the mid-May confirmation of Wanrun's acquisition by *H2-Mobile* (h2-mobile.fr; FR: *"Safra passe sous pavillon chinois"*, EN: *"Safra comes under Chinese ownership"*) is labeled $\mathcal{T}_{\text{late}}$ ($T_T < T_A = T_I$), consolidating the topic with high agreement (e.g., $MA_{\text{case}} \geq 8/11,\ MA_{\text{delay}} = 11/11$).

Vallourec DELPHY launch. This topic covers the certification and commercialization of Vallourec's vertical hydrogen storage system *DELPHY*, with coverage concentrated around 5 June 2025. For most models, topic creation occurs at $T_T =$ 5 June 2025. Two early articles anticipate the topic by multiple weeks: an interview in *La Tribune Dimanche* on $T_A =$ 18 May 2025 is labeled $\mathcal{TOA}_{\text{first}}$ by a majority of models ($MA_{\text{case}} = 8/11$; $MA_{\text{delay}} = 9/11$), anticipating the topic by 18 days and satisfying $T_A < T_T = T_I$. Similarly, *La Voix du Nord* publishes on $T_A =$ 21 May 2025 and is labeled $\mathcal{TOA}_{\text{first}}$ with higher agreement ($MA_{\text{case}} = 9/11$; $MA_{\text{delay}} = 10/11$), anticipating the topic by 15 days and again matching $T_A < T_T = T_I$. On the announcement day ($T_A = T_T =$ 5 June 2025), multiple press and market outlets (e.g., *Fortuneo (AOF)*, *Boursorama*, *Investir*, *Les Échos*, *Connaissance des Énergies*, *GlobeNewswire*) form and stabilize the cluster, typically labeled $\mathcal{T}_{\text{first}}$ ($T_A = T_T = T_I$) with near-unanimous agreement ($MA_{\text{case}} \approx 11/11$; $MA_{\text{delay}} = 11/11$). In this topic, we did not observe $\mathcal{TOD}_{\text{late}}$ items; integrated documents are either anticipatory outliers ($\mathcal{TOA}$), immediate integrations ($\mathcal{T}_{\text{first}}$), or later integrations ($\mathcal{T}_{\text{late}}$).

Across these cases, $\mathcal{TOA}_{\text{first}}$ documents appear

strictly before T_T and integrate when the topic forms ($T_A < T_T = T_I$), providing early evidence that precedes later topic consolidation. Subsequent $\mathcal{T}_{\text{first}}$ and $\mathcal{T}_{\text{late}}$ articles stabilize these developments into coherent topics, illustrating how the taxonomy distinguishes anticipatory signals from later mainstream uptake.

8. Conclusion

We introduced a taxonomy of news trajectories based on three events: document appearance, topic creation, and first integration, and applied it to the HydroNewsFr corpus using a cumulative clustering framework. This taxonomy formalizes *anticipatory outliers*: documents that predate topic formation yet later integrate, in contrast to those that reinforce existing topics or persist as noise. Across eleven embedding models and parameter settings, agreement was fair to moderate for the binary classification task, with only 1% of documents unanimously labeled as anticipatory outliers. The best configuration achieved 0.95 majority agreement and a Fleiss' kappa of 0.33.

The agreement analysis showed that combining multiple models helps mitigate label instability across individual models. A small, high-agreement core of articles suggests that inter-model agreement may serve as a useful proxy for label robustness. The case studies further suggest that anticipatory outliers may provide early signals, often appearing days or weeks before a topic forms.

Beyond retrospective analysis, this work provides a basis for prospective early-warning analysis. Future work will extend the approach to predictive modeling of anticipatory outliers as weak signals of emerging topics. We will also examine the drivers of these transitions and the dynamics of other taxonomy cases. Finally, we will assess whether the framework scales to larger corpora and captures broader thematic patterns beyond specific events.

Limitations

Our analysis is limited to French-language news. This limits the generalizability of the findings beyond this setting. We use daily aggregation to match the news cycle; this granularity may miss slower and longer-term topic shifts. Future work should test the framework on other languages, domains, larger corpora, and alternative time scales.

Ethical Considerations

This study was conducted in line with open-science principles of transparency and reproducibility. When possible, we favor open-source and smaller models in order to reduce computational and environmental costs. The analysis uses short quoted fragments only when necessary for commentary and interpretation, in accordance with the French *exception de courte citation*.

Data and Code Availability

To support reproducibility, we make the scripts for corpus construction and the experimental setup available in a dedicated GitHub repository: https://github.com/evangeliazve/lrec_from_noise_to_signal. The dataset and experimental results can be shared for research purposes on request.

Acknowledgments

We thank the three anonymous reviewers for their insightful comments and suggestions. EZ gratefully thanks Infopro Digital for supporting her PhD at LIP6 by allocating part of her work time to it.

Author Contributions

EZ, GB, JGG conceived the research problem. EZ designed and conducted the experiments. GB and JGG supervised the research and provided guidance. EZ wrote the paper, and all authors reviewed it and revised it.

Bibliographical References

James Allan. 2002. Introduction to topic detection and tracking. In James Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, pages 1–16. Springer, Boston, MA.

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60.

H. Igor Ansoff. 1980. Strategic issue management. *Strategic Management Journal*, 1(2):131–148.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Allaa Boutaleb, Jerome Picault, and Guillaume Grosjean. 2024. BERTrend: Neural topic modeling for emerging trends detection. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 1–17, Miami, Florida, USA. Association for Computational Linguistics.

Clément Christophe, Julien Velcin, Jairo Cugliari, Manel Boumghar, and Philippe Suignard. 2021. Monitoring geometrical properties of word embeddings for detecting the emergence of new topics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 994–1003.

Jason Chuang, Sonal Gupta, Christopher Manning, and Jeffrey Heer. 2013. Topic model diagnostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*, pages 612–620. PMLR.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with pre-trained language models. *Journal of Machine Learning Applications*, 17(3):45–62.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics.

John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108.

Elina Hiltunen. 2008. The future sign and its three dimensions. In *Proceedings of the XXIX Annual Conference of the Finnish Future Society*, pages 1–17.

Sarah Kaplan and Keyvan Vakili. 2015. The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, 36(10):1435–1457.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2024. Antm: Aligned neural topic models for exploring evolving topics. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems LVI: Special Issue on Data Management-Principles, Technologies, and Applications*, pages 76–97. Springer.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE.

Carmen K Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. 2014. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd international conference on World wide web*, pages 527–538.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM.

Evangelia Zve, Benjamin Icard, Alice Breton, Lila Sainero, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2025. From outliers to topics in language models: Anticipating trends in news corpora. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 385–398. Association for Computational Linguistics.

A. Appendix

A.1. Case Frequencies

Table 8 reports the share of anticipatory outliers ($\mathcal{TOA}$) among all outlier documents ($\mathcal{TO} \cup \mathcal{O}$) across UMAP dimensionalities at $\theta_{\text{align}} = 0.30$.

A.2. Agreement Metrics

Table 9 reports Fleiss' κ inter-dimensionality agreement by embedding model and θ_{align}.

Model	HDBSCAN							OPTICS						
	2D	3D	5D	10D	20D	30D	40D	2D	3D	5D	10D	20D	30D	40D
Avg. Model	0.29	0.30	0.32	0.31	0.27	0.29	0.29	0.27	0.33	0.30	0.30	0.31	0.30	0.29
bge-m3	0.25	0.41	0.33	0.33	0.32	0.33	0.34	0.27	0.29	0.33	0.32	0.36	0.31	0.27
gemini	0.36	0.35	0.34	0.37	0.33	0.31	0.31	0.25	0.37	0.28	0.31	0.31	0.29	0.34
LaBSE	0.32	0.30	0.37	0.31	0.23	0.24	0.30	0.31	0.29	0.27	0.24	0.30	0.26	0.30
mistral-embed	0.25	0.33	0.30	0.30	0.31	0.32	0.30	0.28	0.33	0.36	0.34	0.37	0.35	0.33
multilingual-e5-large	0.29	0.27	0.42	0.39	0.34	0.40	0.40	0.30	0.37	0.31	0.35	0.32	0.35	0.29
paraphrase-multilingual-MiniLM-L12-v2	0.39	0.38	0.27	0.11	0.14	0.17	0.22	0.29	0.30	0.27	0.24	0.23	0.25	0.23
paraphrase-multilingual-mpnet-base-v2	0.25	0.13	0.18	0.46	0.19	0.22	0.19	0.25	0.29	0.29	0.24	0.25	0.28	0.24
sentence-camembert-base	0.28	0.30	0.31	0.34	0.28	0.29	0.36	0.24	0.27	0.20	0.22	0.27	0.26	0.25
snowflake-arctic-embed-l-v2.0	0.36	0.31	0.43	0.28	0.36	0.35	0.28	0.28	0.44	0.34	0.42	0.35	0.35	0.26
Solon-embeddings-large-0.1	0.23	0.27	0.28	0.24	0.22	0.29	0.25	0.25	0.31	0.26	0.27	0.32	0.29	0.30
text-embedding-3-small	0.26	0.28	0.28	0.26	0.27	0.29	0.29	0.30	0.33	0.39	0.33	0.36	0.35	0.33

Table 8: Ratio of $\mathcal{TOA}$ among outlier documents across UMAP dimensionalities for HDBSCAN and OPTICS, at $\theta_{\text{align}} = 0.30$.

Model	$\theta_{\text{align}} = 0.2$		$\theta_{\text{align}} = 0.3$		$\theta_{\text{align}} = 0.4$		$\theta_{\text{align}} = 0.5$		$\theta_{\text{align}} = 0.6$		$\theta_{\text{align}} = 0.7$	
	κ	κ_{delay}	κ	κ_{delay}	κ	κ_{delay}	κ	κ_{delay}	κ	κ_{delay}	κ	κ_{delay}
HDBSCAN												
bge-m3	**0.51**	0.38	**0.51**	0.38	**0.51**	0.38	**0.51**	0.38	**0.51**	0.38	0.51	0.38
sentence-camembert-base	0.47	0.32	0.47	0.32	0.47	0.32	0.47	0.32	0.47	0.32	0.47	0.32
gemini	0.47	0.43	0.47	**0.43**	0.47	0.43	0.47	0.43	0.47	0.43	0.45	0.41
multilingual-e5-large	0.49	0.42	0.49	0.42	0.49	0.42	0.49	0.42	0.49	0.42	0.49	**0.42**
mistral-embed	0.51	**0.44**	0.51	0.41	0.50	**0.43**	0.50	**0.43**	0.50	0.41	**0.51**	0.41
text-embedding-3-small	0.46	0.41	0.47	0.40	0.46	0.41	0.46	0.41	0.46	**0.43**	0.45	0.42
Solon-embeddings-large-0.1	0.50	0.39	0.50	0.39	0.50	0.39	0.50	0.39	0.50	0.39	0.50	0.39
LaBSE	0.46	0.38	0.46	0.38	0.46	0.38	0.46	0.38	0.46	0.38	0.46	0.38
paraphrase-multilingual-MiniLM-L12-v2	0.37	0.20	0.37	0.20	0.37	0.20	0.37	0.20	0.37	0.20	0.37	0.20
paraphrase-multilingual-mpnet-base-v2	0.41	0.26	0.41	0.26	0.41	0.26	0.41	0.26	0.41	0.26	0.41	0.26
snowflake-arctic-embed-l-v2.0	0.46	0.40	0.46	0.40	0.46	0.40	0.46	0.40	0.46	0.40	0.46	0.40
Mean	0.46	0.37	0.47	0.36	0.46	0.37	0.46	0.37	0.46	0.36	0.46	0.36
OPTICS												
bge-m3	**0.41**	0.33	**0.41**	0.33	**0.41**	0.33	**0.41**	0.33	**0.41**	0.33	**0.41**	0.33
sentence-camembert-base	0.34	0.24	0.34	0.24	0.34	0.24	0.34	0.24	0.34	0.24	0.34	0.24
gemini	0.36	0.33	0.36	0.33	0.36	0.33	0.36	0.33	0.36	0.33	0.36	0.33
multilingual-e5-large	0.38	0.32	0.38	0.32	0.38	0.32	0.38	0.32	0.38	0.32	0.38	0.32
mistral-embed	0.37	0.33	0.37	0.33	0.39	**0.35**	0.38	0.33	0.37	0.33	0.38	0.33
text-embedding-3-small	0.38	0.32	0.37	0.31	0.37	0.30	0.37	0.31	0.37	0.32	0.37	0.29
Solon-embeddings-large-0.1	0.37	0.29	0.37	0.29	0.37	0.29	0.37	0.29	0.37	0.29	0.37	0.29
LaBSE	0.38	0.28	0.38	0.28	0.38	0.28	0.38	0.28	0.38	0.28	0.38	0.28
paraphrase-multilingual-MiniLM-L12-v2	0.35	0.26	0.35	0.26	0.35	0.26	0.35	0.26	0.35	0.26	0.35	0.26
paraphrase-multilingual-mpnet-base-v2	0.35	0.31	0.35	0.31	0.35	0.31	0.35	0.31	0.35	0.31	0.35	0.31
snowflake-arctic-embed-l-v2.0	0.36	**0.34**	0.36	**0.34**	0.36	0.34	0.36	**0.34**	0.36	**0.34**	0.36	**0.34**
Mean	0.37	0.30	0.37	0.30	0.37	0.30	0.37	0.30	0.37	0.30	0.37	0.30

Table 9: Per-model Fleiss' κ for agreement across UMAP dimensionalities, shown for alignment thresholds $\theta_{\text{align}} \in [0.2, 0.7]$.