Predicting Emerging Topics from Outliers: A Prospective Study of Weak Signals in Embedding Space

Evangelia Zve[1,2]*, Gauvain Bourgne[1], Jean-Gabriel Ganascia[1]
[1]LIP6, Sorbonne Université, CNRS, {name.surname}@lip6.fr
[2]Infopro Digital

Abstract

Some documents that embedding-based topic models initially classify as noise later become founding members of emerging topics. At publication time, however, they appear as scattered points in embedding space and are difficult to distinguish from ordinary noise without the benefit of hindsight. We study whether such *anticipatory outliers* can be predicted prospectively, using only information available when a document first appears. We derive labels from the subsequent trajectories of outlier documents, distinguishing those that anticipate new topics from those that reinforce existing topics or remain isolated, and estimate label confidence through agreement across multiple embedding models. On two French news corpora, anticipatory outliers prove predictable at publication time. Under cross-validation, F_1 rises from about 0.77 over the full eligible population to above 0.90 on high-consensus subsets, and remains at 0.76–0.80 under a strictly chronological evaluation. Predictive performance is driven mainly by geometric features capturing each outlier's position in embedding space.

1 Introduction

In document streams such as news or scientific literature, the challenge is not only to summarize topics that are already visible, but to identify topics as they begin to emerge (Allan, 2002; Boutaleb et al., 2024; Ebadi et al., 2026). Despite recent advances in language models and representation learning, early topic detection remains difficult for embedding-based topic models.

Methods such as BERTopic (Grootendorst, 2022) place documents in a semantic space and cluster them, often using density-based algorithms such as HDBSCAN (McInnes et al., 2017). Documents outside dense regions are usually treated as noise and excluded. Yet these outliers do not all share the same fate (Zve et al., 2025). Some remain isolated, some later join topics that already existed, and some precede topics that had not yet formed.

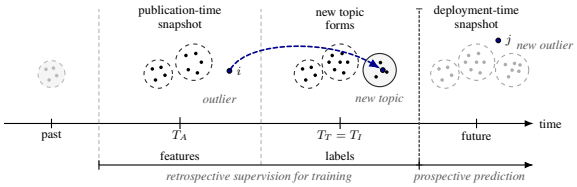


Figure 1: Prospective prediction under retrospective supervision. Labels are reconstructed from later trajectories. Predictors are computed at publication time T_A. Article i illustrates an anticipatory outlier.

Building on the temporal taxonomy of document trajectories introduced by Zve et al. (2026), we focus on documents that appear before the topics they later join. Retrospectively, this anticipatory status is observable once the future topic has formed. Prospectively, however, the same document is still only an outlier in the current embedding space. This gap raises a concrete prediction problem. Can publication-time evidence distinguish ordinary outliers from documents that will later become part of a topic that has not yet formed?

To address this, we reconstruct article trajectories from cumulative daily snapshots of a news stream. On each day, articles observed so far are clustered in embedding space, and the resulting clusters are aligned across consecutive snapshots, to track continuing topics and detect newly formed ones. This reconstruction allows us to determine whether publication-time outliers remain isolated, later join existing topics, or anticipate topics that form only after the article appears.

We formulate anticipatory outlier detection as a supervised binary classification task in which retrospective topic trajectories provide labels, while predictors are restricted to information available when each document first appears. Retrospective

*Corresponding author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1898–1913
November 6-10, 2026 ©2026 Association for Computational Linguistics

topic formation provides the supervision needed to train a prospective model.

We make three contributions. First, we frame anticipatory outlier detection as a prospective supervised article-level prediction task, with labels reconstructed from retrospective topic trajectories across multiple embedding models and confidence estimated through inter-model agreement. Second, we propose and discuss three complementary families of publication-time predictors: geometric position, textual content, and early social circulation of outlier articles. Third, we show that anticipatory outliers are predictable at publication time, both under cross-validation and under a strictly chronological evaluation, and identify the features that drive these predictions.

2 Related Work

Weak-signal analysis studies sparse, ambiguous observations whose significance becomes clear only in retrospect (Hiltunen, 2008; Ansoff, 1975). Related NLP work on Topic Detection and Tracking (Allan, 2002), first-story detection (Petrović et al., 2010), burst detection (Kleinberg, 2002), and social-stream event detection (Becker et al., 2011) seeks early evidence of emerging developments, typically at the event, story, burst, or topic level. In contrast, work on topic-model outliers treats initially unclustered articles as the unit of analysis and asks whether they become early members of emerging topics (Zve et al., 2025, 2026).

Our task is also related to novelty and outlier detection, which identify deviations from regularities (Pimentel et al., 2014; Aggarwal, 2017). Here, outlier status is only the starting point. Rather than detecting outliers per se, we ask which publication-time outliers later become precursors of new topics.

Our work also relates to temporal and embedding-based topic modeling. Classical dynamic topic models track latent topics from word co-occurrence patterns over time (Blei and Lafferty, 2006; Wang and McCallum, 2006). Neural and embedding-based topic models instead represent documents in dense semantic spaces, improving topic discovery through contextual representations (Dieng et al., 2020; Grootendorst, 2022). Recent emerging-topic systems track embedding-space topical structure by aligning clusters across time windows and monitoring nascent topics (Christophe et al., 2021; Boutaleb et al., 2024). These methods are well suited to tracking topics once repeated evidence has accumulated (Ebadi et al., 2026). We instead focus on an earlier stage, when articles remain unclustered and their future topical role is unclear, asking whether these trajectories can be predicted at publication time.

3 Trajectory-Based Supervision

We derive supervised labels from the later trajectories of outlier articles, following the taxonomy of Zve et al. (2026). On each day, all articles published so far are clustered in embedding space, which assigns each article either to a topic or to the outlier set. Aligning topics across consecutive days then tells us whether each topic is continuing or newly created. An article that was an outlier at publication time can therefore end up in one of three situations: it joins a topic that already existed when it appeared, it remains an outlier, or it joins a topic that only formed after its publication. The label is positive in this last, anticipatory case. We run this procedure separately for each embedding model and combine the results in Section 3.3. Figure 1 illustrates the setup with one anticipatory outlier.

3.1 Cumulative Clustering Setting

For each day t, the snapshot contains all articles observed up to and including t. Article texts are embedded, projected with UMAP (McInnes et al., 2018), and clustered using a density-based method. Such methods, including HDBSCAN (McInnes et al., 2017) and OPTICS (Ankerst et al., 1999), can identify low-density observations that do not belong to any cluster. These observations form the candidate outlier cases considered below.

Topic continuity is recovered by aligning clusters in consecutive daily snapshots. Each cluster is represented by the centroid of its article embeddings in the reduced space. The one-to-one matching problem between clusters at $t-1$ and t is solved using the Hungarian algorithm (Kuhn, 1955), with cosine distance between centroids as the matching cost. A cluster at time t inherits the identity of a previous topic if its best match lies below the alignment threshold θ_{align}; otherwise, it is treated as a new topic. This aligned sequence provides the temporal reference used to assign outlier trajectories.

3.2 Trajectories and Prediction Target

Using the notation of Zve et al. (2026), each article trajectory is characterized by three temporal

variables: its publication time T_A, the creation time T_T of the topic it eventually joins, if any, and its first integration time T_I into that topic. The relative ordering of these events determines the trajectory category and defines the classification target.

The *positive class* consists of *anticipatory outliers*: articles that are unassigned at publication time and later join a topic that did not yet exist when they appeared, i.e., $T_A < T_T \leq T_I$. This includes $\mathcal{TOA}_{\text{first}}$, where integration occurs when the topic is created, $T_A < T_T = T_I$, and $\mathcal{TOA}_{\text{late}}$, where the topic is created after publication but the article joins it only later, $T_A < T_T < T_I$.

The *negative class* is the complement of this target within the set of articles that are outliers at T_A. It includes articles that later join a topic that already existed when they appeared, denoted by $\mathcal{TOD}_{\text{late}}$, and articles that remain unassigned throughout the observation window, denoted by $\mathcal{O} = \mathcal{O}_{\text{recent}} \cup \mathcal{O}_{\text{old}}$. The supervised task is therefore the binary distinction $\mathcal{TOA} = \mathcal{TOA}_{\text{first}} \cup \mathcal{TOA}_{\text{late}}$ versus $\neg\mathcal{TOA} = \mathcal{TOD}_{\text{late}} \cup \mathcal{O}$, restricted to articles that are outliers at publication time.

3.3 Consensus-Based Filtering and Labeling

A central difficulty is that trajectory assignment is model-dependent. The embedding space, density structure, and cross-time topic alignment all depend on the upstream embedding model. We therefore do not treat any single model-specific trajectory assignment as definitive supervision. Instead, we run the trajectory-labeling pipeline separately for each embedding model and use inter-model agreement in two ways: to filter reliable publication-time outliers and to assign anticipatory versus non-anticipatory labels. This is aligned with previous work showing that agreement across imperfect labeling sources can serve as a proxy for label reliability (Snow et al., 2008; Ratner et al., 2017; Strehl and Ghosh, 2002).

Let $\mathcal{M}$ denote the set of embedding models. For each article i, let o_i be the number of models that classify the article as an outlier at publication time, and let a_i be the number of models that assign it to an anticipatory trajectory. We define labels using three consensus thresholds (k_o, k_a, k_n), an outlier-eligibility threshold k_o, a positive-label threshold k_a, and a negative-label threshold k_n.

$$y_i = \begin{cases} 1, & o_i \geq k_o \text{ and } a_i \geq k_a, \\ 0, & o_i \geq k_o \text{ and } a_i \leq k_n, \\ \text{uncertain}, & \text{otherwise.} \end{cases}$$

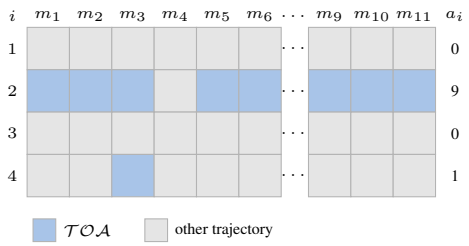


Figure 2: Consensus labeling example. Rows are articles $i = 1, \ldots, 4$, columns m_1–m_{11} are the embedding models in $\mathcal{M}$, and a_i is the number of models assigning the $\mathcal{TOA}$ trajectory to article i (shaded cells). The four articles have already passed the eligibility filter $o_i \geq k_o$. Under $(k, k, 0)$, article 2 is positive for $k \leq 9$, article 4 only for $k = 1$, and articles 1 and 3 are negative for every k.

The condition $o_i \geq k_o$ defines the eligible population: articles reliably identified as publication-time outliers. Within this population, positives require at least k_a anticipatory votes, while negatives require at most k_n such votes. Articles satisfying neither rule are treated as *uncertain* and excluded from training and evaluation. Figure 2 illustrates the rule on four articles.

In our experiments, we use the conservative setting $(k_o, k_a, k_n) = (k, k, 0)$. Thus, both outlier status and positive labels must be supported by at least k models, whereas negatives must receive no anticipatory vote. Varying k quantifies the coverage–confidence tradeoff. Larger values retain fewer articles but provide higher-confidence supervision. More general asymmetric rules are possible, but varying the three thresholds independently introduces additional calibration choices; we leave their systematic investigation to future work.

4 What Predicts Anticipatory Outliers?

If anticipatory outliers are meaningful early signals rather than residual noise, they should already leave detectable traces at T_A. We therefore propose and discuss three complementary families of T_A-based predictors, each capturing a different article-level source of evidence about later topic formation.

Geometric features locate the article in the publication-time embedding space, measuring whether it sits at the margins of established topic regions or in sparse residual areas where new structure may later form. Textual features characterize the article's content, tone, and form, testing whether anticipatory trajectories are associated with distinctive framing, entity anchoring, or stylis-

tic cues. Social features capture early circulation on X, testing whether articles that later join emerging topics already leave observable traces in audience reach or audience-overlap structure. All predictors use only T_A features, while labels are derived from later trajectories to prevent future-topic leakage.[1]

4.1 Geometric Features

Geometric features measure where an outlier lies in the UMAP-reduced embedding space of the cumulative topic snapshot at T_A. Let x_i denote the reduced embedding of article i, and let C_{T_A} denote the set of topic clusters present at T_A.

We first measure proximity to existing topics. For each cluster $c \in C_{T_A}$, we compute its centroid μ_c and the Euclidean distance from x_i to μ_c. The nearest and second-nearest distances, d_{1i} and d_{2i}, measure how far the article lies from the topic structure at T_A. Their difference, $\Delta_i = d_{2i} - d_{1i}$, defines the centroid margin: large values indicate proximity to one dominant topic, while small values indicate comparable proximity to multiple topics. Figure 3(a) illustrates this geometry.

Centroid distances treat all clusters as equally compact, even though some topics are more dispersed than others. We therefore compute a diagonal Mahalanobis distance to each existing topic cluster containing at least five articles, and retain the minimum value

$$d_i^{\text{Mah}} = \min_{\substack{c \in C_{T_A} \\ |c| \geq 5}} \sum_j \frac{(x_{ij} - \mu_{cj})^2}{\sigma_{cj}^2 + \epsilon},$$

where σ_{cj}^2 is the coordinate-wise variance of cluster c, and ϵ prevents division by zero. This feature measures whether x_i is atypical relative to the dispersion of the nearest plausible topic, rather than only distant in Euclidean space.

We then measure local density around the article. Using only articles observed in the cumulative snapshot at T_A, we compute the mean Euclidean distance $\bar{d}_i^{\text{kNN}}$ from x_i to up to 20 nearest neighbors, depending on the number of articles available in the snapshot; see Figure 3(b). We also compute the standard deviation of these distances to capture neighborhood heterogeneity and distinguish uniformly sparse regions from mixed neighborhoods containing both close and distant neighbors.

A further group of features captures proximity to other outliers. Let $O(T_A)$ denote the set of articles

[1]Full glossary of features is available in Appendix B.2.

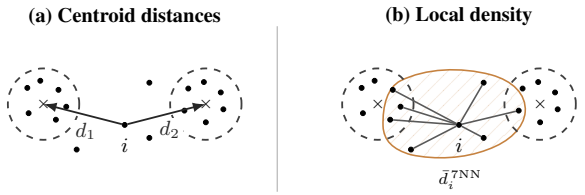


Figure 3: Geometric predictors at publication time T_A: (a) distances from outlier i to the nearest topic centroids and (b) an illustrative local-neighborhood configuration.

classified as outliers in the cumulative snapshot at T_A. We record whether this set is non-empty, its size, and the mean Euclidean distance from x_i to up to ten nearest articles in $O(T_A)$. These features distinguish outliers that are isolated in the embedding space from outliers that lie close to other publication-time outliers.

Finally, we include the *outlierness score* at T_A, computed from the GLOSH score returned by HDBSCAN (Campello et al., 2015; McInnes et al., 2017). It measures how strongly an article is separated from dense regions in the HDBSCAN hierarchy. Higher values indicate that the article is weakly connected to nearby clusters. Unlike centroid-distance features, it captures local density structure rather than distance to topic centers.

Because each embedding model induces its own UMAP space, raw distance values are not directly comparable across models. We therefore convert each distance-based geometric feature to a within-model percentile rank. For feature g, article i, and embedding model m, let $g_{im}(T_A)$ be the publication-time value and let $n_m(T_A)$ be the number of rows in the corresponding model-specific feature table. We compute

$$p_{im}^{(g)}(T_A) = \frac{\text{rank}_{m,T_A}(g_{im}(T_A))}{n_m(T_A) + 1},$$

using average ranks for ties. This yields percentile-ranked features in $(0, 1)$, where larger values indicate larger within-model distances.

Article-level geometric predictors are obtained by aggregating model-specific feature values for each article. For each geometric feature, we compute the mean, median, and standard deviation across embedding models to capture average position, central tendency, and cross-model variation.

4.2 Textual Features

Textual features measure whether anticipatory outliers are distinctive in content or form. They are computed once per article from its textual content.

We first compute tone features. Subjectivity is obtained with the French version of TextBlob (Loria, 2018). Neutrality is derived from the French VADER compound score (Hutto and Gilbert, 2014). Articles with strong positive or negative polarity receive lower neutrality scores, whereas articles with weak affective polarity receive higher scores.

We then compute length and surface-complexity features. Character and word count measure the amount of textual information available. Average sentence length in words, average word length in characters, total syllables, and average syllables per word describe the surface complexity and density of this information.

Finally, we extract named entities with the French `spaCy` pipeline (Honnibal et al., 2020). The total number of entities measures the degree of entity anchoring in the article. The number of distinct entities captures how varied this anchoring is. We also retain counts for persons, organizations, locations, and miscellaneous entities, which indicate whether the article is tied to specific actors, institutions, places, or other named references. Such anchoring may precede cluster formation when later articles repeatedly refer to the same entities.

4.3 Social Features

Social features capture early circulation on X using URL-sharing events observed no later than the publication-time cumulative snapshot at T_A. If an article has no observed social trace by T_A, all social variables are set to zero, so that the absence of early circulation is encoded explicitly.

We first measure direct sharing activity. For each article URL, we record the number of distinct users who shared it, the median follower count, tweet count, and listed count of these users to capture the breadth and typical visibility of early circulation.

We then encode audience overlap through a weighted article–article co-sharing graph, obtained as a one-mode projection of the underlying user–article bipartite graph. Nodes correspond to article URLs observed by T_A. Two URLs are connected when at least one user shared both URLs by T_A, and the edge weight is the number of shared users. The derived metrics capture the extent to which new outlier articles circulate through overlapping audiences before a topic is semantically consolidated (Granovetter, 1973; Ugander et al., 2012).

From this graph we retain three measures. The weighted clustering coefficient (Onnela et al., 2005) indicates whether the article belongs to a locally coherent co-sharing neighborhood. The Louvain community size (Blondel et al., 2008) measures the scale of the audience group in which the article circulates; the partition is computed with a fixed random seed, so that community assignments are reproducible. The bridge ratio measures the share of an article's co-sharing weight that connects outside its own community.

5 Experimental Setup

5.1 Datasets

We evaluate the prediction task on two French news corpora. HYDRONEWSFR, described in Zve et al. (2026), focuses on hydrogen-energy narratives. We curated CLIMATENEWSFR, a climate-change corpus collected with the same pipeline. It combines two complementary streams: news articles retrieved from Google News results using the GNews Python library, and X posts linking to news articles, collected through the official X API. The corpus was collected using the French keyword "changement climatique" ("climate change") and includes the available title and lead paragraph.

Corpus	Period	Articles	X posts
HYDRONEWSFR	20 Mar–8 Jun 2025	1,616	1,533
CLIMATENEWSFR	27 Mar–27 May 2025	2,445	1,046

Table 1: Summary of the two main corpora.

Both corpora are temporally ordered, multi-source news streams with limited coverage gaps. This property is important for the present task, since the transition of publication-time outliers into later topics can only be evaluated reliably when the underlying stream is dense enough to support temporal continuity. CLIMATENEWSFR is broader and more heterogeneous, covering political, social, and environmental aspects of climate change, whereas HYDRONEWSFR is centered on hydrogen-related industrial, technological, and policy developments.

5.2 Labeling-Pipeline Configuration

In all experiments, we use an ensemble of 11 French-specific and multilingual embedding models. The full list is given in Appendix B.1. For each model, we concatenate each article's title and lead paragraph, embed the resulting text, project the embeddings to 20 dimensions with UMAP, cluster them daily with HDBSCAN, and align clusters across time using centroid matching with threshold $\theta_{\text{align}} = 0.30$. This configuration follows the

setting reported by Zve et al. (2026), which was found to yield strong inter-model agreement on the $\mathcal{TOA}$ versus non-$\mathcal{TOA}$ distinction. We adapt their publicly available code to derive trajectory labels.[2] We keep the HDBSCAN library defaults, `min_cluster_size=5` and `min_samples=5`.

5.3 Supervised Learning and Evaluation

The supervised dataset pairs the consensus labels from Section 3 with the article-level publication-time predictors from Section 4. Articles with uncertain consensus labels are excluded from supervised training and evaluation. We also discard the first five days of each corpus as a warm-up window before training. We compare five classifiers: logistic regression with ℓ_2 regularization (Hoerl and Kennard, 1970), a linear support vector classifier (Cortes and Vapnik, 1995), a decision tree (Breiman et al., 1984), a random forest (Breiman, 2001), and XGBoost (Chen and Guestrin, 2016).

Because anticipatory outliers are the rarer class, we use F_1 as the primary metric and report recall and precision as complementary measures. We estimate performance with two protocols. The main protocol is 5-fold article-level cross-validation, which measures how much information publication-time features carry about later trajectories. We complement it with a chronological forward-chaining protocol, in which each article is predicted by a model trained only on strictly earlier articles, and which therefore approximates deployment conditions. We also report a constant-positive baseline that predicts $\mathcal{TOA}$ for all eligible articles. The complete setup is provided in Appendix B.3.

6 Results

We examine whether anticipatory outliers can be predicted from information available at publication time. The analysis proceeds in three steps. We first vary the consensus threshold used to filter and label publication-time outliers, assessing how prediction changes as supervision is supported by stronger agreement across embedding models. We then compare multiple classifier families under the selected consensus settings, to determine whether the signal is specific to one learning algorithm or remains stable across modeling choices. Finally, we evaluate prediction chronologically.

[2]Code: https://github.com/evangeliazve/aacl-public

6.1 Effect of Agreement Thresholds

Table 2 reports the coverage–performance tradeoff obtained by varying the embedding-model consensus threshold k, using XGBoost at T_A. As introduced in Section 3.3, k specifies how many embedding models must agree that an article is a publication-time outlier and, for positive labels, that it follows an anticipatory trajectory ($\mathcal{TOA}$).

Increasing k makes the supervision set more selective. Fewer articles are retained, but their labels are supported by stronger inter-model agreement. At $k = 1$, the model already reaches $F_1 = 0.765$ on HydroNewsFr and $F_1 = 0.778$ on ClimateNewsFr. Performance tends to improve under stricter agreement, consistent with reduced label noise, but high thresholds retain few positives and show greater fold-level variability. For the detailed classifier and interpretability analyses, we use $k = 4$ for HydroNewsFr (83 positives, 254 negatives) and $k = 6$ for ClimateNewsFr (68 positives, 330 negatives). These are the largest agreement thresholds before the positive class becomes very small and fold-level estimates become visibly less stable, as shown in Table 2.

6.2 Prediction at Publication Time

Using the selected corpus-specific values of k, we train and compare five classifier families on the corresponding article-level labeled subsets. Table 3 reports cross-validated performance at T_A.

All trained models substantially outperform the constant-positive baseline, indicating that publication-time features contain information about later integration into newly formed topics when trajectory labels are supported by consistent evidence across embedding models. XGBoost obtains the highest F_1-score in both corpora: 0.912 ± 0.069 in HydroNewsFr and 0.969 ± 0.047 in ClimateNewsFr. Linear models are more recall-oriented, with Linear SVC and logistic regression reaching recall 0.941 ± 0.064 in HydroNewsFr, and logistic regression reaching recall 0.985 ± 0.031 in ClimateNewsFr.

6.3 Chronological Evaluation

Cross-validation lets a model see articles published after those it is tested on. To assess performance under deployment conditions, we additionally evaluate XGBoost with chronological forward chaining: at each cutoff t, the model is trained on all labeled articles with $T_A \leq t$ and tested on those published

	HydroNewsFr				ClimateNewsFr			
k	Retained	Positive	Negative	$F_1 \pm$ std	Retained	Positive	Negative	$F_1 \pm$ std
1	1235	544	691	0.765 ± 0.013	1941	875	1066	0.778 ± 0.012
2	773	276	497	0.811 ± 0.056	1390	489	901	0.851 ± 0.016
3	508	150	358	0.895 ± 0.022	1005	273	732	0.895 ± 0.020
4	337	83	254	$\mathbf{0.912 \pm 0.069}$	766	182	584	0.915 ± 0.038
5	219	39	180	0.860 ± 0.116	570	115	455	0.952 ± 0.033
6	145	21	124	0.910 ± 0.092	398	68	330	$\mathbf{0.969 \pm 0.047}$
7	81	13	68	0.917 ± 0.144	270	43	227	0.940 ± 0.016
8	40	6	34	0.800 ± 0.400	161	21	140	0.949 ± 0.063

Table 2: Effect of the agreement threshold k at T_A with XGBoost. Scores are cross-validated means $\pm$ standard deviations across folds. Bold marks the F_1 scores of the selected k values used in the detailed analyses.

	HydroNewsFr			ClimateNewsFr		
Model	Precision	F_1-score	Recall	Precision	F_1-score	Recall
Baseline	0.246	0.395	1.000	0.171	0.291	1.000
Linear SVC	0.853	0.889	**0.941**	0.889	0.906	0.930
LogReg ℓ_2	0.838	0.882	**0.941**	0.895	0.936	**0.985**
Decision Tree	0.803	0.835	0.891	0.867	0.898	0.937
Random Forest	0.867	0.885	0.917	0.905	0.935	0.969
XGBoost	**0.919**	**0.912**	0.917	**0.970**	**0.969**	0.969

Table 3: Article-level precision, F_1-score, and recall at T_A under selected consensus settings: $k = 4$ for HydroNewsFr and $k = 6$ for ClimateNewsFr. Results are averaged over five article-level CV folds.

Corpus	Pos.	Neg.	F_1 [95% CI]	Base
HydroNewsFr	42	211	0.762 [0.65, 0.86]	0.285
ClimateNewsFr	26	316	0.800 [0.66, 0.91]	0.141

Table 4: Forward-chaining results at T_A with XGBoost and $W = 7$-day test windows, under the selected consensus settings. Cells report F_1 with a 95% percentile bootstrap interval and the constant-positive baseline.

in the following W days, so that each evaluated article is predicted exactly once by a model that has seen only strictly earlier articles. Table 4 reports pooled F_1 for $W = 7$ days; the full protocol and other window lengths are given in Appendix E.

As expected, chronological scores are lower than cross-validated ones, since each model is trained on fewer articles and cannot exploit any later structure of the stream. Both corpora nevertheless remain well above the constant-positive baseline even at the lower confidence bound, and results are stable across window lengths on the longer extended corpus. The cross-validated estimates should thus be read as an upper bound on the information carried by publication-time features, and the chronological estimates as a more conservative indication of prospective performance.

7 Interpretability Analysis

We analyze the learned XGBoost models from the cross-validated setting of Section 6.2 to identify which publication-time features contribute to $\mathcal{TOA}$ predictions. We first compare feature families through ablation, then inspect global SHAP importance and local out-of-fold explanations.

7.1 Ablation Study

Table 5 reports feature-family ablations for XGBoost at T_A. "Only" rows use a single feature family, whereas "All w/o" rows remove one family from the full feature set. To test whether the main contrasts are stable across folds, we compare paired fold-level F_1 scores from the same five article-level cross-validation splits using paired t-tests, with Benjamini–Hochberg correction (Benjamini and Hochberg, 1995).

Geometry accounts for most of the predictive signal. Using geometry alone nearly matches the full model in both corpora, with no significant F_1 decrease in HydroNewsFr (0.900 versus 0.912) and no decrease in ClimateNewsFr (0.969 versus 0.969). In contrast, removing geometry yields a large and significant drop, from 0.912 to 0.385 in HydroNewsFr and from 0.969 to 0.226 in ClimateNewsFr. The precision–recall columns show that this effect is not driven by only one side of the tradeoff, as removing geometry lowers both precision and recall in both corpora.

Text and social features are weak as standalone predictors in both corpora, and both perform significantly below the geometry-only model. Removing either family from the full model does not significantly reduce F_1. This suggests that these features provide at most auxiliary, corpus-specific information. The partial exception is social-only prediction in HydroNewsFr, with moderate precision but

	HydroNewsFr			ClimateNewsFr		
Feature set	Precision	F_1	Recall	Precision	F_1	Recall
All features	0.919	0.912	0.917	0.970	0.969	0.969
Only geometry	0.907	0.900	0.905	0.970	0.969	0.969
Only text	0.322	$0.272^{\ddagger}$	0.242	0.279	$0.202^{\ddagger}$	0.165
Only social	0.707	$0.349^{\ddagger}$	0.239	0.235	$0.083^{\ddagger}$	0.212
All w/o geometry	0.406	$0.385^{\dagger}$	0.371	0.285	$0.226^{\dagger}$	0.194
All w/o social	0.930	0.917	0.917	0.970	0.969	0.969
All w/o text	0.919	0.905	0.904	0.969	0.962	0.955

Table 5: Ablations for XGBoost at T_A. Results are 5-fold CV means. Symbols mark significant F_1 drops at $\alpha = 0.05$: † vs. all features; ‡ vs. geometry only.

very low recall. This pattern indicates that social features help identify a small subset of anticipatory cases, but do not characterize the class.

7.2 SHAP Analysis

7.2.1 Global Importance

Table 6 reports the top global SHAP features for the XGBoost models, computed with TreeExplainer (Lundberg et al., 2020). Mean absolute SHAP values measure each feature's average contribution magnitude, while Spearman correlations between feature values and signed SHAP values indicate whether larger feature values push predictions toward or away from $\mathcal{TOA}$.

The top-ranked features are geometric in both corpora. In HydroNewsFr, the largest contribution comes from the median second-nearest-centroid distance ($|\text{SHAP}| = 0.1018$), followed by the median nearest-centroid distance ($|\text{SHAP}| = 0.0544$). In ClimateNewsFr, the strongest feature is the median standard deviation of 20-nearest-neighbor distances ($|\text{SHAP}| = 0.0773$), followed by median nearest- and second-nearest-centroid distances. These features are positively correlated with signed SHAP values, indicating that articles farther from existing topic centroids, or located in sparser and more heterogeneous local neighborhoods, receive higher $\mathcal{TOA}$ scores.

The number of outliers present in the snapshot at T_A shows the opposite pattern. In both corpora, larger outlier counts push predictions away from $\mathcal{TOA}$, with strong negative correlations in HydroNewsFr ($r = -0.81$ and $r = -0.87$) and ClimateNewsFr ($r = -0.91$ and $r = -0.94$). Thus, the model does not treat the number of outliers present at T_A as evidence of anticipation. High $\mathcal{TOA}$ scores are instead associated with articles that tend to be distant from existing topic centroids, lie in sparse and heterogeneous local neighborhoods, and receive higher HDBSCAN outlierness scores.

Textual and social features have smaller and more corpus-specific effects, as shown in Table 7. In HydroNewsFr, the strongest textual feature is miscellaneous entity count, while total and person-entity counts are negatively associated with $\mathcal{TOA}$. In ClimateNewsFr, subjectivity is the strongest textual feature. Social features contribute weakly in HydroNewsFr, led by weighted clustering in the article–article co-sharing graph, and have zero mean absolute SHAP value in ClimateNewsFr. These patterns support the ablation result, with non-geometric features providing auxiliary information rather than the main predictive signal.

Feature	Mean $\lvert\text{SHAP}\rvert$	Spearman r
HydroNewsFr		
`ner_misc`	0.0113	0.82^{***}
`ner_total_ents`	0.0097	-0.79^{***}
`ner_person`	0.0082	-0.81^{***}
`text_subjectivity`	0.0071	0.82^{***}
`media_weighted_clustering`	0.0027	0.53^{***}
`media_community_size`	0.0022	0.49^{***}
ClimateNewsFr		
`text_subjectivity`	0.0027	0.73^{***}
`ner_misc`	0.0025	-0.81^{***}

Table 7: Main non-geometric SHAP features at T_A. Significance coding: $^{***}p < 0.001$.

7.2.2 Local Explanations

We examine one true-positive out-of-fold example from each corpus, selected to illustrate substantively interpretable later topics. Each article is predicted by a fold-specific model that was not trained on it. Retrospective topic reconstruction is then used to interpret the topic the article later joined.

HydroNewsFr. We consider the *H2 Mobile* article on Germany's use of salt caverns for large-scale hydrogen storage.[3] Its out-of-fold predicted probability at T_A is 0.9983. It is labeled positive, with 5 votes across the embedding-model ensemble. In the retrospective `multilingual-e5-large` reconstruction, T_A is 25 April 2025, while $T_T = T_I =$ 9 May 2025. The article therefore appears before the topic becomes a cluster in this reconstruction. The later topic is centered on large-scale underground hydrogen storage in salt formations and includes articles on Storengy's storage projects, German cavern capacity, and pilot demonstrations in saline formations.[4]

[3] *H2 Mobile*, "L'Allemagne mise sur ses cavernes salines pour le stockage massif d'hydrogène," (25 April 2025).

[4] *France 3 Régions*, "Un site pilote européen de l'Ain valide

HydroNewsFr				ClimateNewsFr			
Feature	Cat.	Mean |SHAP|	Spearman r	Feature	Cat.	Mean |SHAP|	Spearman r
d2_second_centroid_pct_median	G	0.1018	0.69^{***} ↑	knn_std_k20_pct_median	G	0.0773	0.28^{***} ↑
d1_nearest_centroid_pct_median	G	0.0544	0.64^{***} ↑	d1_nearest_centroid_pct_median	G	0.0619	0.48^{***} ↑
d2_second_centroid_pct_mean	G	0.0509	0.67^{***} ↑	d2_second_centroid_pct_median	G	0.0495	0.38^{***} ↑
d1_nearest_centroid_pct_mean	G	0.0420	0.84^{***} ↑	n_recent_outliers_mean	G	0.0255	-0.91^{***} ↓
knn_mean_k20_pct_median	G	0.0417	0.48^{***} ↑	d1_nearest_centroid_pct_mean	G	0.0145	0.86^{***} ↑
knn_mean_k20_pct_mean	G	0.0264	0.65^{***} ↑	knn_mean_k20_pct_median	G	0.0139	0.38^{***} ↑
n_recent_outliers_median	G	0.0136	-0.81^{***} ↓	outlier_score_std	G	0.0103	0.67^{***} ↑
outlier_score_median	G	0.0135	0.89^{***} ↑	outlier_score_mean	G	0.0074	0.67^{***} ↑
knn_std_k20_pct_mean	G	0.0129	0.36^{***} ↑	d2_second_centroid_pct_mean	G	0.0067	0.37^{***} ↑
n_recent_outliers_mean	G	0.0119	-0.87^{***} ↓	n_recent_outliers_std	G	0.0066	-0.94^{***} ↓

Table 6: Top XGBoost features, ranked separately by mean absolute SHAP value. Arrows indicate the sign of the Spearman correlation between feature value and SHAP contribution. G = geometry, T = text, S = social. Significance coding: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

The strongest positive contributions come from distances to existing topic structure, especially the median second-nearest-centroid distance $(+0.2034)$, the median nearest-centroid distance $(+0.1254)$, and the median 20-nearest-neighbor distance $(+0.1087)$. The median outlier-score also contributes positively $(+0.0382)$. Together, these features place the article away from existing hydrogen-topic centers and in a sparse neighborhood. Non-geometric effects are small: subjectivity (-0.0165) and the number of named entities (-0.0094) push slightly away from $\mathcal{TOA}$, while the number of unique users sharing the article contributes positively $(+0.0033)$.

ClimateNewsFr. We consider the *CNRS* article on archaeological sites threatened by climate change.[5] Its out-of-fold predicted probability at T_A is 0.9978. It is labeled positive, with 10 votes. In the retrospective `mistral-embed` reconstruction, $T_A = 2$ April 2025, while $T_T = T_I = 17$ April 2025, so this article is a $\mathcal{TOA}_{\text{first}}$ case. The article therefore appears before climate-related cultural heritage and archaeology becomes a cluster in this reconstruction. The later topic includes articles on Greek historical sites threatened by climate change and on the transformation of cultural heritage worldwide under climate change, all published later than the CNRS article.[6]

The largest positive contributions are the median nearest-centroid distance $(+0.2019)$, the median standard deviation of 20-nearest-neighbor distances $(+0.1983)$, and the mean second-nearest-centroid distance $(+0.1038)$. The number of outliers at T_A also supports the prediction: consistent with the negative global correlation reported above, the article was published when the number of outliers at T_A was low, and this low count contributes positively $(+0.0832)$. HDBSCAN outlier-score variability contributes more modestly $(+0.0327)$. These features place the article away from existing climate topics, in a sparse and heterogeneous neighborhood. Textual effects are much smaller: average sentence length is the main negative correction (-0.0109), while subjectivity contributes weakly toward $\mathcal{TOA}$ $(+0.0052)$. Social features do not materially affect this local prediction.

In both examples, $\mathcal{TOA}$ predictions are driven by distance from existing topic centers combined with a non-uniform local neighborhood at T_A.

8 Conclusion

We introduced *anticipatory outlier detection*, a prospective task that predicts if a publication-time outlier will later become an early member of a topic that has not yet formed. On two French news corpora, we show that this trajectory is predictable from publication-time information, especially under stricter inter-model consensus labels, and that the signal persists under a strictly chronological evaluation. Prediction is driven mainly by embedding-space geometry: anticipatory outliers lie far from existing topic centroids, in sparse and uneven neighborhoods. Textual and social variables provide weaker, more corpus-specific signals.

More broadly, this work raises the question of whether anticipatory outliers reflect a general property of dynamic embedding spaces beyond news, and even beyond text. Similar signals may arise in images, video, or audio, when initially isolated representations later become part of emerging semantic or stylistic structures.

le stockage de l'hydrogène dans du sel," (26 May 2025); *EnergyNews*, "Le stockage d'hydrogène en cavités salines : opportunités et défis pour l'industrie," (23 May 2025).

[5] *CNRS*, "Réchauffement climatique : l'archéologie face à la disparition des sites et des fouilles," (2 April 2025).

[6] *Courrier International*, "En Grèce, les sites historiques menacés par le changement climatique," (23 April 2025); *La Croix*, "Le changement climatique transforme notre patrimoine aux quatre coins de la planète," (17 April 2025).

Limitations

Our methodology provides evidence that publication-time outliers can contain signals of later topic formation within an embedding-based framework. The evaluation relies on consensus labels obtained across embedding models, mitigating dependence on any single embedding space. Qualitative examples further indicate that several predicted outliers correspond to coherent later topics. Nevertheless, external validation remains necessary. A systematic human-annotation study would help assess how these signals are perceived in terms of novelty, relevance, and real-world significance.

The study relies on two French news corpora with complementary scopes. HYDRONEWSFR is an existing dataset focused on a specialized industrial and policy domain, while CLIMATENEWSFR is a curated dataset covering a broader and more heterogeneous public issue. Both corpora are temporally ordered news streams built from Google News results and X-sharing activity of news articles, with articles drawn from multiple media sources. This provides relatively dense time series, reduces timeline discontinuities, and limits dependence on a single source while keeping the language and collection pipeline controlled across settings. However, the comparison remains limited to two domains within a single language and does not cover other media ecosystems or longer temporal scales. Replication on larger multilingual corpora and longer observation windows would further assess the transferability of the approach.

The supervised labels depend on the retrospective trajectory-reconstruction pipeline. In the main setting, embeddings are projected with 20-dimensional UMAP, clustered with HDBSCAN, and aligned across cumulative snapshots using a fixed topic-alignment threshold ($\theta_{\text{align}} = 0.30$). This configuration follows prior trajectory-labeling work showing strong inter-model agreement for the anticipatory versus non-anticipatory distinction, with limited variation across UMAP dimensionalities (Zve et al., 2026). However, the present study adds a downstream prospective prediction task, so robustness at the labeling stage may not fully imply robustness of prediction scores. Future work should therefore test whether the observed stability across UMAP dimensionalities also holds for prediction, and should examine alternative dimensionality-reduction methods, clustering algorithms, and alignment thresholds.

Finally, social features are limited to observed X-sharing activity, which provides only a partial view of online circulation. This limited coverage may partly explain the weak contribution of social features in our models. Early diffusion may also occur on other social-media platforms or through channels not captured in the present data. Future work should extend the analysis to additional platforms and circulation channels.

Ethical Considerations

In line with open-science principles, the code for reproducing the experiments is publicly available in a dedicated GitHub repository (Appendix A). The embedding ensemble combines open-source and API-based models to assess robustness across model families used in research and applied settings. Where possible, we favor compact models to reduce unnecessary computational cost and environmental impact.

The study uses news articles retrieved from Google News results using the GNews Python library, and social-media sharing traces collected through the official X API. We do not redistribute raw text or raw social-media traces, since these may contain identifiable user activity and may be subject to publishers' rights, platform terms, personal-data restrictions, or copyright constraints. Data can be shared for research purposes only upon reasonable request.

The method is intended for research on weak-signal detection in dynamic text streams, not as a deployable monitoring or decision system. Its outputs are corpus-level signals, not evidence that an article is objectively important, novel, or likely to shape future events. Human annotation, external validation, and calibration would be required before any applied use. The main practical risk is a biased allocation of attention: some predictions may reflect corpus-specific patterns, and other relevant documents may be missed. Any adaptation should document data provenance, media coverage, language scope, and source selection, and account for uncertainty.

Acknowledgments

EZ thanks Infopro Digital for granting her the time to pursue her PhD thesis alongside her work.

References

Charu C. Aggarwal. 2017. *Outlier Analysis*, 2 edition. Springer, Cham.

James Allan, editor. 2002. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers, Boston, MA.

Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, pages 49–60. ACM.

H. Igor Ansoff. 1975. Managing strategic surprise by response to weak signals. *California Management Review*, 18(2):21–33.

Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 438–441.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

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.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

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

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. 2015. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):1–51.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

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. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Ashkan Ebadi, Alain Auger, and Yvan Gauthier. 2026. WISDOM: An AI-powered framework for emerging research detection using weak signal analysis and advanced topic modelling. *Journal of Informetrics*, 20(1):101759.

Mark S. Granovetter. 1973. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Elina Hiltunen. 2008. The future sign and its three dimensions. *Futures*, 40(3):247–260.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

C. J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Jon Kleinberg. 2002. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 91–101. ACM.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Steven Loria. 2018. TextBlob documentation. `https://textblob.readthedocs.io/`. Release 0.15.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 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.

Jukka-Pekka Onnela, Jari Saramäki, János Kertész, and Kimmo Kaski. 2005. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics.

Marco A. F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A review of novelty detection. *Signal Processing*, 99:215–249.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. 2012. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966.

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, Gauvain Bourgne, Benjamin Icard, and Jean-Gabriel Ganascia. 2026. From noise to signal: When outliers seed new topics. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 7523–7533, Palma de Mallorca, Spain. ELRA Language Resource Association.

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*, pages 385–398. Association for Computational Linguistics.

A Supplementary Materials

We provide supplementary materials at: https://github.com/evangeliazve/aacl-public. The repository includes the pipeline for applying the method to other corpora, together with the scripts used to reproduce the classification, ablation, SHAP, and forward-chaining results reported in the paper.

Data collection used GNews for Google News results[7] and the official X API[8].

B Supplementary Experimental Details

B.1 Embedding Models

Table 8 lists the embedding models used for trajectory reconstruction.

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

Table 8: Embedding models used in the experiments.

B.2 Feature Glossary

Table 9 presents the features proposed and discussed in this study, complementing Section 4. Geometric features are per-model quantities; the classifiers use their mean, median, and standard deviation across the eleven embedding models, aggregated as described in Section 4.1. Social and textual features are computed once per article; social features use the sharing events observed up to T_A, and textual features depend only on the article text.

B.3 Evaluation Protocol

Warm-up period. Before training, we use the first five daily snapshots of each corpus as a warm-up period, so that the cumulative clustering and trajectory-reconstruction pipeline has sufficient prior context.

[7] https://github.com/ranahaani/GNews
[8] https://developer.x.com/en/docs

Family	Feature	Definition / interpretation
Geometric: topic distance	d1_nearest_centroid_pct	Within-model percentile rank of the Euclidean distance from the article embedding to the nearest non-noise topic centroid in the snapshot at T_A.
Geometric: topic distance	d2_second_centroid_pct	Within-model percentile rank of the Euclidean distance to the second-nearest non-noise topic centroid at T_A.
Geometric: topic distance	margin_d2_minus_d1_pct	Within-model percentile rank of the margin between the second-nearest and nearest centroid distances at T_A.
Geometric: topic shape	mahal_nearest_pct	Within-model percentile rank of the minimum diagonal Mahalanobis distance to an existing topic cluster at T_A.
Geometric: local density	knn_mean_k20_pct	Within-model percentile rank of the mean Euclidean distance to up to 20 nearest neighbors in the snapshot at T_A, fewer when the snapshot is smaller.
Geometric: local density	knn_std_k20_pct	Within-model percentile rank of the standard deviation of distances to these same neighbors at T_A.
Geometric: outlier neighborhood	outlier_proto_mean_dist_pct	Within-model percentile rank of the mean distance to up to ten nearest articles classified as outliers in the snapshot at T_A.
Geometric: outlierness	outlier_score	HDBSCAN outlierness score at T_A, from the GLOSH score (Campello et al., 2015; McInnes et al., 2017).
Geometric: outlier pool	has_recent_outliers	Indicator that at least one article is classified as an outlier in the snapshot at T_A.
Geometric: outlier pool	n_recent_outliers	Number of articles classified as outliers in the snapshot at T_A.
Social diffusion	soc_unique_users	Number of distinct X users who shared the article URL up to T_A.
Social diffusion	soc_median_user_public_metrics_followers_count	Median follower count of the X users who shared the article URL up to T_A.
Social diffusion	soc_median_user_public_metrics_tweet_count	Median lifetime tweet count of the X users who shared the article URL up to T_A.
Social diffusion	soc_median_user_public_metrics_listed_count	Median listed count of the X users who shared the article URL up to T_A.
Media co-sharing graph	media_weighted_clustering	Weighted clustering coefficient (Onnela et al., 2005) of the article URL node in the co-sharing graph built from shares observed up to T_A.
Media co-sharing graph	media_bridge_ratio	Share of the article node's co-sharing weight that connects outside its Louvain community (Blondel et al., 2008) in the co-sharing graph at T_A.
Media co-sharing graph	media_community_size	Size of the Louvain community (Blondel et al., 2008) containing the article URL node in the co-sharing graph at T_A.
Text style	text_subjectivity	Subjectivity of the article text, from the French version of TextBlob (Loria, 2018).
Text style	text_neutrality	Neutrality of the article text, from the French VADER compound polarity magnitude (Hutto and Gilbert, 2014).
Readability / length	avg_sentence_len_words	Average sentence length in words.
Readability / length	avg_word_len_chars	Average word length in characters.
Readability / length	total_syllables	Approximate total number of syllables in the article text, counted with a rule-based French heuristic.
Readability / length	avg_syllables_per_word	Approximate average number of syllables per word.
Readability / length	len_chars	Total number of characters in the article text.
Readability / length	len_words	Total number of word tokens in the article text.
Named entities	ner_total_ents	Total number of named entities detected with the French spaCy pipeline (Honnibal et al., 2020).
Named entities	ner_distinct_ents	Number of distinct named-entity strings detected with the same pipeline (Honnibal et al., 2020).
Named entities	ner_person	Count of person entities detected with the same pipeline (Honnibal et al., 2020).
Named entities	ner_org	Count of organization entities detected with the same pipeline (Honnibal et al., 2020).
Named entities	ner_loc	Count of location entities detected with the same pipeline (Honnibal et al., 2020).
Named entities	ner_misc	Count of named entities not assigned to person, organization, or location categories (Honnibal et al., 2020).

Table 9: Glossary of predictors used in the supervised models.

Cross-validation. Performance is estimated with 5-fold article-level cross-validation with approximately 80/20 train–test split in each fold, with four folds used for training and the remaining fold held out for evaluation. Folds are constructed at the article level, so that all records associated with the same article are assigned to the same fold and no article can appear in both the training and test sets.

Preprocessing. All preprocessing is fit within each training fold. Missing values are median-imputed using the training split only. Logistic regression and Linear SVC additionally use standardization fit on the training split. Tree-based models are trained on imputed but unscaled features.

Class imbalance and baseline. For logistic regression, Linear SVC, decision tree, and random forest, we use `class_weight=balanced`. For XGBoost, `scale_pos_weight` is set once per corpus and consensus setting to $N_- / \max(N_+, 1)$, where N_+ and N_- are the positive and negative counts in the selected labeled subset for that setting.

SHAP Analysis. For the XGBoost interpretability analysis, we compute probability-scale SHAP values with TreeExplainer (Lundberg et al., 2020).

B.4 Classifier Hyperparameters

Table 10 details the fixed parameters of all classifiers used in the experiments.

Model	Hyperparameters
Random Forest	`n_estimators=600`, `max_depth=8`, `min_samples_split=20`, `min_samples_leaf=10`, `max_features=sqrt`, `class_weight=balanced`
LogReg ℓ_2	`penalty=l2`, `solver=liblinear`, `max_iter=2000`, `class_weight=balanced`
Linear SVC	`max_iter=5000`, `class_weight=balanced`
Decision Tree	`max_depth=6`, `min_samples_leaf=10`, `class_weight=balanced`
XGBoost	`n_estimators=300`, `max_depth=4`, `learning_rate=0.05`, `subsample=0.8`, `colsample_bytree=0.8`, `objective=binary:logistic`, `eval_metric=logloss`, `scale_pos_weight=`$N_- / \max(N_+, 1)$

Table 10: Classifier hyperparameters used in the experiments. We use `random_state=42` where applicable. For XGBoost, N_+ and N_- denote the positive and negative counts in the selected labeled subset for the corresponding corpus–consensus setting.

C Supplementary Classification Results

C.1 Full Results at Selected k

Table 11 reports precision, recall, and F_1-score with standard deviations across folds for the selected consensus settings, complementing Table 3.

C.2 Classifier Results Across Agreement Thresholds

Table 12 reports classifier F_1-scores across different values of k at T_A. These results complement the main XGBoost threshold analysis in Section 6.1, by showing that the coverage–confidence pattern is stable across classifier families.

C.3 Fold-level Ablation Comparisons

Table 13 reports paired fold-level F_1 comparisons for the main ablation contrasts in Table 5. Tests are computed on the same five article-level cross-validation folds used to produce the ablation results. We report the mean paired difference ΔF_1, the paired t-test p-value, and the Benjamini–Hochberg corrected q-value (Benjamini and Hochberg, 1995).

D Robustness Check: Extended Hydrogen Corpus

We repeated the experiment on an extended hydrogen corpus spanning 1 January 2025 to 8 June 2025. This analysis is not part of the main evaluation because social features were not uniformly available over the full period. We use only geometric and textual features and interpret the results as supporting evidence under a longer observation window. Table 14 shows the same pattern observed in the main experiments: stricter k retain fewer articles but yield stronger predictive performance.

E Forward-Chaining Evaluation

This appendix details the chronological evaluation summarized in Section 6.3. At each cutoff t, the model is trained on all articles with $T_A \leq t$ and tested on those with $T_A \in (t, t + W]$; the cutoff then advances by W, so each evaluated article is predicted exactly once, by a model trained only on strictly earlier articles. Origins are restricted to $t \leq T_A^{\max} - W$, so the final partial window is not evaluated. The warm-up before the first origin follows an a-priori minimum-training-size rule, two weeks for the main corpora and four weeks for the extended corpus of Appendix D. The decision threshold is fixed at 0.5 and XGBoost follows Appendix B.4, with `scale_pos_weight` recomputed on each training set and median imputation fit on training articles only. Articles in the warm-up and in the final partial window are not evaluated, which is why the counts in Table 15 are smaller than in Table 2. Brackets give 95% percentile bootstrap intervals on pooled F_1 (2,000 article resamples). Results come from a single run in one pinned environment.

All three corpora exceed the constant-positive baseline even at the lower confidence bound. The extended corpus reaches the highest scores; it also covers a longer period and contains a higher proportion of positives. Under the window-fit rule, the extended corpus pools 129 positives and 290 negatives at $W = 14$d; with an 8-week warm-up, $F_1 = 0.851$ [0.79, 0.91]. Scores are stable across window lengths on the extended corpus and more variable on the two smaller ones, where each window contains few positives. Per-window tables, the full warm-up $\times$ window grid, and per-subclass predicted-positive rates are available upon request.

	HydroNewsFr			ClimateNewsFr		
Model	Precision	Recall	F_1	Precision	Recall	F_1
Baseline	0.246±0.011	1.000±0.000	0.395±0.014	0.171±0.017	1.000±0.000	0.291±0.025
Linear SVC	0.853±0.095	0.941±0.064	0.889±0.056	0.889±0.083	0.930±0.063	0.906±0.051
LogReg ℓ_2	0.838±0.073	0.941±0.064	0.882±0.040	0.895±0.069	0.985±0.031	0.936±0.044
Decision Tree	0.803±0.083	0.891±0.115	0.835±0.051	0.867±0.085	0.937±0.058	0.898±0.058
Random Forest	0.867±0.103	0.917±0.080	0.885±0.063	0.905±0.079	0.969±0.038	0.935±0.057
XGBoost	0.919±0.112	0.917±0.080	0.912±0.069	0.970±0.037	0.969±0.062	0.969±0.047

Table 11: Full article-level cross-validated performance at T_A, reported as mean ± standard deviation across folds. Results use $k = 4$ for HydroNewsFr and $k = 6$ for ClimateNewsFr.

Corpus	k	Retained	Positive	Negative	Baseline	Linear SVC	LogReg	Tree	Forest	XGBoost
HydroNewsFr	1	1235	544	691	0.611	0.765	0.772	0.712	0.761	0.765
	2	773	276	497	0.525	0.815	0.813	0.743	0.811	0.811
	3	508	150	358	0.455	0.851	0.841	0.820	0.854	0.895
	4	337	83	254	0.395	0.889	0.882	0.835	0.885	0.912
	5	219	39	180	0.298	0.863	0.864	0.764	0.843	0.860
	6	145	21	124	0.248	0.878	0.838	0.894	0.937	0.910
	7	81	13	68	0.330	0.950	0.950	0.867	0.833	0.917
	8	40	6	34	0.258	1.000	1.000	0.700	0.900	0.800
ClimateNewsFr	1	1941	875	1066	0.621	0.779	0.778	0.755	0.780	0.778
	2	1390	489	901	0.519	0.827	0.827	0.769	0.823	0.851
	3	1005	273	732	0.427	0.873	0.887	0.843	0.875	0.895
	4	766	182	584	0.384	0.905	0.903	0.860	0.910	0.915
	5	570	115	455	0.336	0.913	0.913	0.927	0.931	0.952
	6	398	68	330	0.291	0.906	0.936	0.898	0.935	0.969
	7	270	43	227	0.273	0.920	0.915	0.903	0.929	0.940
	8	161	21	140	0.228	0.929	0.927	0.962	0.985	0.949

Table 12: Classifier results at T_A across symmetric consensus thresholds $(k, k, 0)$. Values are cross-validated mean F_1 scores.

Corpus	Comparison	ΔF_1	p	q
HydroNewsFr	All vs. all w/o geometry	$+0.526$	7.19×10^{-4}	1.83×10^{-3}
	All vs. only geometry	$+0.011$	0.536	0.577
	All vs. all w/o social	-0.006	0.374	0.436
	All vs. all w/o text	$+0.006$	0.374	0.436
	Only geometry vs. only text	$+0.628$	3.66×10^{-4}	1.17×10^{-3}
	Only geometry vs. only social	$+0.551$	1.39×10^{-4}	5.74×10^{-4}
ClimateNewsFr	All vs. all w/o geometry	$+0.743$	1.57×10^{-4}	4.06×10^{-4}
	All vs. only geometry	0.000	–	–
	All vs. all w/o social	0.000	–	–
	All vs. all w/o text	$+0.007$	0.374	0.374
	Only geometry vs. only text	$+0.767$	7.74×10^{-5}	3.89×10^{-4}
	Only geometry vs. only social	$+0.886$	3.63×10^{-4}	6.95×10^{-4}

Table 13: Paired fold-level F_1 comparisons for the selected ablation settings. Positive ΔF_1 means that the first condition outperforms the second. Dashes indicate identical fold-level F_1 values or degenerate zero differences, for which the paired t-test is not defined.

k	Retained	Positive	Negative	F_1	Recall
1	2134	1056	1078	0.793	0.773
2	1390	605	785	0.871	0.865
3	965	387	578	0.922	0.931
4	656	240	416	0.923	0.910
5	465	163	302	0.942	0.939
6	321	107	214	0.981	0.981
7	204	71	133	0.981	0.988
8	120	48	72	0.991	0.983

Table 14: XGBoost results on the extended corpus.

Corpus	Pos.	Neg.	$W = 5$d	$W = 7$d	$W = 14$d	$W = 21$d	Base
HydroNewsFr	42	211	0.776 [0.67, 0.86]	0.762 [0.65, 0.86]	0.780 [0.68, 0.87]	0.696 [0.60, 0.79]	0.285
ClimateNewsFr	26	316	0.821 [0.70, 0.92]	0.800 [0.66, 0.91]	0.737 [0.59, 0.85]	0.778 [0.64, 0.89]	0.141
HydroNewsFr Ext.	131	345	0.875 [0.83, 0.92]	0.885 [0.84, 0.92]	0.874 [0.83, 0.91]	0.873 [0.83, 0.91]	0.432

Table 15: Forward-chaining results at T_A, for test windows of W days. Ext. denotes the extended hydrogen corpus of Appendix D. Cells report pooled F_1 with a 95% percentile bootstrap interval; Base is the constant-positive baseline. Counts are the pooled evaluated articles at $W = 7$d and vary slightly with W under the window-fit rule.