From Outliers to Topics in Language Models: Anticipating Trends in News Corpora

Evangelia Zve, Benjamin Icard, Alice Breton, Lila Sainero, Gauvain Bourgne, and Jean-Gabriel Ganascia

LIP6, Sorbonne University, CNRS, France

Abstract

This paper examines how outliers, often dismissed as noise in topic modeling, can act as weak signals of emerging topics in dynamic news corpora. Using vector embeddings from state-of-the-art language models and a cumulative clustering approach, we track their evolution over time in French and English news datasets focused on corporate social responsibility and climate change. The results reveal a consistent pattern: outliers tend to evolve into coherent topics over time across both models and languages.

1 Introduction

As information ecosystems become increasingly dynamic, the early identification of emerging trends in news media remains a key challenge for natural language processing. Topic modeling, which clusters semantically similar documents to uncover latent themes, plays a central role in this task. Early approaches, most notably Latent Dirichlet Allocation (`LDA`) (Blei et al., 2003), introduced a probabilistic framework to infer latent topics from textual documents (Hoyle et al., 2022). More recent embedding-based methods, such as `BERTopic` (Grootendorst, 2022), represent documents as dense vector embeddings, enabling more contextualized representations and yielding more coherent topics in dynamic corpora such as online news content (Babalola et al., 2024).

Unlike partition-based clustering methods often used for clustering vector embeddings, such as `KMeans` (Hartigan and Wong, 1979), or probabilistic topic models like `LDA`, both of which assign every document to a topic, `HDBSCAN` (Campello et al., 2015) is a density-based clustering algorithm that explicitly labels low-density points as *outliers*. These documents, which do not fit into any topical cluster, are often treated as noise and excluded from downstream analysis.

Challenging the assumption that outliers are mere noise, we explore the hypothesis that outliers, documents not assigned to any cluster, may serve as early signals of emerging topics. We employ a cumulative clustering approach using `BERTopic` with `HDBSCAN`, tracing how isolated documents evolve and whether they are gradually integrated into clusters as their narratives gain salience. To aid interpretability, we also analyze lexical and stylistic features of outliers and their role in cluster integration.

To conduct our analysis, we use two news corpora. The first, in French, is a manually curated dataset documenting a corporate social responsibility dispute which serves as a pilot study. The second, in English, focuses on climate change and is used for replication. Both corpora are topically constrained, span continuous time periods, and provide full-text coverage, allowing to control for topical and timeline gaps.

Section 2 reviews related work. Section 3 details the full experimental setting, with a particular focus on the methodology. Section 4 presents the French study and results on outlier conversion. Section 5 reports replication results in English. Findings in both languages are discussed and compared in Section 6. Section 7 concludes and outlines future directions.

2 Related Work

Topic modeling is widely applied across various domains, including corporate social responsibility (Lee et al., 2023) and climate change (Ylä-Anttila et al., 2022), in both traditional and social media contexts (Laureate et al., 2023). The field's methodological evolution, from probabilistic approaches like `LDA` (Blei et al., 2003) to embedding-based methods such as `BERTopic` (Grootendorst, 2022), has improved semantic coherence. However, while outliers have been often treated as noise (Alattar and Shaalan, 2021), their role in sig-

naling emerging topics remains an underexplored area of research.

Research in temporal topic analysis has evolved from early techniques like burst detection (Chen et al., 2016) and term-frequency-based change point identification (Yao et al., 2021) to more recent approaches tracking semantic drift (Jung et al., 2020) and transformer-based dynamic modeling (Karakkaparambil et al., 2024; Boutaleb et al., 2024). While these methods effectively capture shifts in established topics, they typically overlook sparse outliers, documents that may precede and predict emerging themes before they coalesce into detectable clusters.

This relates to clustering methodology. While probabilistic topic models like `LDA` assign soft cluster memberships, and partition-based algorithms such as `KMeans` (Hartigan and Wong, 1979) enforce hard assignments, both approaches assume that every document belongs to a cluster. In contrast, density-based methods like `HDBSCAN` (Campello et al., 2015) and `OPTICS` (Ankerst et al., 1999) explicitly identify outliers as low-density points that do not belong to any cluster. Unlike general anomaly detection techniques (e.g., Isolation Forest (Liu et al., 2008), Local Outlier Factor (Breunig et al., 2000)), which detect outliers without considering the topical coherence of thematically structured corpora, `HDBSCAN`'s built-in outlier detection aligns more closely with semantic structure. This allows to track how semantically isolated documents may evolve into coherent topic clusters over time.

This paper examines whether outliers can serve as early signals of emerging topics. By tracking their integration into clusters over time via cumulative clustering, we aim to complement existing work focused on stable topic structures.

3 Experimental Setting

3.1 Hypothesis

While topic modeling and document clustering have been extensively studied, the role of outliers in the dynamic formation of topics has not yet been explored. To address this gap, we propose the following hypothesis:

$\mathcal{H}$: *In topic-based cumulative clustering of news articles, topics emerge or are reinforced in part through the assimilation of outliers—that is, documents initially unclustered that later become part of coherent topic clusters.*

This hypothesis assumes that topic formation in cumulative clustering reflects a gradual process of semantic integration, in which outliers may act as early signals of emerging or evolving topics.

3.2 Models

To test $\mathcal{H}$, we use nine open-source embedding models with diverse transformer architectures and language capabilities. Model selection was guided by performance on the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022), as reported on the Hugging Face leaderboard[1] as of September 16, 2024. Table 3 (Appendix A.2) summarizes the selected models.

3.3 Methodology

The methodology involves four main steps. First, we project news articles into a semantic space using language model embeddings. We then apply dimensionality reduction to enable efficient clustering and address the *curse of dimensionality* (Köppen, 2000). Subsequently, we perform cumulative clustering over 20 monthly time windows and evaluate clustering quality to determine the optimal experimental configuration. Based on this setup, we test $\mathcal{H}$ concerning outlier-to-topic conversion and assess its robustness through inter-model agreement. Finally, we analyze lexical and stylistic features to interpret differences between converted and non-converted outliers.

3.3.1 Data Preparation

Each news article is represented using dense vector embeddings generated from nine pre-trained language models. For each document, we compute embeddings from three variants: body text, headline, and full article (both headline and body text). This projects articles into a high-dimensional semantic space, where distances reflect semantic similarity. We apply Uniform Manifold Approximation and Projection (`UMAP`) (McInnes et al., 2018) to reduce the dimensionality of embeddings prior to clustering. Output dimensions are varied across 2D, 3D, 5D, and 10D. `UMAP` is chosen over Principal Component Analysis (`PCA`) (Wold, 1987) due to its ability to preserve both local and global structure, which is important for identifying fine-grained topic distinctions and local outliers (Atzberger et al., 2023).

[1] https://huggingface.co/spaces/mteb/leaderboard

3.3.2 Cumulative Clustering

We employ cumulative clustering (iterative topic modeling over expanding time windows) across 20 monthly intervals. At each step, documents from the current and all prior months are clustered jointly using `BERTopic` with `HDBSCAN` (McInnes et al., 2017). This density-based algorithm assigns documents to clusters or labels them as outliers via the `GLOSH` algorithm (Campello et al., 2015), which identifies low-density regions by comparing a point's local density to its neighbors. Documents labeled `-1` are classified as outliers and excluded from clusters. To test $\mathcal{H}$, we track whether these outliers transition to inliers (i.e., join a cluster) in subsequent windows, thereby signaling emergent topics.

The clustering quality is evaluated using the silhouette score (Shahapure and Nicholas, 2020), which measures cluster cohesion and separation. Scores above 0.7 are considered strong, 0.5–0.7 moderate, and below 0.25 weak. To evaluate clustering over time, we compute the mean and median silhouette scores across all time windows, and then aggregate these globally across all models. We compare the nine selected embedding models, content variants (headline, body, full article), and `UMAP` settings to ensure robustness. Based on these comparisons, we select the configuration with the highest silhouette score and proceed with our methodology to verify our hypothesis.

3.3.3 Outlier-to-Topic Conversion

Under hypothesis $\mathcal{H}$, we evaluate whether outliers contribute to the formation of new topics or the reinforcement of existing ones. We compute, for each model, the proportion of outliers that later become topic inliers, and assess robustness via the rescaling method of Icard et al. (2024), which measures whether $\mathcal{H}$ is consistently validated for the *same* outliers across models. Specifically, for articles identified as outliers by *all* models (at some point in their time window), we compute the proportion x of models that validate $\mathcal{H}$ and rescale it as $a = |2x - 1|$. This transformation captures consensus independently of polarity (as in Cohen's kappa): both $x = 1$ (unanimous validation) and $x = 0$ (unanimous rejection) yield maximal agreement $a = 1$, while $x = 0.5$ corresponds to minimal agreement $a = 0$, since models are evenly split in this case.

3.3.4 Lexicon and Writing Style Analysis

As an attempt to explain the conversions observed, we first controlled for potential topical differences between converted and non-converted outliers using word-level `TfidfVectorizer` scores (Qaiser and Ali, 2018), hereafter referred to as TF-IDF. Let w be a word and let $\text{TFIDF}_g(w)$ denote its average TF-IDF score in group $g \in \{\mathcal{H}, \text{not } \mathcal{H}\}$, where $\mathcal{H}$ corresponds to outliers that were integrated into topic clusters ("converted"), and not $\mathcal{H}$ to those that remained isolated ("non-converted"). To capture the differential lexical salience of word w across the two groups, we define the delta TF-IDF as:

$$\Delta\text{TFIDF}(w) = \text{TFIDF}_{\mathcal{H}}(w) - \text{TFIDF}_{\text{not } \mathcal{H}}(w) \quad (1)$$

In addition, we investigated variation beyond lexical content by analyzing stylistic differences between converted and non-converted outliers using the stylometric framework introduced by Terreau et al. (2021), which quantifies eight core stylistic dimensions. These include the relative frequency of *function words* (e.g., prepositions, conjunctions, auxiliaries), *punctuation marks* (e.g., periods, commas), *numbers*, and *named entities* (e.g., persons, organizations) per sentence; distributions of *part-of-speech tags* (e.g., nouns, verbs, adjectives); and averages of *structural features* (e.g., word length, word frequency, syllables per word). The framework also incorporates *lexical complexity metrics* (e.g., Yule's K (Yule, 2014), Shannon entropy (Shannon, 1948)) and *readability indices* (e.g., Flesch-Kincaid Grade Level (Kincaid et al., 1975)).

4 Pilot Study

4.1 French Dataset

We constructed a dataset for the pilot study, referred to as TP, consisting of 102 French news articles that we manually collected and curated. The articles document a controversy involving the major energy company *TotalEnergies* and the prestigious French Grande École *École Polytechnique*, who planned to build a research center on the university's Saclay campus. The project drew both support, citing its contribution to energy research, and criticism, focused on academic independence and environmental impact. The TP dataset covers the full timeline of media coverage, from December 2018 to August 2024, and includes documents from official sources, mainstream media, partisan

outlets, opinion sections, and NGOs. It captures the entire development of the story, without topical or temporal gaps.

4.2 Topic-Based Clustering

We applied topic-based clustering to the TP dataset using the methodology described in subsection 3.3. Figure 1 presents the cumulative clustering output generated by the `Solon-embeddings-large-0.1` model. The figure shows a 2D representation derived from 10D UMAP projections of document embeddings across nine time windows, illustrating topic structure and outlier transitions over time.

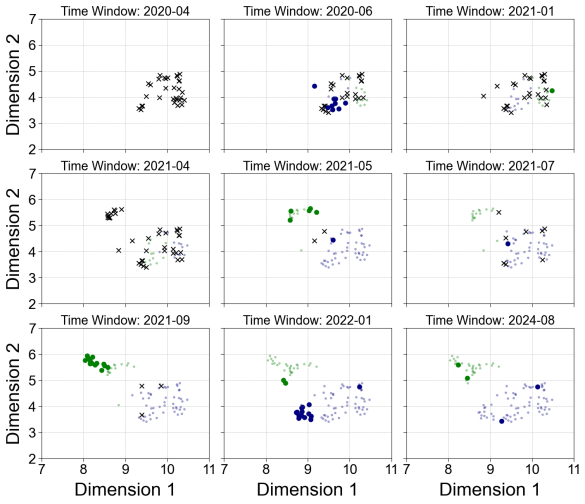


Figure 1: 2D Scatter plot of the cumulative clustering obtained on TP (after UMAP 10D reduction) over nine time windows, using `Solon-embeddings-large-0.1`. Outliers are indicated with black × and topics in blue and green.

Across all nine models and UMAP dimensions, clustering quality is consistent, with mean and median silhouette scores above 0.5 (range: -1 to 1). On average, body-text embeddings yield higher-quality clusters than headline or full-article representations. UMAP with 10 dimensions outperforms the 2D, 3D, and 5D settings. Among models, `Solon-embeddings-large-0.1` achieves the highest scores, while `xlm-roberta-large` performs the worst. Based on these findings, we evaluate Hypotheses $\mathcal{H}$ on TP using UMAP-10D and body-text embeddings.

4.3 Outlier Behavior

To evaluate Hypothesis $\mathcal{H}$, we computed, for each model, the proportion of outliers that later became inliers during cumulative clustering. Figure 2 shows the mean validation score per model.

Table 1: Mean silhouette scores per model for UMAP 10D using the body text of the TP dataset. Bold values indicate the models achieving the best silhouette score for each document type. (See full results in A.3.1.)

Model	UMAP 10D		
	Headline	Body	Full Article
`multilingual-e5-large`	**0.6065**	0.5519	0.5689
`e5-base-v2`	0.5592	0.5350	0.4846
`sentence-camembert-base`	0.5990	0.5850	0.6167
`all-MiniLM-L12-v2`	0.5654	0.5846	0.5349
`Solon-embeddings-large-0.1`	0.5772	0.6694	0.5553
`xlm-roberta-large`	0.4941	0.4802	0.4424
`all-roberta-large-v1`	0.5525	0.6258	0.5759
`multilingual-mpnet-base-v2`	0.5391	0.5923	0.6865
`distilbert-base-uncased`	0.3670	**0.9373**	**0.8895**
Mean	0.5400	**0.6180**	0.5993
Median	0.5417	**0.6183**	0.5756

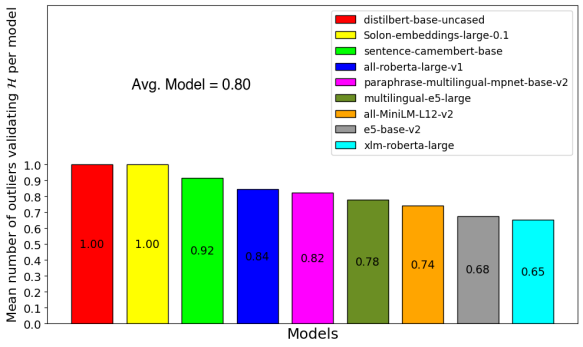


Figure 2: Mean number of outliers per model that validate prediction $\mathcal{H}$ on TP by converting into topic inlier at some time point (specific to each model). Each colored bar represents the mean of each model.

The average validation score of $\mathcal{H}$ across models on TP is high, with a mean of 0.80. As expected, models trained or fine-tuned on French perform strongly: `Solon-embeddings-large-0.1` achieves perfect validation (1.0), and `sentence-camembert-base` scores 0.92. Among English-language models, `e5-base-v2` shows intermediate performance (0.68), while several others yield unexpectedly strong results: `all-MiniLM-L12-v2` (0.74), `all-roberta-large-v1` (0.84), and `distilbert-base-uncased` (1.0). Multilingual models show mixed performance: `xlm-roberta-large` scores moderately (0.65), whereas `paraphrase-multilingual-mpnet-base-v2` (0.82) and `multilingual-e5-large` (0.78) achieve high scores. Overall, model-level validation of $\mathcal{H}$ ranges from moderate to perfect, with a relatively uniform distribution.

Across models, outlier-to-inlier conversion rates are highest in early clustering phases (64.58%–100% in late 2020), followed by a ta-

pering trend with persistent outliers in later periods. As detailed in Appendix A.3.2, some models exhibit stable integration while others decline over time. Despite these intra-model fluctuations, the general pattern of early integration supports $\mathcal{H}$ across temporal windows.

To test whether this holds beyond model variation, we computed inter-agreement using the rescaling method mentioned in Section 3.3.3. The result $a = 0.7002$ shows strong agreement that $\mathcal{H}$ is validated across all models based on converting the same outliers. This suggests that despite inconsistencies in how individual models integrate outliers over time, their validation of $\mathcal{H}$ remains broadly aligned. Accordingly, the average model $x = 0.80$ is a good consensus model regarding the validation of $\mathcal{H}$. For this reason, we proceed using the average model representation for our interpretability analysis.

In the next section, we examine whether writing style, beyond semantic similarity, helps explain why some outliers are eventually integrated into clusters, while others remain isolated.

4.4 Lexicon and Writing Style Analysis

As an attempt to explain the conversion of outliers into topics, we controlled for topical alignment among outliers to assess their influence on topic formation (see Subsection 3.3.4). For each word appearing in outlier documents, we computed the difference in average TF-IDF scores between those validating $\mathcal{H}$ and those not validating $\mathcal{H}$. Specifically, we used the lexical salience metric $\Delta\text{TFIDF}(w)$, as defined in (1), and its inverse. Among the top 20 words in each class, the mean difference was 0.0088 for outliers validating $\mathcal{H}$ and -0.0126 for those not validating $\mathcal{H}$. Both differences were statistically significant at the 0.05 level using the Kruskal–Wallis test.

A closer examination of the top 20 terms most prevalent among outliers validating $\mathcal{H}$ reveals words associated with institutional support for the project (e.g., "*cabinet*", "*total*", "*lobbying*", "*saclay*") or individuals endorsing it (e.g., "*brunelle*", "*nathalie*"). In contrast, the top 20 terms more prevalent among outliers *not* validating $\mathcal{H}$ include words reflecting opposition to the project (e.g., "*recours*", "*victoire*"), as well as references to activist groups (e.g., "*greenpeace*", "*militant*") and opposing figures (e.g., "*julliard*", "*jean*"). In both sets, the majority of these words were statistically distinctive.[2] These results suggest that conversion of outliers into topics is partly influenced by their alignment with dominant themes in the TP dataset, which is consistent with the role of semantic similarity in reinforcing or generating topical structure.

To evaluate our qualitative observation that the lexicon of outliers not validating $\mathcal{H}$ tends to be more subjectively framed or opinion-laden, we carried out a quantitative analysis to test this hypothesis. Specifically, we assessed whether lexical salience defined in (1) correlated with the degree of subjectivity or neutrality in the documents where each word occurred. For each word w, we computed the average subjectivity and neutrality scores across all documents D_w in which it appeared:

$$\text{Subjectivity}(w) = \frac{1}{|D_w|} \sum_{d \in D_w} s(d) \tag{2}$$

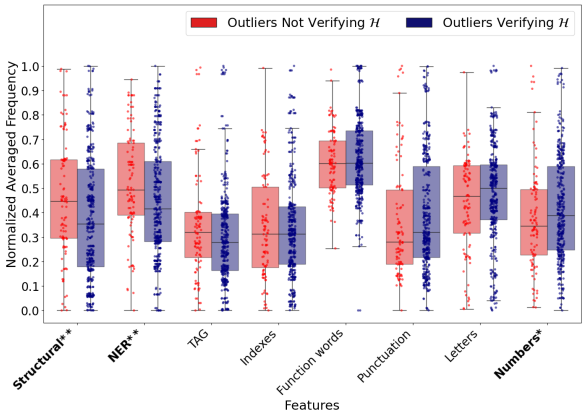
$$\text{Neutrality}(w) = \frac{1}{|D_w|} \sum_{d \in D_w} n(d) \tag{3}$$

where $s(d)$ and $n(d)$ denote the subjectivity and neutrality scores of document d, computed using `TextBlob` (Loria et al., 2018) and `VADER` (Hutto and Gilbert, 2014), respectively.

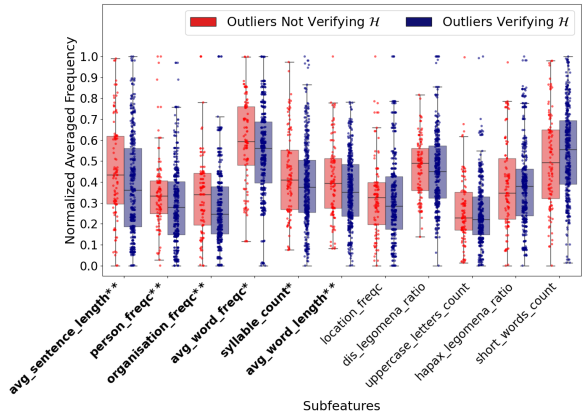
We then computed Spearman's correlation between $\Delta\text{TFIDF}(w)$ and the subjectivity and neutrality scores of the corresponding documents. The analysis revealed a *moderate negative correlation* with subjectivity ($r = -0.223$, $p < 0.01$) and a *weak positive correlation* with neutrality ($r = 0.105$, $p < 0.01$). These patterns indicate that words more prominent among converted outliers tend to appear in more neutral, less subjective contexts and thus that outliers more likely to become topics are characterized by a lexicon that is more factual in nature.

To evaluate broader stylistic effects, we applied the stylometric framework of Terreau et al. (2021) to measure differences across eight core stylistic features between converted and non-converted outliers: *function words*, *punctuation marks*, *numbers*, *named entities*, *part-of-speech tags*, *structural features*, *lexical complexity indexes*, and *readability metrics*. Figure 3 summarizes the results for both main categories (Fig. 3a) and significant subfeatures (Fig. 3b). We omit a detailed analysis for

[2] Three words among outliers validating $\mathcal{H}$ —"*public*", "*direction*", and "*palaiseau*"—were not statistically significant, while only one word ("*ecole*") lacked significance among outliers not validating $\mathcal{H}$.



(a) Differences in mean frequencies for the eight main features.



(b) Differences in mean frequencies for subfeatures, based on observed significance in Figure 3a.

Figure 3: Differences for TP in the eight stylistic features and subfeatures from Terreau et al. (2021), between outliers validating $\mathcal{H}$ and outliers not validating $\mathcal{H}$. Statistical significance is measured using the Kruskal–Wallis test, with * and ** indicating p values < 0.05 and < 0.01, respectively.

significant differences in Numbers, as this feature consists solely of single-digit values (ranging from 0 to 9), making any further breakdown not directly interpretable.

At the level of the eight main features (see Figure 3a), statistically significant differences were observed only for Named Entities (NER), Structural features, and Numbers. Structural features and NER were less frequent in outliers validating $\mathcal{H}$ than in those not validating $\mathcal{H}$, whereas Numbers were more frequent in outliers validating $\mathcal{H}$. No significant differences were found for TAG, Punctuation, Letters, Indexes, or Function Words.

A closer examination of the significant subfeatures (Fig. 3b) shows that, for NER, names of persons and organizations appear significantly less often in outliers validating $\mathcal{H}$. No difference was found for location markers. For Structural subfeatures, outliers validating $\mathcal{H}$ exhibit shorter sentences and words, fewer syllables per word, and higher average word frequency. No other structural subfeatures showed significant variation.

These stylistic differences observed for the average model may be explained by the fact that more structural features introduce complexity, and thus stylistic simplification may support the integration of outliers into topic clusters. Specifically, shorter and simpler text, with fewer named entities, may make it easier for the average model to associate such outliers with broader topic structures, thus facilitating the validation of $\mathcal{H}$. Conversely, a higher frequency of Numbers, particularly single-digit ones, may reflect more patterned or categorical language that also facilitates topic clustering. No clear effects were found for TAG, Punctuation, Letters, Indexes, or the remaining structural subfeatures.

5 Replication Study

5.1 English Dataset

To validate and generalize our findings, we used an existing larger English dataset of climate change news articles, *climate-news-db*.[3] This dataset originally comprised 27,877 news articles from global media outlets, spanning January 2015 to November 2024. To ensure topical consistency, we curated a focused subset of 312 articles, referred to as GHG, by filtering for content explicitly addressing Greenhouse Gas Emissions (GHG). Articles were selected based on the presence of the terms "*Greenhouse Gas*" or "*Greenhouse Emissions*", and sampled across 20 monthly time windows between January 2022 and August 2023. For consistency, we retained only articles from major U.S.-based outlets (e.g., *The Washington Post*, *The New York Times*, *Fox News*, and *CNN*).

5.2 Topic-Based Clustering

We applied topic-based clustering to the body text of the GHG articles using 10D UMAP projections. With the exception of `e5-base-v2`, Table 2 shows that all nine models achieved strong silhouette scores, with both mean and median values at or above 0.5 (on a scale from –1 to 1). These results are slightly lower than, but broadly consistent with, those obtained for the TP dataset under the same configuration.

[3] https://www.climate-news-db.com

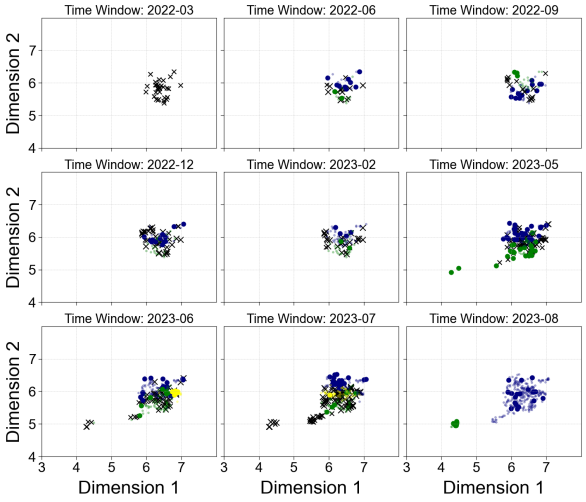


Figure 4: 2D Scatter plot of the UMAP 10D cumulative clustering obtained on GHG over nine time windows, using e5-base-v2. Outliers are indicated with black ×, topics in blue, green and yellow.

Table 2: UMAP 10D silhouette scores obtained on the GHG dataset for the body text of articles, sorted from best to worst.

Model	Mean Silhouette Score
e5-base-v2	**0.5661**
multilingual-e5-large	0.5490
all-MiniLM-L12-v2	0.5416
...-multi..-mpnet-base-v2	0.5387
xlm-roberta-large	0.5376
Solon-embeddings-large-0.1	0.5159
sentence-camembert-base	0.5092
all-roberta-large-v1	0.5044
distilbert-base-uncased	0.4998
Mean	0.5291
Median	0.5376

5.3 Outlier Behavior

Figure 5 shows the mean validation score per model for $\mathcal{H}$ on GHG. The results indicate a high average validation across models, with a mean score of 0.81. As expected, English-specialized models: distilbert-base-uncased, e5-base-v2, and all-MiniLM-L12-v2, achieve perfect validation (1.0), followed by all-roberta-large-v1 (0.85). Among French-specialized models, sentence-camembert-base performs more weakly (0.58), as anticipated, while the perfect score of Solon-embeddings-large-0.1 (1.0) is less expected. Multilingual models show mixed results: paraphrase-multilingual-mpnet-base-v2 and xlm-roberta-large perform poorly (both 0.41), while multilingual-e5-large again achieves perfect validation. The distribution of scores appears bimodal: five models achieve perfect validation, while the remaining four show moderate to low scores. This sharp divide may reflect potential overfitting among English-specialized models that integrate all outliers into topics.

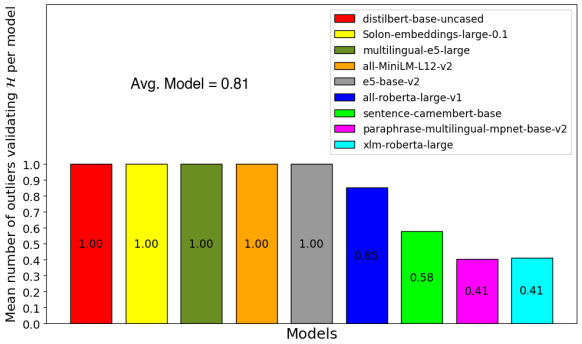


Figure 5: Mean number of outliers per model that validate prediction $\mathcal{H}$ on GHG by converting into topic inlier at some time point (specific to each model). Each colored bar represents the mean for each model.

This consistency in temporal dynamics (see Appendix A.4.1 for a detailed time-window analysis) aligns with the high average validation score of 0.81 (Figure 5). Most models follow a similar pattern: strong early outlier-to-topic conversion, reduced integration in mid-phases, and stabilization with persistent outliers. While some models, particularly multilingual ones and sentence-camembert-base, show greater fluctuation, the overall trend supports $\mathcal{H}$. As in the Pilot experiment, we computed inter-agreement across models with respect to $\mathcal{H}$, using the rescaling method of Icard et al. (2024). Again, the result $a = 0.6783$ strongly supports that models validate $\mathcal{H}$ based on converting the exact same outliers. The average model $x = 0.81$ is then a good consensus model regarding the validation of $\mathcal{H}$.

5.4 Lexicon and Writing Style Analysis

As part of our interpretability analysis, we sought to understand why some outliers aligned with topics while others did not. We first examined the top 20 words with the highest ΔTFIDF scores in outliers validating $\mathcal{H}$ compared to those not validating it, and vice versa. As defined in (1), $\Delta\text{TFIDF}(w)$ captures the difference in average TFIDF scores for word w between the two outlier classes. The mean difference was 0.0031 for (1), and 0.0023 for the reverse. Neither difference was statistically significant (Kruskal–Wallis test, $p > 0.05$), suggesting that thematic lexical content does not meaningfully distinguish the two outlier classes in GHG. How-

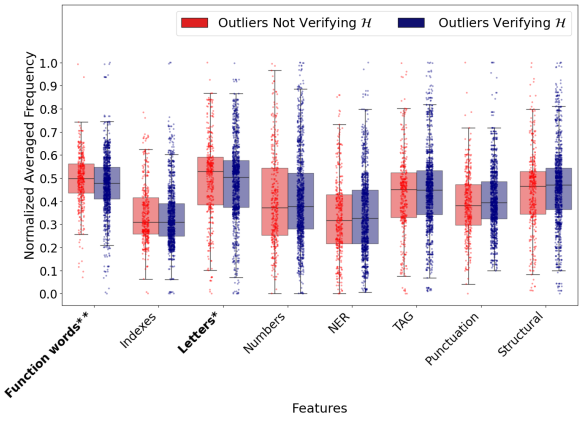


Figure 6: Differences for GHG in the eight stylistic features from Terreau et al. (2021), between outliers validating $\mathcal{H}$ and outliers not validating $\mathcal{H}$. Statistical significance is measured using the Kruskal–Wallis test, with * and ** indicating p values < 0.05 and < 0.01, respectively.

ever, this finding does not rule out the possibility that stylistic or other non-topical lexical and linguistic features influence outlier conversion.

To address this gap, we analyzed the differences in stylistic characteristics between the two outlier classes, using the framework proposed by Terreau et al. (2021). The results for GHG are given in Figure 6 for the eight main features. We do not provide a detailed analysis of Function words and Letters, as Letters consist solely of single-character values (ranging from A to Z), and Function Words gain significance from their overall distribution rather than their individual occurrences, making a further breakdown not directly interpretable.

Among the eight features, significant frequency differences were found only for Function words and Letters, which were notably less frequent in outliers verifying $\mathcal{H}$ compared to outliers not verifying $\mathcal{H}$. This may be explained by the fact that function words (e.g., prepositions, conjunctions) and letters (e.g., A, B, C) lack semantic content, so their reduction helps the average model recognize topics in outliers and validate $\mathcal{H}$. In contrast, Indexes, Numbers, NER, Punctuation, TAG, and Structural features do not appear to have a particular effect on this recognition.

6 Discussion

We observed consistent outlier-to-topic conversion across two linguistically distinct datasets, confirming that the phenomenon generalizes. Validation of $\mathcal{H}$ is robust across topic domains (social responsibility and climate change), languages (French and English), and dataset sizes (102 and 312 articles), with a stable mean score around 0.80. Inter-model agreement remains high (with $a = 0.7002$ for French, $a = 0.6783$ for English), suggesting that topic-based clustering reliably integrates outliers under varied conditions.

In lexical analysis, TF-IDF differences between converted and non-converted outliers were significant in TP but not in GHG. In TP, converted outliers were more strongly associated with lower subjectivity and higher lexical neutrality. This reflects a structural difference: TP focuses on a defined controversy with a polarized lexicon, while GHG likely follows a more neutral, report-oriented style, as it was not curated under controversy criteria.

The stylistic features analysis revealed that writing style has a significant impact on the conversion of outliers into topics, though the relevant features differ by language. In TP, conversion is influenced by structural features, named entities, and numbers; in GHG, by function words and letter distributions. This suggests that embedding models rely on language-specific stylistic cues when integrating outliers.

These differences align with model training: French-trained models perform better on TP, English-trained ones on GHG, while multilingual models show mixed results, reflecting their training data (see Table 3 in Appendix A.4.1 for details).

7 Conclusion

Our findings demonstrate that outlier-to-inlier conversion is a consistent mechanism in topic emergence within cumulative, density-based clustering frameworks. The effect is robust across nine language models, two typologically distinct languages, and datasets with varying topical scope. In the French dataset (TP), focused on a well-defined controversy, average model validation reached 0.80; in the English dataset (GHG), covering broader climate discourse, the score was similarly high at 0.81. Inter-model agreement exceeded 0.65 in both cases, indicating stable clustering dynamics across architectures and domains.

Future work will distinguish between outliers that act as precursors to new topics and those that reinforce existing structures. We aim to quantify their predictive value and examine their temporal behavior across phases of topic development.

We also plan to scale our analysis to larger and

more heterogeneous corpora, particularly in domains where informational risks, such as discursive conflict and disinformation, are likely to emerge or escalate. In parallel, we will evaluate alternative clustering algorithms with integrated outlier detection (e.g., `OPTICS`) and broaden our assessment across additional model architectures. These extensions aim to test the generality and deepen the explanatory power of our findings.

Limitations

This study was designed as a controlled pilot to explore the predictive role of outliers in topic emergence under well-defined experimental conditions. Although the number of raw articles was relatively limited (102 in French and 312 in English), each document was processed with nine distinct language models, resulting in 918 French and 2,808 English data points. This mitigated the limitations typically associated with small corpus sizes.

High inter-model agreement ($a = 0.7002$ for French and $a = 0.6783$ for English) and consistent clustering quality (silhouette scores of 0.61 and 0.52, respectively) further support that the results are robust within the bounds of this setup.

The decision to prioritize depth over breadth at the expense of dataset size was deliberate: it enabled the construction of a high-quality, manually curated corpus with full-text availability, temporal continuity (i.e., no temporal gaps), and source diversity. This design helped control for confounding factors such as incomplete timelines and uneven topic coverage, which often affect large-scale datasets whose compilation processes are not fully transparent.

While these constraints were necessary to ensure experimental clarity and interpretability, they naturally limit the generalizability of the findings. Future work will scale the analysis to larger corpus of news articles to test its applicability in more complex and dynamic information environments.

Ethics Statement

Our research adheres to the ethical principles of open science, transparency, and sustainability. We ensure reproducibility by making our code accessible in a dedicated GitHub repository, with data and results available upon request. We comply with intellectual property and data protection regulations by sharing only vector embeddings generated by language models. This approach aligns with the principles of 'transformative fair use". We promote AI transparency by contributing to the interpretability of language models, supporting the responsible and explainable use of these models. To support efficiency and sustainability, we prioritize the use of small, open-source language models.

Declaration of contribution

EZ, BI, and JGG conceptualized the research problem and designed the experiments. EZ managed the data collection process. AB and LS managed data cleaning and annotation. EZ was responsible for the technical aspects: coding, model selection, and building the experimental framework. EZ, BI, GB, and JGG analyzed and discussed the results. EZ and BI wrote the paper, which all authors read and revised together. EZ and BI share first authorship. Correspondence: evangelia.zve@lip6.fr, benjamin.icard@lip6.fr, jean-gabriel.ganascia@lip6.fr.

References

Fuad Alattar and Khaled Shaalan. 2021. Emerging research topic detection using filtered-lda. *AI*, 2(4):578–599.

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.

Daniel Atzberger, Tim Cech, Willy Scheibel, Matthias Trapp, R. Richter, J. Dollner, and Tobias Schreck. 2023. Large-scale evaluation of topic models and dimensionality reduction methods for 2d text spatialization. *DOI: 10.1109/TVCG.2023.3326569*. Accessed: January 2025.

Olusola Babalola, Bolanle Ojokoh, and Olutayo Boyinbode. 2024. Comprehensive evaluation of lda, nmf, and bertopic's performance on news headline topic modeling. *Journal of Computing Theories and Applications*, 2(2):268–289.

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. *arXiv preprint arXiv:2411.05930*.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104. ACM.

Ricardo JGB 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 (TKDD)*, 10(1):1–51.

Yubo Chen, Liheng Xu, Kang Liu, and Jun Zhao. 2016. Event detection with burst information networks. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 3317–3327.

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

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.

Amos Hoyle, Pratyusha Goel, Cynthia Phillips, Jordan Boyd-Graber, and Philip Resnik. 2022. Is automated topic model evaluation broken? the incoherence of coherence. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 478–493.

Clayton 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.

Benjamin Icard, François Maine, Morgane Casanova, Géraud Faye, Julien Chanson, Guillaume Gadek, Ghislain Atemezing, François Bancilhon, and Paul Égré. 2024. A multi-label dataset of french fake news: Human and machine insights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 812–818.

Sukhwan Jung, Rituparna Datta, and Aviv Segev. 2020. Identification and prediction of emerging topics in news media. In *IEEE International Conference on Big Data*.

Charu Karakkaparambil, Mayank Nagda, Nooshin Haji Ghassemi, Marius Kloft, and Sophie Fellenz. 2024. Evaluating dynamic topic models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Mario Köppen. 2000. The curse of dimensionality. In *5th online world conference on soft computing in industrial applications (WSC5)*, volume 1, pages 4–8.

Caitlin Doogan Poet Laureate, Wray Buntine, and Henry Linger. 2023. A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, 56(12):14223–14255.

Haein Lee, Seon Hong Lee, Kyeo Re Lee, and Jang Hyun Kim. 2023. Esg discourse analysis through bertopic: comparing news articles and academic papers. *Computers, Materials & Continua*, 75(3):6023–6037.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Steven Loria et al. 2018. textblob documentation. *Release 0.15*, 2(8):269.

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, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Shahzad Qaiser and Ramsha Ali. 2018. Text mining: use of tf-idf to examine the relevance of words to documents. *International journal of computer applications*, 181(1):25–29.

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.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Enzo Terreau, Antoine Gourru, and Julien Velcin. 2021. Writing style author embedding evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 84–93. Association for Computational Linguistics.

H. Wold. 1987. Principal component analysis. *Technometrics*, 38(3):235–238.

Lili Yao, Yue Zhang, and Xisen Wang. 2021. Dynamic term frequency analysis for topic shift detection. *ACL*.

Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. 2022. Topic modeling for frame analysis: A study of media debates on climate change in india and usa. *Global Media and Communication*, 18(1):91–112.

C Udny Yule. 2014. *The statistical study of literary vocabulary*. Cambridge University Press.

A Appendix

A.1 Supplementary Materials

The code and visualizations supporting this paper are available at: https://github.com/evangeliazve/outliers-to-topics-icnlsp. The datasets and experimental results can be provided upon request. The repository includes Python scripts for reproducing our experiments, as well as statistical analyses and visualizations corresponding to key figures and tables in the paper. The BERTopic framework is documented at: https://maartengr.github.io/BERTopic/. Further details on HDBSCAN can be found in its official documentation: https://hdbscan.readthedocs.io/en/latest/, and information on UMAP dimensionality reduction is available at: https://umap-learn.readthedocs.io/en/latest/basic_usage.html. For TF-IDF, we used the TfidfVectorizer from scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html. The sentiment analysis tools employed in this study are TextBlob (https://textblob.readthedocs.io/en/dev/index.html) and VADER (https://github.com/cjhutto/vaderSentiment)..

A.2 Models

Table 3 presents the nine sentence embedding models used in our experiments for topic-based clustering, detailing their underlying architectures, embedding dimensionality, language coverage, and model sizes.

A.3 Pilot Study Appendix

A.3.1 Detailed Silhouette Scores

In this appendix, we provide detailed results from the pilot study evaluating the effectiveness of different sentence embedding models for topic-based clustering on the TP dataset. Specifically, Table 4 reports the mean silhouette scores obtained for each model under varying dimensionality reductions (2D, 3D, 5D, and 10D using UMAP) and different text sample types (headline, body, and combined text). These results offer insights into how model selection, dimensionality, and text granularity impact clustering quality.

A.3.2 Validation or invalidation of $\mathcal{H}$ per model over different time windows for TP

For a more detailed examination of model variations, both across and within models, Table 5 presents the validation or invalidation of $\mathcal{H}$ per model over different periods of cumulative clustering. Among French models, Solon-embeddings-large-0.1 is fully consistent, achieving complete integration early, while sentence-camembert-base shows non-monotony, with conversion dropping from 95.83% to 50% and some persistent outliers. English models exhibit pronounced inconsistency: e5-base-v2 weakens over time (68.75% to 36.67%), all-MiniLM-L12-v2 and all-roberta-large-v1 show fluctuating progress despite strong early conversion (66.67% and 85.42%), and distilbert-base-uncased remains fully stable with no outliers through the whole. Multilingual models vary widely, with paraphrase-multilingual-mpnet-base-v2 and xlm-roberta-large starting strong (85.42%, 64.58%) but leaving substantial outliers later (12.74%, 42.15%), while multilingual-e5-large follows an unstable trajectory, declining from 83.33% to 48.15%.

Across models, a general pattern emerges: strong early conversion of outliers into topic inliers, slowing integration in the mid-phase, and eventual stabilization with persistent outliers in 2023. Early clustering is largely consistent, with conversion rates ranging from 64.58% (xlm-roberta-large) to 100% (Solon-embeddings-large-0.1) in 2020-11. By the mid-phase (2021-07), some models, like all-MiniLM-L12-v2 (73.91%) and all-roberta-large-v1 (71.43%), sustain moderate integration, while others, like e5-base-v2 (42.86%), decline. Late-stage variations are more pronounced, with paraphrase-multilingual-mpnet-base-v2 retaining 46.47% of outliers as topic inliers, while xlm-roberta-large and e5-base-v2 drop to 26.67% and 36.67%, respectively. sentence-camembert-base, despite an early peak (95.83%), declines to 50.00%.

A.3.3 Top 10 Distinguishing Terms Based on TF-IDF Differences Between Outliers Validating and Not Validating $\mathcal{H}$

Table 6 lists the top 10 terms whose TF-IDF scores most strongly differentiate outliers that validate hypothesis $\mathcal{H}$ from those that do not, highlighting

Model	Architecture	Dimensions	Language	Parameters
Solon-embeddings-large-0.1	RoBERTa	1024	French	560M
sentence-camembert-base	CamemBERT	768		111M
all-roberta-large-v1	RoBERTa	1024	English	355M
e5-base-v2	E5	768		109M
distilbert-base-uncased	DistilBERT	768		67M
all-MiniLM-L12-v2	MiniLM	384		33.4M
xlm-roberta-large	XLM-RoBERTa	1024	Multilingual	561M
multilingual-e5-large	E5	1024		560M
paraphrase-multilingual-mpnet-base-v2	MPNet	768		278M

Table 3: Description of the nine sentence embedding models used to conduct the topic-based clustering experiments.

Model	UMAP 2D			UMAP 3D			UMAP 5D			UMAP 10D		
	Headline	Body	All	Headline	Body	All	Headline	Body	All	Headline	Body	All
multilingual-e5-large	0.6235	0.6002	0.5914	0.6121	0.5480	0.5713	0.6020	0.5481	0.5692	0.6065	0.5519	0.5689
e5-base-v2	0.6133	0.5556	0.4668	0.5718	0.5627	0.4671	0.5580	0.5479	0.5030	0.5592	0.5350	0.4846
sentence-camembert-base	0.6120	0.5616	0.5994	0.6083	0.5791	0.6302	0.5934	0.5877	0.6354	0.5990	0.5850	0.6167
all-MiniLM-L12-v2	0.6039	0.5858	0.5465	0.5570	0.6197	0.5243	0.5702	0.4962	0.5608	0.5654	0.5846	0.5349
Solon-embeddings-large-0.1	0.5573	0.6340	0.6497	0.6056	0.6351	0.6031	0.5660	0.6153	0.5778	0.5772	0.6694	0.5553
xlm-roberta-large	0.5416	0.3729	0.3294	0.4996	0.3812	0.3226	0.5348	0.3694	0.3848	0.4941	0.4802	0.4424
all-roberta-large-v1	0.5294	0.6701	0.5862	0.5427	0.6255	0.6121	0.5536	0.6361	0.6040	0.5525	0.6258	0.5759
..-multilingual-mpnet-base-v2	0.5259	0.6062	0.7324	0.5221	0.5918	0.6429	0.5517	0.5977	0.6754	0.5391	0.5923	0.6865
distilbert-base-uncased	0.4872	0.7907	0.8535	0.5232	0.9233	0.8509	0.4413	0.9575	0.8670	0.3670	0.9373	0.8895
Mean	0.5660	0.5975	0.5945	0.5603	0.6074	0.5816	0.5523	0.5951	0.6008	0.5400	0.6180	0.5993
Median	0.5588	0.6056	0.5929	0.5692	0.5984	0.5718	0.5544	0.6029	0.6040	0.5417	0.6183	0.5756

Table 4: Mean silhouette scores per model, dimensionality and text samples types obtained on dataset TP.

key lexical features associated with each group.

A.4 Replication Study Appendix

A.4.1 Validation or invalidation of $\mathcal{H}$ per model over different periods for GHG

For a detailed examination of model variations, Table 7 presents the validation or invalidation of $\mathcal{H}$ for each model over different periods of cumulative clustering. Among English models, distilbert-base-uncased, e5-base-v2, and all-MiniLM-L12-v2 show complete consistency, achieving full integration early. all-roberta-large-v1 follows a steady trajectory, with conversion decreasing slightly from 93.62% to 88.17%. Among French models, sentence-camembert-base, a French model, shows instability, with a conversion fluctuating from 46.43% to 58.14% before dropping to 43.18%. Solon-embeddings-large-0.1, despite being a French model, integrates all outliers early, aligning with its high absolute validation score. Multilingual models exhibit mixed behaviors., with multilingual-e5-large achieving full integration like English models, while paraphrase-multilingual-mpnet-base-v2 and xlm-roberta-large retain substantial outliers (with 40.70% and 43.91%, respectively). multilingual-mpnet-base-v2 initially increases its conversion (26.32% to 38.10%) before stabilizing. xlm-roberta-large exhibits a downward trend, with conversion dropping from 45.00% to 22.00%.

That said, trends across models reveal a broadly consistent trajectory: high early conversion of outliers into topic inliers (ranging from 26.32% to 100% in 2022-10), followed by a mid-phase slowdown with moderate-to-low integration (10.95%–65.71% in 2023-02), and eventual stabilization with persistent outliers in the final stage (10.25%–43.91%). Most models adhere to this pattern, with strong early conversion seen in all-roberta-large-v1 (93.62%) and e5-base-v2 (100%), followed by a gradual decline in mid-phase integration for models like sentence-camembert-base (fluctuating from 46.43% to 58.14%) and multilingual-mpnet-base-v2 (increasing from 26.32% to 38.10%). By the final stage, outlier retention converges to similar rates across models, such as sentence-camembert-base stabilizing at 25.64% and xlm-roberta-large retaining 43.91% of outliers.

A.4.2 Top 10 Distinguishing Terms Based on TF-IDF Differences Between Outliers Validating and Not Validating $\mathcal{H}$

Table 8 lists the top 10 terms whose TF-IDF scores most strongly differentiate outliers that validate hypothesis $\mathcal{H}$ from those that do not, highlighting key lexical features associated with each group.

Model	Measures	Time			
		2020-11 (50%)	**2021-07 (70%)**	**2023-09 (90%)**	**Remaining (100%)**
Solon-..-large-0.1	Nb Outliers / All Articles at t	48/48	8/69	0/100	0/102
	% Becoming Inliers at $(t+n)$	100%	100%	0.00	Converted on 2022-01
..-multi..-mpnet-..	Nb Outliers / All Articles at t	48/48	69/69	15/100	13/102
	% Becoming Inliers at $(t+n)$	85.42%	84.06%	46.47%	-
sentence-camembert-..	Nb Outliers / All Articles at t	48/48	33/69	8/100	4/102
	% Becoming Inliers at $(t+n)$	95.83%	90.91%	50.00%	-
multi..-e5-large	Nb Outliers / All Articles at t	48/48	25/69	27/100	25/102
	% Becoming Inliers at $(t+n)$	83.33%	64.00%	48.15%	-
xlm-roberta-large	Nb Outliers / All Articles at t	48/48	39/69	30/100	43/102
	% Becoming Inliers at $(t+n)$	64.58%	64.10%	26.67%	-
all-MiniLM-L12-v2	Nb Outliers / All Articles at t	48/48	69/69	16/100	26/102
	% Becoming Inliers at $(t+n)$	66.67%	73.91%	12.50%	-
all-roberta-large-v1	Nb Outliers / All Articles at t	48/48	21/69	10/100	12/102
	% Becoming Inliers at $(t+n)$	85.42%	71.43%	30.00%	-
distil..-base-uncased	Nb Outliers / All Articles at t	0/48	0/69	0/100	0/102
	% Becoming Inliers at $(t+n)$	0.00%	0.00%	0.00%	Converted on 2020-06
e5-base-v2	Nb Outliers / All Articles at t	48/48	21/69	30/100	41/102
	% Becoming Inliers at $(t+n)$	68.75%	42.86%	36.67%	-

Table 5: Proportion of outliers converting to clusters in TP, for each model and along four time windows.

Word	ΔTFIDF(w)	ΔOcc(w)	**Word**	ΔTFIDF(w)	ΔOcc(w)
cabinet	0.0122*	93	totalenergies	-0.0328**	-28
total	0.0119*	2613	recours	-0.0185**	-14
brunelle	0.0106*	136	greenpeace	-0.0173**	-265
nathalie	0.0104*	139	victoire	-0.0155**	-6
lobbying	0.0103**	122	ecole	-0.0143	-162
public	0.0098	428	julliard	-0.0129**	-14
direction	0.0097	563	jean	-0.0126**	-20
palaiseau	0.0095	60	décision	-0.0124**	-127
saclay	0.0089*	740	conseil	-0.0116**	-626
quartier	0.0086*	40	militant	-0.0112**	-47

Table 6: Top 10 absolute values of ΔTFIDF(w) for TP. Words with positive values are more characteristic of converted outliers ($\mathcal{H}$), and those with negative values are more typical of non-converted outliers (not $\mathcal{H}$). Statistical significance is based on the Kruskal-Wallis test; * and ** indicate p-values < 0.05 and < 0.01, respectively. ΔOcc(w) indicates the difference in word occurrence counts between the two groups.

Model	Measures	Time			
		2022-10 (50%)	**2023-02 (70%)**	**2023-06 (90%)**	**Remaining (100%)**
	Nb Outliers / All Articles at t	79/79	18/105	96/236	0/312
`Solon-embeddings-large-0.1`	% Becoming Inliers at $(t+n)$	100%	100%	100%	Converted on 2023-07
	Nb Outliers / All Articles at t	19/79	21/105	81/236	127/312
`...-multi..-mpnet-base-v2`	% Becoming Inliers at $(t+n)$	26.32%	38.10%	33.33%	-
	Nb Outliers / All Articles at t	28/79	43/105	88/236	80/312
`sentence-camembert-base`	% Becoming Inliers at $(t+n)$	46.43%	58.14%	43.18%	-
	Nb Outliers / All Articles at t	23/79	26/105	49/236	0/312
`multi..-e5-large`	% Becoming Inliers at $(t+n)$	100%	100%	100%	Converted on 2023-07
	Nb Outliers / All Articles at t	20/79	60/105	76/236	137/312
`xlm-roberta-large`	% Becoming Inliers at $(t+n)$	45.00%	45.00%	22.00%	-
	Nb Outliers / All Articles at t	79/79	69/105	90/236	0/312
`all-MiniLM-L12-v2`	% Becoming Inliers at $(t+n)$	100%	100%	100%	Converted on 2023-07
	Nb Outliers / All Articles at t	47/79	40/105	93/236	32/312
`all-roberta-large-v1`	% Becoming Inliers at $(t+n)$	93.62%	92.50%	88.17%	-
	Nb Outliers / All Articles at t	42/79	33/105	87/236	0/312
`distilbert-base-uncased`	% Becoming Inliers at $(t+n)$	100%	100%	100%	Converted on 2023-07
	Nb Outliers / All Articles at t	13/79	27/105	58/236	0/312
`e5-base-v2`	% Becoming Inliers at $(t+n)$	100%	100%	100%	Converted on 2023-07

Table 7: Proportion of outliers converting to clusters in GHG, for each model and along four time windows.

Word	ΔTFIDF(w)	ΔOcc(w)	**Word**	ΔTFIDF(w)	ΔOcc(w)
climate	0.0067	17851	amazon	-0.0034	-98
report	0.0051*	2656	pakistan	-0.0033	-130
degree	0.0035	2576	china	-0.0031	-480
said	0.0035	9134	child	-0.0027	-227
bill	0.0033*	804	thunberg	-0.0024	-63
company	0.0032	2577	reactor	-0.0023	-65
would	0.0031	3668	protest	-0.0023	-87
republican	0.0030	728	soil	-0.0023	-185
energy	0.0030	5651	granholm	-0.0023	-51
nice	0.0029	895	art	-0.0023	-172

Table 8: Top 10 absolute values of ΔTFIDF(w) for GHG. Words with positive values are more characteristic of converted outliers ($\mathcal{H}$); words with negative values are more typical of non-converted outliers (not $\mathcal{H}$). Statistical significance is based on the Kruskal-Wallis test; * indicates $p < 0.05$. ΔOcc(w) shows the difference in word frequency between the two groups.