Value Creation Radar: Interactive Schema-guided Emerging Topic Detection

Yuri Campbell[1*] **Jan Peuckert**[2] **Michael Prilop**[3] **Sebastian Haugk**[1] **Elena Senger**[1,4]

[1]Fraunhofer ISI, Germany
[2]Helmut Schmidt University, Germany
[3]Institute for Applied Informatics (InfAI), Germany
[4]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
yuri.campbell@isi.fraunhofer.de, jan.peuckert@hsu-hh.de
prilop@infai.de, sebastian.haugk@isi.fraunhofer.de, elena.senger@cis.lmu.de

Abstract

We present an interactive topic modeling system with schema-grounded representations for large, evolving document collections. The system addresses two limitations of prior work on emerging topic detection: detected topics are typically modeled as purely distributional patterns without grounding in a domain-specific schema, and outputs such as topic labels, popularity scores, and summaries are not designed for iterative human interpretation. Our system combines BERTopic-based emerging topic detection with a document-level relevance classifier and a schema-aligned topic summarizer, both anchored in an explicit schema of sectoral value creation. An interactive interface lets analysts inspect, group, split, and label clusters, turning model outputs into curated, contextualized topics. We demonstrate the system on scientific abstracts and news articles, where it supports the identification of emerging, low-frequency semantic patterns relevant to horizon scanning within strategic foresight.

1 Introduction

Detecting emerging, low-frequency semantic patterns in large, temporally evolving document collections is a recurring challenge in NLP (Maitre et al., 2019; El Akrouchi et al., 2021). Such patterns are sparse, ambiguous, and distributed across heterogeneous sources, making them difficult to capture with frequency-based or static topic modeling approaches (Yoon, 2012). They are particularly relevant for strategic foresight—especially in horizon scanning (Ishigaki et al., 2022)—where companies, policymakers, and analysts seek to anticipate changes in sectoral value creation (shifts in actors, processes, business models, or institutional arrangements) before they become widely visible. In this setting, the relevance of an emerging topic depends not only on its temporal dynamics but also on whether it can be interpreted with respect to a specific economic or industrial context.

Consider a group of documents in a corpus of scientific abstracts, described by the keywords *diesel*, *biodiesel*, *fuel*. Summarizing them adds technical detail—nanoparticle additives blended into fuels and lubricants to improve combustion in compression-ignition engines—but still describes a matter of engine performance. Read against a schema of value creation, the same material introduces new *actors* (nanoparticle suppliers) contributing new material-development and testing *resources* to formulation and qualification *processes*, yielding *artefacts* such as nano-additive blends that serve a *demand* shaped by tightening emission standards. What the keywords present as an established fuels research area becomes a visible reconfiguration of a sectoral value network—the kind of signal horizon scanning is meant to surface.

Despite substantial progress in topic modeling (Boutaleb et al., 2024; Zve et al., 2025) and in interactive topic exploration (Smith et al., 2018; Pham et al., 2024; Katz et al., 2024), two limitations remain across both threads (§2) in the context of emerging topic detection for strategic foresight. First, detected topics are modeled as purely distributional patterns without grounding in structured domain representations, leaving the interpretive step illustrated above entirely to the analyst. Second, system outputs are produced as one-shot predictions, with little support for the iterative inspection, refinement, and labeling that domain experts perform in practice.

We present *Value Creation Radar*, an interactive system that addresses both limitations.[1] Our key idea is to treat detected topics as intermediate representations that are further mapped to an explicit schema of sectoral value creation, enabling

* Corresponding author.

[1]The system is available from the corresponding author on request at https://wsrweb.imw.fraunhofer.de/.

context-sensitive interpretation; an interactive interface then lets analysts inspect, refine, and curate the resulting topics (§4).

Contributions.

- A schema of sectoral value creation, operationalized as two NLP components: a distantly supervised document-level relevance classifier and a schema-conditioned topic summarizer (§4.2).
- Their integration with emerging topic detection and interactive cluster refinement in a single human-in-the-loop workflow (§4).
- A human evaluation with domain experts, in which schema-grounded representations are rated higher than plain LLM summaries on usefulness and interpretability, and the relevance classifier matches expert consensus on 75% of documents (§6).

2 Related Work

Recent NLP approaches to emerging topic detection model the problem using topic modeling, representation learning, and temporal signal analysis. Early work relied on probabilistic topic models and lexical statistics (Maitre et al., 2019; El Akrouchi et al., 2021); more recent methods leverage supervised retrieval and comment generation (Ishigaki et al., 2022), embedding-based neural topic models with temporal scoring (Boutaleb et al., 2024; Zve et al., 2025), and AutoML pipelines for trend prediction (Abolfadl et al., 2025). These approaches advance the detection of temporally salient topics, but the resulting representations remain purely distributional and are not anchored to any explicit domain schema.

A parallel line of work has explored interactive topic exploration (e.g. Smith et al., 2018; Pham et al., 2024; Wang et al., 2023; Katz et al., 2024). These systems support human-in-the-loop exploration of topic models, but they are typically built around static corpora and generic topic representations, with no mechanism for emerging topic detection or grounding in a domain schema. Our system sits at the intersection of these two threads: it combines temporal emerging topic detection with schema-grounded interpretation and interactive refinement in a single workflow.

Element	Definition
Actor	an individual or a collective that contributes a resource at its disposal to a process of creating an artefact
Resource	a capability of an actor that is used in the process of creating an artefact
Process	an activity for the creation of an artefact to which actors contribute their resources
Pattern	an organising principle through which a process integrates the resources to create an artefact
Artefact	a good that is created by integrating resources according to a pattern and that serves a demand
Demand	an articulated or anticipated need that an artefact serves
Revenue	a return an actor receives for contributing to the process of creating an artefact
Framework	a context factor that influences how value creation takes place

Table 1: Elements of Value Creation Systems

3 Domain Schema: Value Creation

To move beyond purely distributional representations, we introduce a schema-guided representation layer that links textual signals to a structured schema of value creation. Inspired by frame semantics (Fillmore, 1982; Baker et al., 1998) [2], we model value creation as a domain-specific interpretive schema characterized by a set of abstract schema elements and implicit relations among them (e.g., actors, resources, processes; see Table 1). The schema provides semantic constraints that guide the interpretation of clusters and documents.

4 System

Our system (Figure 1) is designed to support the scanning phase of horizon scanning, a structured foresight process with three phases: scoping, in which domain experts define a search domain and assemble a corpus; scanning, in which the corpus is systematically explored to surface candidate emerging topics; and sense-making, in which ex-

[2]We make use of the general intuition that domains can be structured around participant roles. However, our interpretative schema does not follow all directives of a conceptual frame in its strict sense.

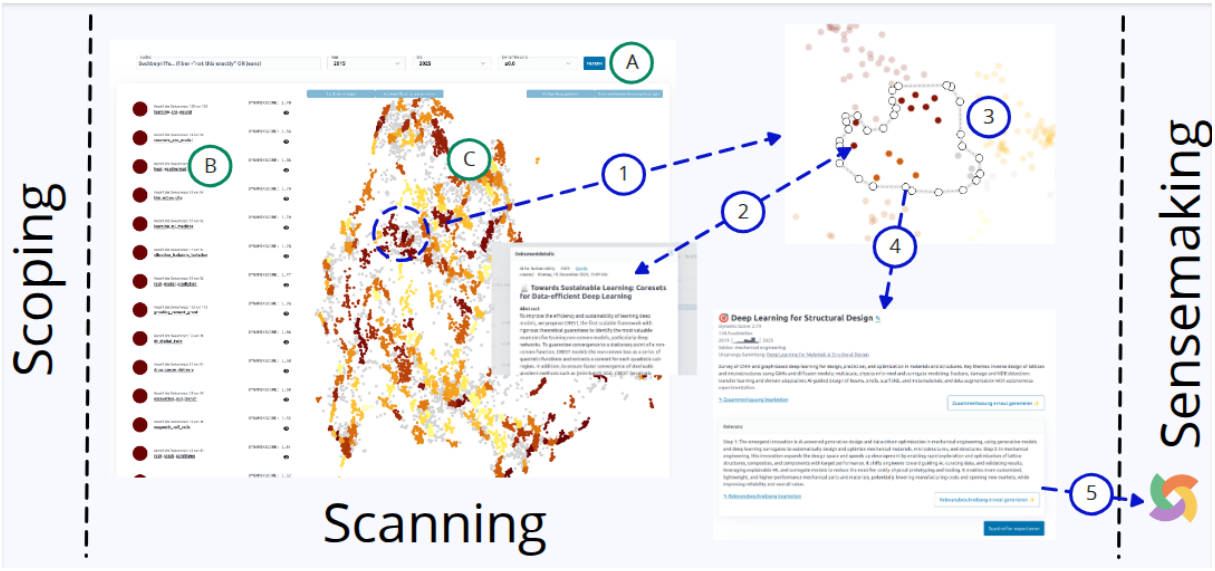

Figure 1: Overview of the main user interface consisting of: (A) Filter options for the topic map documents by content (fulltext search), timeframe or score. (B) List of clusters visible in current viewport (ranked by growth score and colored using a heatmap approach). (C) Actual topic map with points, each representing a document. Typical *map* interactions (zoom, pan) are available. A typical user interaction is to (1) zoom in on an area of interest, (2) inspect a document and its value creation relevance, and to (3) interact and directly manipulate topics by applying splitting, grouping or creating a new topic using a free-form selection of documents ("lasso-select-tool"). Each topic constructed in this human-machine cooperation can be (4) explored in a new view that shows its schema-guided topic representation and value-creation-relevant documents. Detected topics can be (5) saved by the user and exported for collaborative discussion among domain experts on Samarbeid[a], a third-party forum platform.

[a]https://www.samarbeid.org/

perts jointly interpret the surfaced topics to arrive at shared insights. The system targets scanning specifically: it ingests a scoped corpus, supports individual analysts in exploring and refining emerging topics, and produces curated topic collections that feed into downstream sense-making. Beyond standard topic detection (§4.1), our system adds schema-grounded analysis and interactive refinement (§4.2).

4.1 Topic detection pipeline

We model emerging topic detection as a two-stage NLP pipeline. Given a corpus of time-stamped documents, the pipeline produces a ranked list of emerging topic clusters, each consisting of a distributional topic representation, a growth score, and a set of documents with 2D-reduced coordinates. First, documents are encoded using contextual embeddings and grouped into semantically coherent clusters using BERTopic (Grootendorst, 2022). Second, clusters are ranked by a growth score, defined as the ratio of a cluster's corpus share in the current period to its mean share over preceding periods (Appendix A.1.2).

4.2 Schema-Grounded Components and Interaction

We frame the demo as an interactive NLP system where users iteratively explore and refine model-generated representations. Rather than presenting emerging topics as isolated, static distributional representations, the system provides schema-guided (i) Value Creation Relevance indication and (ii) Contextualized Topic Representations, which can be further manipulated in a (iii) Human-in-the-loop manner.

i) Value Creation Relevance At the document level, we model value creation relevance as a text classification task, estimating whether a document expresses information related to the schema. This assessment helps filter out material that may be topically novel but uninformative for sectoral analysis, so that exploration is guided not only by textual similarity and temporal dynamics, but also by whether the underlying source texts relate to one or more elements of the schema. The classifier is an encoder model trained with distant LLM supervision. We first sample documents from the corpus using inverse-frequency sampling over preliminary

BERTopic clusters, which avoids overrepresenting dense regions. Sampled documents are then annotated for value creation relevance by `gpt-5-nano`, which receives the schema in Table 1 and reasons over its elements when assigning binary labels. Finally, we fine-tune `ModernBERT` (Warner et al., 2024) on the resulting annotated dataset (see Appendix A.2 for hyperparameters).

ii) Contextualized Topic Representation At the cluster level, we augment topic representations with schema-aligned summaries that highlight their semantic relation to the schema. Instead of standard cTFIDF-based topic labels, the system generates structured interpretations that map clusters to schema elements. This makes emerging topics easier to interpret and compare across datasets, and supports a transition from generic topic inspection to context-aware signal assessment. In order to generate such contextualized topic representations, we feed: 1) the distributional representation, 2) the 10 most representative documents of the cluster (MMR-selection), 3) the value creation schema in Table 1 and 4) the search context with which the corpus was gathered, this can be a search query or a free-text definition. With these four elements, `gpt-5` generates a contextualized topic summary, which is schema-aligned with the domain. Table 2 shows an example of both representations for the topic introduced in §1.

iii) Human in the Loop The system implements a human-in-the-loop NLP workflow, where users interact with model outputs to refine clusters and construct higher-level signals. In an interactive clustering and labeling process, users manually curate clusters by grouping, splitting, and annotating them. Thus, detected topics emerge from the interaction between model-generated structure and user-driven refinement. As depicted in Figure 1 a typical user flow begins after scoping: the collected corpus is ingested into the system, and domain experts interact with it individually to identify, refine, and save topics along with their contextualized representations. The saved topics then serve as input to collaborative sense-making, where experts construct shared interpretations from the individual selections.

5 Data

Our system is designed to be data-agnostic, meaning it is suitable for analyzing a variety of text sources and formats. Currently, we operate on heterogeneous text corpora, including scientific abstracts and news articles, treating each as a domain-specific corpus. These domains capture research-driven and market-oriented developments, respectively. Each source is processed separately and retrieved through targeted Boolean queries from Scopus[3] and commercial data providers. Future extensions include patents, pre-prints, and firm-level data to broaden coverage and diversify detectable emerging topics.

6 Human Evaluation

We conduct a small human evaluation of the two schema-grounded components with two domain experts. See Appendix B for full evaluation protocol. For this evaluation, we constructed a corpus at the intersection of Artificial Intelligence and Sustainability Research. Our search yielded 9,655 documents; the full search procedure and term lists are given in Appendix B.1. On this corpus, we had 3,475 LLM-annotated documents for value creation relevance (32% positive).

Topic Representations (N=13) We sampled 13 topics of mixed value creation relevance (30–70% classifier-positive documents). For each topic, annotators independently rated two representations on three 1–5 Likert dimensions (*Value Creation Adherence*, *Usefulness*, *Interpretability*): cTFIDF keywords paired with (a) a plain LLM summary (Boutaleb et al., 2024; Achkar et al., 2025), a commonly used LLM-based baseline, and (b) our schema-grounded summary. Representations were presented blinded and in randomized order.

Relevance Classification (N=56) From the same 13 topics, we drew 56 documents almost evenly distributed across topics and balanced between classifier-positive and classifier-negative predictions. Annotators independently labeled each document as value-creation-relevant or not. We report inter-annotator agreement (Cohen's κ) and classifier accuracy on the consensus subset.

Results Table 3 reports the findings. Value Creation Adherence functions here as a manipulation (*i.e.* schema application) check rather than as a comparison: only one condition receives the schema, so the large margin on this dimension (2.1 vs. 4.6) confirms that schema conditioning takes effect in the intended way, but it does not on its own

[3] www.scopus.com

Keywords	diesel, biodiesel, fuel
+Summary	Diesel and biodiesel blends are enhanced with nanoparticles (metal oxides, CNTs, graphene, etc.) to improve CI engine performance, combustion efficiency, emissions, stability, and tribology. The scope spans feedstocks like waste cooking oil, algae, jatropha, pongamia, neem, and cashew, plus nano-fluid fuels and nano-catalysts. Analyses cover energy/exergy, sustainability, and enviroeconomic impacts using experiments, simulations, and optimization (DOE/RSM/AI).
+Schema	The emergent innovation is the widespread use of nanoparticle additives in diesel and biodiesel fuels and lubricants to improve combustion performance, efficiency, and emissions in mechanical engines. By enabling cleaner emissions and higher efficiency, it creates value by serving *demand* shaped by environmental standards and customer expectations for better engine performance. It changes the value network by adding new *actors*, such as nanoparticle suppliers and additive developers, which contribute material-development, testing, and analytical capabilities as *resources*. Within this value-creation *pattern*, the key *processes* include additive formulation, testing, materials qualification, compatibility assessment, and data-driven optimization—including exergy and sustainability analyses and design of experiments—to tailor blends to specific engines and fuels. These processes create *artefacts* such as nano-additive blends, specialty fuels, and retrofit solutions, influencing costs and pricing while opening potential new *revenue* streams.

Table 2: Both representations for one detected topic (§1). The plain LLM summary is not inaccurate: it characterizes the research area precisely, in the vocabulary of the source documents. It answers a different question, however, than the one horizon scanning asks. The schema-grounded summary re-reads the same documents as a reconfiguration of a value network, making explicit which *actors*, *resources*, *processes*, and *demands* are implicated (schema elements italicized).

establish that the resulting representations are better. The comparative claim rests on Usefulness and Interpretability, where neither condition holds a definitional advantage. On both, the schema-grounded representation was rated higher ($2.5 \rightarrow 4.0$ and $2.7 \rightarrow 4.0$). Given the sample size, we read these as consistent supporting evidence rather than as a precise effect estimate. Inter-rater agreement on the Likert ratings was substantial (ICC = 0.61). For the classification task, inter-annotator agreement was fair (κ =0.35), reflecting its difficulty; nevertheless the classifier matched annotator consensus on 75% of documents.

Dimension	**+Summary**	**+Schema**
Topic representations (N=13, Likert 1–5)		
Value Creation Adherence	2.1 (0.5)	4.6 (0.3)
Usefulness	2.5 (0.4)	4.0 (0.4)
Interpretability	2.7 (0.6)	4.0 (0.5)
Inter-rater ICC	0.61	
Relevance classification (N=56)		
Inter-annotator κ	0.35	
Classifier vs. consensus	75%	

Table 3: Human evaluation results, means over 13 topics of the two annotators' per-topic mean rating, with the standard deviation across topics in parentheses. Both representation conditions include cTFIDF keywords; "+Summary" uses a plain LLM summary, "+Schema" uses our schema-grounded summary.

7 Conclusion

We presented an interactive, schema-grounded system for emerging topic detection that combines distributional modeling with structured interpretation and human-in-the-loop refinement. By linking topics to an explicit value creation schema and enabling iterative exploration, the system transforms raw topic outputs into contextualized signals for horizon scanning. A small human evaluation indicates that schema-grounded summaries are rated more useful and more interpretable than plain LLM summaries, and that the relevance classifier matches expert consensus on the majority of agreed cases. This demonstrates how integrating schema grounding and interaction can make emerging topic detection more interpretable and practically useful for strategic foresight.

Limitations

Data A key limitation concerns data quality, which directly affects the validity of the analysis. While scientific abstracts exhibit relatively high and homogeneous text quality, making them well-suited for this approach, online news data vary substantially in form and reliability. In several cases, datasets proved unusable due to poor text quality, including documents consisting of link lists, job search results, student curricula, or stock price sequences. These formats do not provide coherent textual content. Effective data preparation is

therefore essential. Documents must be filtered to ensure they contain continuous, meaningful text, and timestamps must accurately reflect the time of publication. This is particularly critical for web-derived data, where metadata (e.g., from aggregated search results) may not correspond to actual publication dates.

Evaluation scale and subjectivity. Our human evaluation is small in scale (13 topics for the representation study, 56 documents for the classifier study, two annotators in both cases) and should be read as qualitative evidence rather than a basis for statistical claims. Inter-annotator agreement on the relevance classification was fair (κ =0.35), reflecting the inherent subjectivity of judging value creation relevance even for domain experts familiar with the schema. While the classifier matched annotator consensus on 75% of cases, both this number and the Likert results would benefit from replication with more annotators, more topics, and ideally annotators drawn from outside the development team.

Absence of a system-level comparison. Our evaluation ablates the schema-grounded summarizer against a plain LLM summarizer, but does not compare the system as a whole against existing interactive or emerging-topic systems. Such a head-to-head comparison is not currently well-defined: there is no shared benchmark for schema-grounded emerging topic detection, and the closest systems differ in both inputs and outputs. BERTrend (Boutaleb et al., 2024) and Zve et al. (2025) produce temporal topic signals without interactive refinement; TopicGPT (Pham et al., 2024) and Knowledge Navigator (Katz et al., 2024) support interactive exploration of static corpora without temporal emergence or domain grounding. Comparing them would therefore require first agreeing on a task definition and an output format that all systems can produce. Constructing such a benchmark, with a shared time-stamped corpus and expert-annotated emerging topics, is in our view the main open problem for evaluating systems of this kind, and a prerequisite for the comparative claims we do not make here.

Evaluation of the interface. Our evaluation targets the two schema-grounded components, not the interactive interface itself. We do not measure whether grouping, splitting, and free-form selection help analysts arrive at better or faster curated topics, nor do we report usability or task-completion data from the annotation sessions. The evidence we present therefore concerns the representations the system produces, not the interaction design through which analysts explore them. A task-based study comparing curation with and without the schema-grounded views, together with standard usability instrumentation, is a necessary next step before claims about the interactive workflow can be supported.

Schema and language scope. The value creation schema is a deliberate abstraction designed to apply across sectors, but its eight elements may not capture sector-specific dynamics with equal fidelity in all domains. Both the schema-grounded summarizer and the relevance classifier are also dependent on the schema in its current form: changes to the schema would require regenerating LLM annotations and retraining the classifier. Our current evaluation is restricted to English-language scientific abstracts and news articles; performance on other languages, document genres (e.g., patents, regulatory filings), or rapidly shifting domains remains untested.

Reliance on proprietary LLMs. The schema-grounded summarizer uses `gpt-5`, and classifier training annotations are produced with `gpt-5-nano`. This introduces dependencies on a closed model whose behavior, availability, and cost may change over time, and limits full reproducibility. Replacing the LLM components with open-weight alternatives is a natural next step.

Learning from curation actions. Our human-in-the-loop interface supports manual curation through grouping, splitting, and labeling, but user interactions do not currently feed back into retraining / fine-tuning an underlying topic model or classifier. Refinements are persisted as user-specific views rather than used as supervision to update model parameters. Closing this loop, for instance through online learning from curation actions, is a direction we leave for future work.

Acknowledgments

This research was funded by the German Federal Ministry of Research, Technology and Space (BMFTR) under grant 02J21A00X. We thank all project partners for their collaboration and the Scientific Advisory Board for their valuable guidance and support.

References

Ahmed Abolfadl, Marwa Mahmoud Abla, Mervat Abu-Elkheir, and Maggie Mashaly. 2025. AutoCluster, AutoTopicModeling, AutoTrendAnalysis: A complete AutoML pipeline for predicting emerging trends. In *2025 International Conference on Computer and Applications (ICCA)*, pages 1–6, Bahrain Campus, Arab Open University, Bahrain. IEEE.

Pierre Achkar, Satiyabooshan Murugaboopathy, Anne Kreuter, Tim Gollub, Martin Potthast, and Yuri Campbell. 2025. From keyterms to context: Exploring topic description generation in scientific corpora. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pages 102–122, Hybrid. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

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 (FuturED)*, pages 1–17, Miami, Florida, USA. Association for Computational Linguistics.

Manal El Akrouchi, Houda Benbrahim, and Ismail Kassou. 2021. End-to-end LDA-based automatic weak signal detection in web news. *Knowledge-Based Systems*, 212:106650.

Charles J. Fillmore. 1982. Frame semantics. In The Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul. Reprinted in *Cognitive Linguistics: Basic Readings*, ed. Dirk Geeraerts, Mouton de Gruyter, 2006, pp. 373–400, https://doi.org/10.1515/9783110199901.373.

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

Tatsuya Ishigaki, Suzuko Nishino, Sohei Washino, Hiroki Igarashi, Yukari Nagai, Yuichi Washida, and Akihiko Murai. 2022. Automating horizon scanning in future studies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 319–327, Marseille, France. European Language Resources Association.

Uri Katz, Mosh Levy, and Yoav Goldberg. 2024. Knowledge navigator: LLM-guided browsing framework for exploratory search in scientific literature. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8838–8855, Miami, Florida, USA. Association for Computational Linguistics.

Julien Maitre, Michel Ménard, Alain Bouju, and Guillaume Chiron. 2019. Recherche de signaux faibles dans un contexte d'investigation numérique. In *De l'hypertexte aux humanités numériques: actes de H2PTM'19*, Systèmes d'information, web et société, pages 200–215, Montbéliard, France. ISTE Editions.

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.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Alison Smith, Varun Kumar, Jordan L. Boyd-Graber, Kevin D. Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI 2018)*, pages 293–304, Tokyo, Japan. Association for Computing Machinery (ACM).

Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. 2023. Prompting large language models for topic modeling. In *Proceedings of the 2023 IEEE International Conference on Big Data (BigData)*, pages 1236–1241, Sorrento, Italy. IEEE.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Janghyeok Yoon. 2012. Detecting weak signals for long-term business opportunities using text mining of web news. *Expert Systems with Applications*, 39(16):12543–12550.

Evangelia Zve, Benjamin Icard, Alice Breton, Lila Sainero, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2025. From outliers to topics in language models: Anticipating trends in news corpora. In *Proceedings of the 8th International Conference on*

Natural Language and Speech Processing (ICNLSP-2025), pages 385–398, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

A Implementation Details

A.1 Emerging Topic Detection

A.1.1 Topic Detection

For this approach, we use the standard BERTopic package[4]. This standard pipeline consists of mainly four stages: Embedding, Dimensionality Reduction, Clustering and Topic Representation. For embedding, we use the nomic-ai/modernbert-embed-base[5] model (Nussbaum et al., 2024). For the next two stages, we opt for the standard pairing with UMAP (McInnes et al., 2018) and HDBSCAN (McInnes et al., 2017). Finally, in the representation stage, we use class-TFIDF, which was introduced in (Grootendorst, 2022). We tune the following hyper-parameters, in order to arrive at most 100 topics: UMAP - number of neighbors (5—50) and min. distance (0.0—0.5); HDBSCAN - min. cluster size (10—50). For UMAP, we fix the metric to cosine, and euclidean for HDBSCAN. Moreover, we fix the UMAP components to 2, in order to plot the reduced vectors in a 2D plane. All other hyper-parameters use standard values from their implementations.

A.1.2 Topic Ranking

Topics are ranked and color-coded by the ratio of their share of publications at the dataset's leading edge (i.e., over the past two years) to their mean share over the preceding periods. This metric captures recent dynamics in attracting attention while remaining robust to irregular or incomplete time series and leveraging all available observations. As a result, the indicator is not unduly sensitive to temporal gaps and supports limited comparability across datasets with differing coverage or sampling structures. Although we have so far worked exclusively with annual data, the approach could in principle also be applied to other observation frequencies.

The metric computes a growth score for each cluster based on the growth of the document frequency over time. Let: k be a cluster, t be a time period (normally a year), $c_{k,t}$ be the number of documents in k in period t, N_t be the total number of corpus documents in period t. For each period t and cluster k, we normalize the absolute document frequency, where the raw document count is divided by the size of the complete corpus for that period:

$$f_{k,t} = \frac{c_{k,t}}{N_t}$$

Here, $f_{k,t}$ is the normalized frequency, or corpus share, of cluster k in period t. In order to avoid period seasonality, we combine the last two periods into one current period. For the final two periods $T-1$ and T:

$$c_{k,\text{current}} = c_{k,T-1} + c_{k,T}$$

The corresponding corpus denominator is combined in the same way:

$$N_{\text{current}} = N_{T-1} + N_T$$

Normalization is then applied to the combined values:

$$f_{k,\text{current}} = \frac{c_{k,T-1} + c_{k,T}}{N_{T-1} + N_T}$$

Earlier periods are left unchanged.

Now, after normalization, we calculate the mean normalized frequency over all periods before the current period. Given T original periods, the historical mean over the $T-2$ periods preceding the current period is:

$$\bar{f}_k = \frac{1}{T-2} \sum_{t=1}^{T-2} f_{k,t}$$

The growth score P_k of cluster k is the ratio of the current normalized frequency to this historical mean:

$$P_k = \frac{f_{k,current}}{\bar{f}_k + \epsilon}$$

The implementation uses:

$$\epsilon = 10^{-3}$$

to avoid division by zero when a cluster had zero document count on all previous periods. From that definition, we take the following behaviors:

- $P_k = 1$: current activity matches the historical average.
- $P_k > 1$: current activity is higher than the historical average.
- $0 < P_k < 1$: current activity is below the historical average.

[4] https://maartengr.github.io/BERTopic/index.html

[5] https://huggingface.co/nomic-ai/modernbert-embed-base

A.2 Classifier Training

Table 4 lists the hyperparameters used to fine-tune the value creation relevance classifier. The classifier is built on top of the ModernBERT-base encoder (Warner et al., 2024) and trained on the LLM-annotated dataset described in Section 4.2.

Table 4: Classifier fine-tuning hyperparameters.

Parameter	Value	Meaning
Model	ModernBERT-base	Base Hugging Face model for fine-tuning
Max seq. length	512	Maximum tokenized sequence length
Dropout	0.1	Hidden and attention dropout in the classifier head
Learning Rate	2e-5	Training learning rate
Epochs	2	Number of training epochs
Batch Size	32	Per-device training and evaluation batch size
Weight Decay	0.01	Weight decay used by the trainer
Early Stop Patience	5	Early-stopping patience (in evaluation steps)

In addition to the values in Table 4, the trainer uses a fixed set of training arguments: evaluation is performed every 25 steps, checkpoints are saved every 200 steps, 10% of the annotated data is held out as a validation split, and the best checkpoint is selected by F1 score on the validation set.

B Human Evaluation Details

B.1 Corpus Construction

The corpus was retrieved from December 2025. The search is performed by combining two terms list. If the title or the abstract of a document have at least one term from a *methods* vocabulary *and* at least one term from an *application* vocabulary, then this document was collected.

The methods vocabulary comprised 15 terms: *artificial intelligence*, *machine learning*, *deep learning*, *neural network*, *computer vision*, *robotics*, *reinforcement learning*, *data-driven*, *optimization algorithm*, *predictive model*, *smart system*, *iot*, *internet of things*, *autonomous system*, *digital twin*. The application vocabulary comprised 27 terms covering circular economy (*circular economy*, *circularity*, *waste management*, *recycling*, *remanufacturing*, *reverse logistics*, *material recovery*), sustainability and environment (*sustainability*, *sustainable development*, *green technology*, *carbon footprint*, *emission reduction*, *environmental impact*, *climate change*, *ecology*), and energy and efficiency (*energy efficiency*, *energy saving*, *power consumption*, *smart grid*, *renewable energy*, *photovoltaic*, *wind power*, *battery management*, *electric vehicle*, *smart building*, *resource efficiency*, *resource allocation*).

The criterion returned 9,655 records.

B.2 Annotators and Protocol

Both annotators are domain experts with backgrounds in Innovation, Production and Futures Studies and were familiar with the value creation schema (Table 1). Before the main annotation, they completed a brief calibration exercise on one additional topic (not included in the reported sample), with discussion permitted, to align on the interpretation of the Likert dimensions and the relevance criterion. All subsequent annotations were performed independently.

B.3 Sample Selection

We selected 13 topics from the corpus using a mixed-relevance criterion: each chosen topic contains between 30% and 70% classifier-positive documents. This ensures that the evaluation targets hard cases rather than trivially homogeneous clusters. Topic IDs were fixed before annotators received any materials. For the classifier evaluation, we drew 56 documents evenly distributed across the 13 topics, stratified 50/50 between classifier-positive and classifier-negative predictions. Because documents are drawn from mixed-relevance topics, this evaluation measures classifier behavior on difficult decision boundaries rather than corpus-wide prevalence.

B.4 Representation Evaluation Protocol

For each of the 13 topics, annotators were shown three representative documents along with two representations: (a) cTFIDF keywords paired with a plain LLM summary, and (b) cTFIDF keywords paired with our schema-grounded summary. The plain LLM summary is produced by prompting the same model as the schema-grounded summarizer (`gpt-5`) with the representative documents only, omitting the schema; summary length and register were matched between the two conditions to avoid surface-level confounds. Representations were presented as "A" and "B" with assignment randomized per topic, so annotators did not know which system produced which output. Annotators rated each

representation independently on three 1–5 Likert dimensions without access to the ratings of their counterpart. We do not rate cTFIDF keywords in isolation, as keyword lists by design cannot express schema-aligned content.

B.5 Likert Dimensions

Ratings were collected on the following three dimensions:

- **Value Creation Adherence:** Does the output adhere to value creation as the intended analysis domain? Specifically, does it capture the concepts, relations, and system elements defined by the schema in Table 1, rather than drifting into unrelated categories? Since only the schema-grounded condition receives the schema, this dimension serves as a manipulation (schema application) check (§6) rather than as a comparative measure.
- **Usefulness:** Does the output provide actionable information for a horizon scanning process? Specifically, does it help analysts identify relevant developments, emerging patterns, or areas requiring further investigation?
- **Interpretability:** Is the output easy to understand and interpret? Specifically, are the labels, descriptions, and groupings clear, coherent, and sufficiently self-explanatory for analysts to use without extensive further explanation?

The 1–5 scale is defined as: (1) *Very low:* does not satisfy the dimension; (2) *Low:* satisfies the dimension only weakly, with major gaps; (3) *Moderate:* partially satisfies the dimension, with notable limitations; (4) *High:* satisfies the dimension well, with only minor issues; (5) *Very high:* satisfies the dimension fully, with no noticeable issues.

B.6 Classification Evaluation Protocol

Annotators received the 56 documents in shuffled order and independently labeled each as value-creation-relevant or not. Documents were presented without classifier predictions to avoid anchoring. From the resulting labels we compute inter-annotator agreement (Cohen's κ) and, on the subset where annotators agree, the accuracy of the ModernBERT classifier against this consensus.