BERTrend: Neural Topic Modeling for Emerging Trends Detection

Anonymous EMNLP submission

Abstract

Detecting and tracking emerging trends and weak signals in large, evolving text corpora is vital for applications such as monitoring scientific literature, managing brand reputation, and surveilling critical infrastructure. Existing solutions often fail to capture the nuanced context or dynamically track evolving patterns over time. BERTrend, a novel method, addresses these limitations using neural topic modeling in an online setting. It introduces a new metric to quantify topic popularity over time by considering both the number of documents and update frequency. This metric classifies topics as noise, weak, or strong signals, flagging emerging, rapidly growing topics for further investigation. Evaluations on two large real-world datasets demonstrate BERTrend's ability to accurately detect and track meaningful weak signals while filtering out noise, offering a comprehensive solution for monitoring emerging trends in large-scale, evolving text corpora.

1 Introduction

The concept of weak signals, introduced by Ansoff (1975), refers to early indicators of emerging trends that can have significant implications across various domains. These include shifts in public opinion in social trends, early disruptive technologies in innovation, changes in activist groups and public sentiment in politics, and potential disease outbreaks in healthcare. Monitoring and analyzing weak signals offers valuable insights for organizations, researchers, and decision-makers, aiding in informed decision-making.

Key data sources for identifying these trends include large text corpora such as news, social media, research and technology journals or reports. Detecting emerging trends involves challenges like distinguishing meaningful weak signals from irrelevant noise, dealing with context ambiguity, and tracking the extended period over which weak signals may gain significance.

With advances in NLP and AI, researchers have developed various techniques to detect weak signals across different fields (Rousseau et al., 2021), including statistics-based methods, graph theory, machine learning, semantic-based approaches, and expert knowledge. However, most solutions fall short in fully addressing the challenge of detecting emerging trends, either by relying solely on keyword-based analysis, which misses contextual nuances, or by being static and unable to dynamically track evolving weak signals.

In this work, we introduce BERTrend, a novel framework for detecting and monitoring emerging trends and weak signals in large, evolving text corpora. BERTrend leverages neural topic modeling, specifically BERTopic, in an online learning setting to identify and track topic evolution over time. Its key contribution lies in dynamically classifying topics as noise, weak signals, or strong signals based on their popularity trends. The proposed metric quantifies topic popularity over time by considering both the number of documents within the topic and its update frequency, incorporating an exponentially growing decay if no updates occur for an extended period. By combining neural topic modeling with a dynamic popularity metric and adaptive classification thresholds, BERTrend provides a comprehensive solution for detecting and monitoring emerging trends in large-scale, evolving text corpora.

Section 3 details the BERTrend algorithm. In section 4, we introduce the two comprehensive datasets used for experiments and the hyperparameters utilized. Section 5 presents qualitative results, including the overall evolution of trends, specific case studies, enhanced trend interpretability using Large Language Models (LLMs) and the impact of zero-shot topic modeling for targeted monitoring of emerging trends. Finally, we discuss potential future directions and acknowledge the limitations of BERTrend.

2 Background

Weak signal detection and monitoring has been an active research area in recent years, with various methods proposed to identify and analyze early indicators of potential future changes in large datasets. This section provides an overview of the existing approaches and their key characteristics.

One of the most most widely adopted approaches are portfolio maps, pioneered by Yoon (2012), used to visually track several weak signals simultaneously. This technique involves constructing keyword emergence maps (KEM) and keyword issue maps (KIM) based on two key metrics: degree of visibility (DoV) and degree of diffusion (DoD). DoV quantifies the frequency of a keyword within a document set, while DoD measures the document frequency of each keyword. Weak signals are identified as keywords with low frequency but high growth potential. Numerous studies, such as (Park and Cho, 2017), (Donnelly et al., 2019), (Lee and Park, 2018), (Roh and Choi, 2020), (Yoo and Won, 2018), (Griol-Barres et al., 2020), have extended and refined this approach by incorporating aspects like multi-word analysis, signal transformation analysis, and domain-specific applications. However KEMs and KIMs present two major drawbacks: by focusing on keywords only, they can miss the context surrounding a weak signal ; and the output is a single snapshot, which does not gives clear clues of evolution over time.

Several machine Learning techniques have also found applications in weak signal detection: Thorleuchter et al. (2014) developed a semantic weak signal tracing approach using latent semantic indexing (LSI) and singular value decomposition (SVD) to identify signals based on evolving semantic patterns. Yoo and Won (2018) combined agent-based simulation with text mining to forecast innovation and investigate weak signals dynamically. Irimia et al. (2018) proposed a gradient descent-based approach, leveraging supervised learning to identify signals detectable by human experts.

Topic modeling has emerged as a promising approach for weak signal detection, particularly in large textual datasets. Thus, Krigsholm and Riekkinen (2019) and Kim et al. (2019) apply text mining and Latent Dirichlet Allocation (LDA) (Blei et al., 2003), to identify future signals in the domain of land administration and policy research databases. Maitre et al. (2019) integrates LDA and Word2Vec to detect weak signals in weakly structured data. El Akrouchi et al. (2021) introduce furthermore two functions for deep filtering: Weakness, which measures the significance, similarity, and evolution of topics using coherence, closeness centrality, and autocorrelation metrics; and Potential Warning, which further filters the terms of the previously filtered topics to identify potential weak signals.

While traditional topic modeling methods like LDA have been useful for weak signal detection, they have notable limitations: it heavily relies on pre-set topic numbers and fails to benefit from the sophisticated, contextual embeddings provided by modern pre-trained models, resulting in less nuanced analysis. Additionally, it operates on a static basis, overlooking the crucial temporal dynamics of weak signals. In contrast, our approach leverages dynamic, high-quality contextual embeddings from pre-trained models. Unlike keyword-based methods, which can miss the subtleties of context and evolution in signal detection due to their reliance on mere term frequency, our embedding-based technique provides a richer, more adaptive analysis that does not require preset topic counts. This shift from static, keyword-based methods to dynamic, embedding-based analysis allows for a more granular and accurate tracking of the evolution and significance of weak signals over time.

3 BERTrend

In this section, we describe BERTrend (Figure 1), a method for identifying and tracking weak signals in large, evolving text corpora. It leverages the power of BERTopic (Grootendorst, 2022), a state-of-the-art topic model, and wraps it in an online learning framework. In this setting, new data arrives on a regular basis, allowing BERTrend to capture the dynamic evolution of topics over time. The method employs a set of metrics to characterize these topics as noise, weak signals, or strong signals based on their popularity trends. By combining the strengths of neural topic modeling with a dynamic, incremental learning approach, BERTrend enables the real-time monitoring and analysis of emerging trends and weak signals in vast, continuously growing text datasets.

BERTopic leverages pre-trained large embedding models to generate high-quality contextual embeddings of documents, enabling the discovery of meaningful and coherent topics. It utilizes HDBSCAN (McInnes et al., 2017), a hierarchical density-based clustering algorithm, which is robust

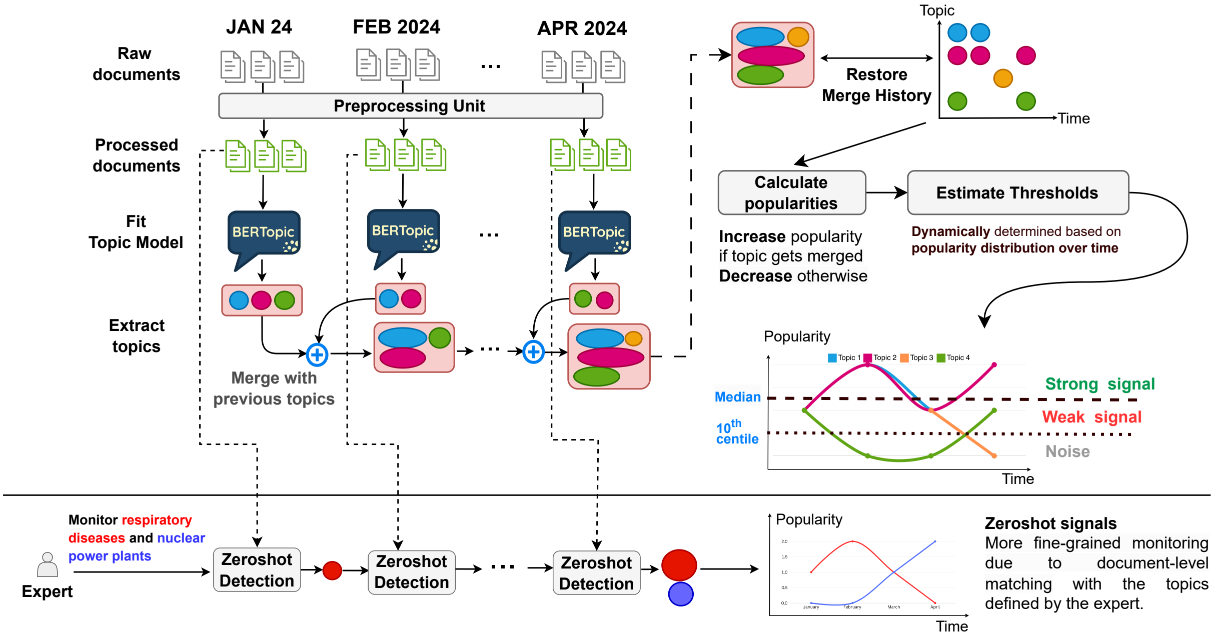


Figure 1: The BERTrend Framework processes data in time-sliced batches, undergoing preprocessing that includes unicode normalization and paragraph segmentation for very long documents. It applies a BERTopic model to extract topics for each batch, which are merged with prior batches using a similarity threshold to form a cumulative topic set. This data helps track topic popularity over time, identifying strong and weak signals based on dynamically chosen thresholds. Additionally, the framework includes a zero-shot detection feature for targeted topic monitoring, providing more fine-grained results due to document-level matching with topics defined by the expert.

to outliers and does not require the number of topics to be specified in advance, allowing the model to automatically determine the optimal number of topics based on the inherent structure of the data.

One of the key advantages of BERTopic is its ability to simulate online learning through model merging. Different BERTopic models can be fitted on documents from non-overlapping time periods and then merged together based on the pairwise cosine similarity between topics of consecutive models, enabling a form of dynamic topic modeling in an online learning setting, where the model can continuously adapt and incorporate new data as it becomes available.

3.1 Data Preprocessing and Time-based Document Slicing

BERTrend preprocesses the input text data by normalizing the text using the NFKC method from the `unicodedata` Python package to handle Unicode characters, special characters, and inconsistencies. Stop word removal and lemmatization are avoided, as BERTopic's underlying framework effectively handles these aspects, and retaining these elements is useful for calculating contextual embeddings using pretrained models.

To accommodate the maximum token lengths recommended by pretrained embedding models and avoid input truncation, lengthy documents are segmented into paragraphs. Each paragraph is treated as an individual document, with a system in place to maintain traceability to its original long document source. This ensures accurate calculation of a topic's popularity over time by considering the original number of documents rather than the inflated number of paragraphs. Abnormally short paragraphs, which often lack sufficient context, are filtered out.

After preprocessing, the entire text corpus D, consisting of N documents, is divided into document slices based on a selected time granularity (e.g., daily, weekly, monthly). A document slice D_t is defined as a subset of documents from D that fall within a specific time interval $[t, t + \Delta t)$, where $t \in \{t_1, t_2, \ldots, t_M\}$, Δt is the chosen time granularity, and M is the total number of document slices. This slicing is crucial for analyzing the temporal dynamics of topics within the corpus.

3.2 Topic Extraction using BERTopic

For each document slice D_t, BERTopic extracts a set of topics $\mathcal{T}_t = \{\tau_t^1, \tau_t^2, \ldots, \tau_t^{K_t}\}$, where K_t is the number of topics in D_t. The process involves:

1. *Document Embedding*: Each document $d \in D_t$ is transformed into a dense vector $\mathbf{e}_d \in \mathbb{R}^h$ using a pre-trained sentence transformer model (Reimers and Gurevych, 2019), where h is the embedding dimension. A topic τ_t^j is described as a set of words $W_{\tau_t^j} = \{w_t^{j,1}, w_t^{j,2}, \dots, w_t^{j,M_j}\}$, where M_j is the number of words representing the topic.

2. *Dimensionality Reduction*: The embeddings are reduced to a lower-dimensional space using UMAP (McInnes et al., 2018), resulting in reduced embeddings $\mathbf{e}'_d \in \mathbb{R}^r$, where $r < h$.

3. *Document Clustering*: The reduced embeddings are clustered using HDBSCAN (McInnes et al., 2017), to group semantically similar documents into clusters. Each cluster $\mathcal{C}_t^j \in \mathcal{C}_t$ is associated with a centroid embedding $\mathbf{c}_t^j \in \mathbb{R}^r$. These clusters represent preliminary groupings of documents that will later be labeled as topics.

4. *Cluster Labeling*: BERTopic assigns labels to clusters to form topics using class-based TF-IDF (c-TF-IDF), considering the frequency and specificity of words within each cluster. Various methods, including LLMs, KeyBERT, and Maximal Marginal Relevance (MMR), can be used to refine the representation of topics. After labeling, each cluster $\mathcal{C}_t^j$ becomes a topic τ_t^j.

3.3 Topic Merging

BERTrend merges topics across document slices to capture their evolution. For each time-based document slice D_{t+1}, the extracted topics $\mathcal{T}_{t+1}$ are compared with the topics from the previous slice $\mathcal{T}_t$ as follows:

1. *Similarity Calculation*: Compute the cosine similarity between each topic embedding $\mathbf{c}_{(t+1)}^j \in \mathcal{T}_{t+1}$ and all topic embeddings $\mathbf{c}_t^k \in \mathcal{T}_t$.
2. *Topic Matching*: If the maximum similarity between $\mathbf{c}_{(t+1)}^j$ and any $\mathbf{c}_t^k$ exceeds a threshold α (e.g., $\alpha = 0.7$), merge the topics and add the documents associated with $\tau_{(t+1)}^j$ to τ_t^k.
3. *New Topic Creation*: If the maximum similarity is below α, consider $\tau_{(t+1)}^j$ as a new topic and add it to $\mathcal{T}_t$.

To maintain topic embedding stability, the embedding of the first occurrence of a topic is retained, preventing drift and over-generalization.

3.4 Popularity Estimation

BERTrend estimates topic popularity over time and classifies them into signal categories based on popularity dynamics. The popularity of topic τ_t^k for document slice D_t is denoted as p_t^k and calculated as follows:

1. *Initial Popularity*: For a new topic τ_t^k of document slice D_t, its initial popularity is set to the number of associated documents: $p_t^k = |D_t^k|$, where D_t^k is the set of documents associated with τ_t^k at time t.
2. *Popularity Update*: For subsequent document slices $D_{t'}$ $(t' > t)$:
   - If τ_t^k is merged with a topic in $\mathcal{T}_{t'}$, its popularity is incremented by the number of new documents: $p_{t'}^k = p_{t'-1}^k + |D_{t'}^k|$.
   - If τ_t^k is not merged with any topic in $\mathcal{T}_{t'}$, its popularity decays exponentially: $p_{t'}^k = p_{t'-1}^k \cdot e^{-\lambda \Delta t^2}$, where λ is a constant decay factor (e.g., $\lambda = 0.01$) and Δt is the number of days since τ^k last received an update.

3.5 Trend Classification

To classify topics into signal categories, BERTrend calculates percentiles of popularity values over a rolling window of size W. For each document slice D_t, two empirical thresholds - the 10th percentile (P_{10}) and the 50th percentile (P_{50}) of popularity values within the window $[t - W, t]$ - are computed. Trend classification is performed based on the topic's popularity p_t^k and its recent popularity trend:

- If $p_t^k < P_{10}$, τ_t^k is classified as a "noise" signal.
- If $P_{10} \leq p_t^k \leq P_{50}$:
  - If the topic's popularity has been increasing over the past few days, as determined by a positive slope of the linear regression line fitted to the topic's popularity values within the window $[t - W, t]$, τ_t^k is classified as a "weak" signal.
  - If the topic's popularity has been decreasing, as determined by a negative slope of the linear regression line, τ_t^k is classified as a "noise" signal, as it likely represents a previously popular topic that is losing relevance.
- If $p_t^k > P_{50}$, τ_t^k is classified as a "strong" signal.

By considering the recent popularity trend in addition to the popularity thresholds, BERTrend ensures that weak signals represent emerging trends with increasing popularity rather than previously popular topics that are losing relevance. This approach helps anticipate and filter out fast the signals that would be considered weak but are instead strong signals that are fading away.

Algorithm 1: BERTrend Algorithm

```
Input: Text corpus $D$, retrospective window size $W$,
       time granularity $G$, similarity threshold $\tau$,
       decay factor $\lambda$
Output: Topics $\mathcal{T}$, popularity $p$, signal classifications
        $S$
Initialize $\mathcal{T} = \emptyset$, $p = \emptyset$, $S = \emptyset$;
$t_{\text{now}} = \text{current time}$;
$t_{\text{start}} = t_{\text{now}} - W$;
$\text{time slices} = \text{slice data}(D, t_{\text{start}}, t_{\text{now}}, G)$;
for $D_t \in$ time slices do
    $\mathcal{T}_t = \text{BERTopic}(D_t)$;
    for $\tau_t^j \in \mathcal{T}_t$ do
        $\text{sim}_{\text{max}} = \max_{\tau_t^k \in \mathcal{T}} \text{Similarity}_{cos}(\mathbf{c}_t^j, \mathbf{c}_t^k)$;
        if $sim_{max} \geq \tau$ then
            $k^* = \arg\max_k \text{Similarity}_{cos}(\mathbf{c}_t^j, \mathbf{c}_t^k)$;
            $D_t^{k^*} = D_t^{k^*} \cup D_t^j$;
            $p_t^{k^*} = p_{t-1}^{k^*} + |D_t^j|$;
        else
            $\mathcal{T} = \mathcal{T} \cup \{\tau_t^j\}$;
            $p_t^j = |D_t^j|$;
    for $\tau_t^k \in \mathcal{T}$ do
        if $\tau_t^k \notin \mathcal{T}_t$ then
            $p_t^k = p_{t-1}^k \cdot e^{-\lambda \Delta t^2}$;
    $\mathbf{P}_{\text{all}} = \bigcup_{\tau^k \in \mathcal{T}} \{p_j^k \mid j \in [t - W + 1, t]\}$;
    $\mathbf{P}_{\text{all}} = \text{sort}(\mathbf{P}_{\text{all}})$;
    $P_{10} = \mathbf{P}_{\text{all}}[\lfloor 0.1 \cdot |\mathbf{P}_{\text{all}}| \rfloor]$;
    $P_{50} = \mathbf{P}_{\text{all}}[\lfloor 0.5 \cdot |\mathbf{P}_{\text{all}}| \rfloor]$;
    for $\tau_t^k \in \mathcal{T}$ do
        if $p_t^k < P_{10}$ then
            $S_t^k = $ "noise";
        else
            if $P_{10} \leq p_t^k \leq P_{50}$ then
                if $slope(\{p_j^k \mid j \in$
                  $[t - W + 1, t]\}) > 0$ then
                    $S_t^k = $ "weak";
                else
                    $S_t^k = $ "noise";
            else
                $S_t^k = $ "strong";
```

Using percentiles calculated dynamically over a sliding window offers several advantages:

1. *Adaptability to datasets*: The retrospective parameter allows the method to adapt to the input data's velocity and production frequency.
2. *Forget gate mechanism*: The sliding window avoids the influence of outdated signals on current threshold calculations.
3. *Robustness to outliers*: Calculating thresholds based on the popularity distribution reduces sensitivity to outlier popularities and prevents thresholds from approaching zero when many signals have faded away.

3.6 Targeted Zero-shot Topic Monitoring

BERTrend includes an optional zero-shot detection feature that allows domain experts to define a set of topics $\mathcal{Z} = \{z_1, z_2, \dots, z_L\}$, each represented by a textual description. The embeddings of these topics and the documents in each slice D_t are calculated using the same embedding model. For each document $d \in D_t$, the cosine similarity between its embedding $\mathbf{e}_d$ and the embedding of each defined topic z_l is computed. Documents with a similarity score above a predefined low threshold β (typically 0.4-0.6) for any of the defined topics are considered relevant and included in the corresponding topic's document set $D_t^{z_l}$. The low threshold accounts for the presumed vagueness and generality of the expert-defined topics, as they have incomplete knowledge that would be supplemented by new emerging information. Finally, the popularity and trend classification for the zero-shot topics are performed in the same manner as for the automatically extracted topics, using the document sets $D_t^{z_l}$ instead of D_t^k.

4 Experimental Setup

4.1 Datasets

We evaluated our approach on two datasets: the arXiv dataset, containing scientific paper abstracts in the computer science category (cs.*) (Cornell-University, 2023), and the New York Times (NYT) news dataset (Tumanov, 2023). The arXiv dataset spans from January 2017 to December 2023 and includes 367,248 abstracts, while the NYT dataset covers the period from January 2019 to January 2023 and includes 184,811 articles. These datasets were chosen for their diverse content and potential to contain topics that could be considered weak signals, such as early warnings about the COVID-19 pandemic. Additionally, they have been used in prior works, providing a basis for comparison and validation of our approach.

4.2 Algorithm parameters

In our experiments, we used the BERTopic framework with carefully selected hyperparameters to optimize weak signal detection performance. We chose the "all-mpnet-base-v2" [1] sentence transformer for document embedding because of its strong performance on various natural language understanding tasks (Reimers and Gurevych, 2019).

In the UMAP dimensionality reduction step, the number of components is set to 5 (default value), and the number of neighbors to 15, which allows UMAP to balance local and global structure in the

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

data, as lower values focus more on local structure while higher values emphasize broader patterns (McInnes et al., 2018). In the HDBSCAN clustering step, we set the minimum cluster size to 2, the smallest possible value, to detect fine-grained clusters. This choice, combined with the "leaf" cluster selection method instead of "excess of mass," generates precise clusters suitable for weak signal detection. The minimum sample size was set to 1, the smallest possible value, to reduce the likelihood of points being declared as noise, as the high number of clusters obtained reduces the need for conservative clustering (McInnes et al., 2017).

Topics were represented by top unigrams and bigrams based on their c-TF-IDF scores, and a minimum similarity threshold of 0.7 (empirically chosen) was used for merging topics across time slices. This threshold ensures the coherence and consistency of the detected topics while allowing room for topics to semantically fluctuate and not be too rigid in the merging process. For the granularity of the time slices, we chose 2 and 7 days for the NYT News the arXiv datasets respectively (values selected empirically to accommodate the rapidly evolving nature of world news compared to the slower pace of research papers).

In the zero-shot example (subsection 5.4), we used a lower similarity threshold of 0.45 for merging topics to accommodate the vague and incomplete nature of the user-defined topics, allowing for a more flexible merging process. This approach maximizes the recall in detecting potentially relevant documents of weak signals.

5 Results

This section provides a qualitative analysis of our method's results, focusing on key aspects to highlight its effectiveness and potential applications.

5.1 Overall results

Figure 2 illustrates the evolution of signal type counts and topic counts in the NYT News dataset and the arXiv cs.* papers dataset We observe striking differences in the signal type distributions between these datasets, which can be attributed to the very nature of their respective domains.

In the NYT News dataset, the number of weak signals remains relatively stable over time, with a manageable quantity of 10 to 20 signals every 2 days. This is well-suited for real-time monitoring and trend detection in fast-paced news cycles, where emerging signals quickly evolve into hot topics of discussion. The occasional spikes in strong signals likely correspond to major events or trending news stories that capture significant attention.

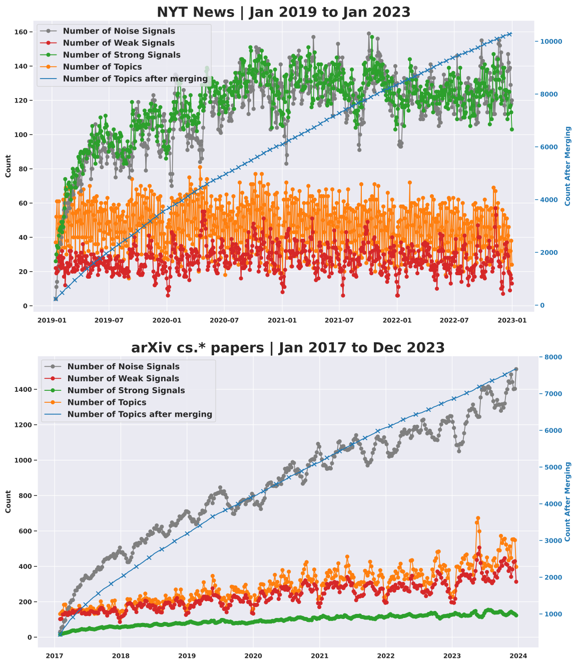


Figure 2: Evolution of Signal Types and Topic Counts in the NYT News and arXiv cs.* Datasets

Conversely, the arXiv cs.* papers dataset exhibits a consistently higher number of weak signals, reflecting the diverse range of emerging research topics in the computer science domain. The number of strong signals is comparatively lower, as only a subset of novel ideas and approaches eventually gain traction and become widely adopted. This aligns with the nature of scientific research, where numerous proposals emerge, but only a few ultimately make a significant impact.

Interestingly, while the number of topics per time slice in the NYT News dataset fluctuates but remains overall stable, the arXiv cs.* papers dataset shows an increasing trend in the number of topics detected per 7-day interval. This can be attributed to the exponential growth of research papers in recent years, leading to a more diverse and rapidly evolving research landscape. The total number of topics after merging (blue line) steadily increases over time in both datasets, reflecting the accumulation of new topics as the datasets grow.

5.2 Case study

In this section, we conduct a qualitative analysis of the results, We focus on a subset of illustrative topics and zoom into key periods to observe their behavior more closely. The examples are selected for their ease for interpretation.

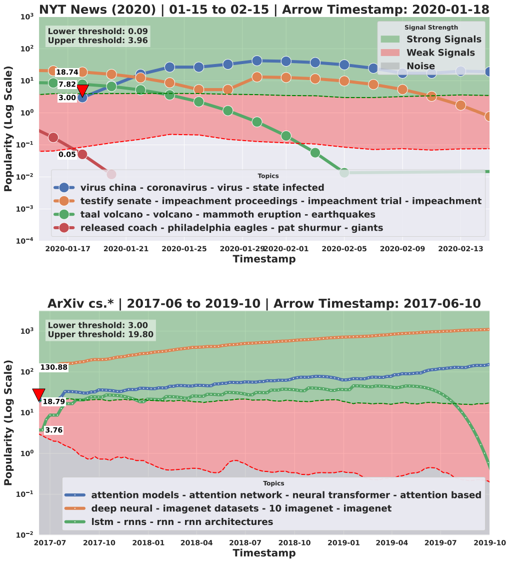


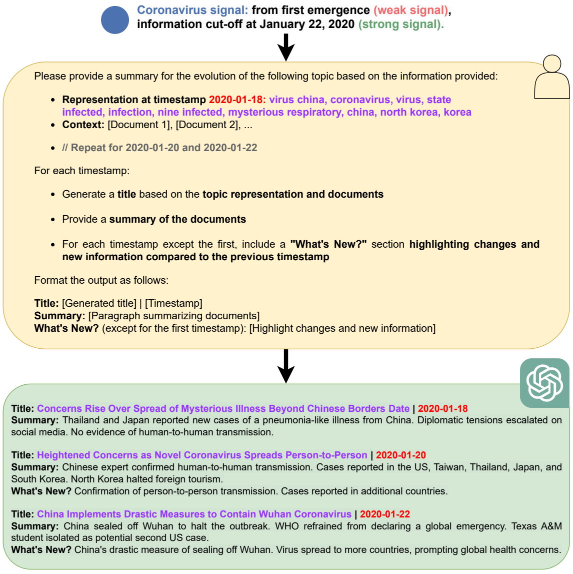
Figure 3: Log-scaled popularity of selected topics from (a) the NYT News dataset and (b) arXiv cs.* papers.

Figure 3a focuses on the period from 01/2020 to 02/2020, when news media began reporting on the COVID-19 outbreak. We observe the appearance of a new topic (blue signal), due to its dissimilarity with pre-existing topics. Initially, the blue signal is classified as weak because of the low number of articles discussing it. Shortly after, it gains traction, transitioning from a weak to a strong signal within a matter of days, as evidenced by its exponential rise in popularity on the log-scaled y-axis. Concurrently, other strong signals during this period include topics related to the impeachment trial of President Trump (orange signal) and the Taal Volcano eruption (Philippines) in Jan 2020 (green signal), while a topic discussing American football teams (red signal) is classified as noise.

In Figure 3b, we showcase the evolution of three selected topics from the arXiv cs.* papers dataset from 06/2017 to 10/2019. The blue signal, representing attention models, was initially a weak signal before June 2017, as attention methods were being used in conjunction with recurrent networks. However, the introduction of the transformer architecture (Vaswani et al., 2017) in June 2017 marked a turning point, after which the topic quickly gained traction, transitioning into a strong signal and eventually becoming a mega-trend. This rise of transformers largely replaced RNNs (Rumelhart et al., 1986) and LSTMs (Hochreiter and Schmidhuber, 1997) (green signal) in NLP tasks, leading to a decline in the popularity of the green signal. In contrast, papers related to computer vision, especially those mentioning ImageNet (Deng et al., 2009), a widely-used dataset in computer vision, were classified as strong signals in June 2017 and continued to exhibit growth. This analysis demonstrates our method's ability to identify potentially impactful research topics early on, track their evolution, and capture the dynamics between related topics.

5.3 Interpretation of signals with LLMs

Topic modeling methods often output topics as sets of keywords, which can be difficult to interpret and may not fully capture the semantic meaning of the topic (Rijcken et al., 2023; Rüdiger et al., 2022).


Coronavirus signal: from first emergence (weak signal), information cut-off at January 22, 2020 (strong signal).

Please provide a summary for the evolution of the following topic based on the information provided:

- **Representation at timestamp 2020-01-18: virus china, coronavirus, virus, state infected, infection, nine infected, mysterious respiratory, china, north korea, korea**
- **Context:** [Document 1], [Document 2], ...
- **// Repeat for 2020-01-20 and 2020-01-22**

For each timestamp:

- Generate a **title** based on the **topic representation and documents**
- Provide a **summary of the documents**
- For each timestamp except the first, include a **"What's New?"** section **highlighting changes and new information compared to the previous timestamp**

Format the output as follows:

Title: [Generated title] | [Timestamp]
Summary: [Paragraph summarizing documents]
What's New? (except for the first timestamp): [Highlight changes and new information]

Title: Concerns Rise Over Spread of Mysterious Illness Beyond Chinese Borders Date | 2020-01-18
Summary: Thailand and Japan reported new cases of a pneumonia-like illness from China. Diplomatic tensions escalated on social media. No evidence of human-to-human transmission.

Title: Heightened Concerns as Novel Coronavirus Spreads Person-to-Person | 2020-01-20
Summary: Chinese expert confirmed human-to-human transmission. Cases reported in the US, Taiwan, Thailand, Japan, and South Korea. North Korea halted foreign tourism.
What's New? Confirmation of person-to-person transmission. Cases reported in additional countries.

Title: China Implements Drastic Measures to Contain Wuhan Coronavirus | 2020-01-22
Summary: China sealed off Wuhan to halt the outbreak. WHO refrained from declaring a global emergency. Texas A&M student isolated as potential second US case.
What's New? China's drastic measure of sealing off Wuhan. Virus spread to more countries, prompting global health concerns.

Figure 4: Enhancing Signal Interpretation using LLMs

Figure 4 demonstrates how LLMs can be leveraged to enhance the interpretation of signals detected by BERTrend and of their evolution over time. In this example, we use the GPT-3.5 Turbo model to generate insightful summaries and highlight new information at each timestamp for the COVID-19 signal. To engineer the prompt for the LLM, we pass the topic representation at each timestamp, along with the associated documents for added context. This approach provides the LLM with a rich understanding of the topic's evolution and enables it to generate more accurate and informative summaries.

However, the maximum token length limitation of 16,385 tokens imposed by GPT-3.5 Turbo presents a challenge when dealing with long-running topics. To mitigate this issue, we deliberately select a date close to the signal's emergence for the example, ensuring that the context fits within the token limit. The generated summaries offer a concise yet comprehensive overview of the topic's evolution over time, effectively capturing the dynamic nature of the signal. By highlighting the new information at each timestamp, the LLM helps identify the key developments and changes in the topic, providing valuable insights for trend analysis and decision-making.

5.4 Impact of zero-shot Topic Modeling

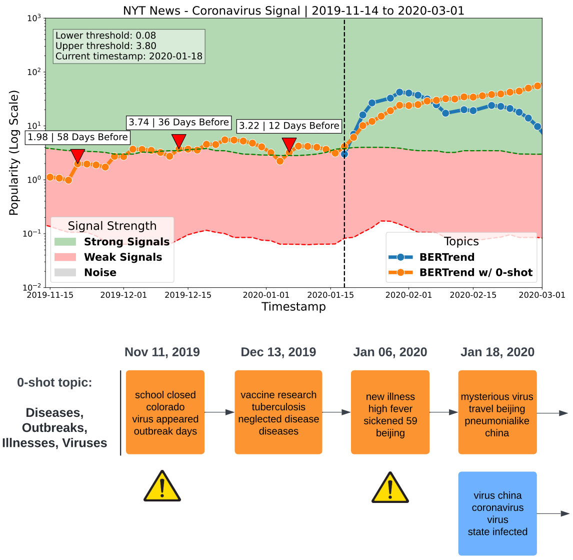


Figure 5: Comparison of COVID-19 Signal Detection with and without zero-shot Topic Modeling

Figure 5 illustrates the impact of incorporating zero-shot topic modeling in the BERTrend algorithm. In this approach, an expert defines a general topic of interest, and each document from a slice is compared against this topic using embedding similarity. Documents that surpass a certain similarity threshold are captured, allowing for targeted weak signal detection. This method enables experts to focus on specific topics of interest while offering higher precision and sensitivity in weak signal detection. By performing document-level comparisons using embeddings, the zero-shot approach minimizes the risk of missing relevant documents during the topic modeling pipeline.

In the provided example, we chose the generic zero-shot topic `"Diseases, Outbreaks, Illnesses, Viruses,"` to detect the COVID-19 signal, simulating a scenario where an expert has a general idea of what to monitor but lacks precise knowledge of an impending outbreak. Remarkably, the zero-shot method identified the earliest article in the dataset mentioning the coronavirus pandemic on January 6th, 2020, referring to it as a "pneumonia-like mysterious virus" alongside "coronavirus". This detection occurred 12 days before the automatic BERTrend usage without zero-shot. Furthermore, the zero-shot approach captured potential weak signals even earlier, such as a November 2019 article reporting school closures in Colorado due to a virus outbreak. While these signals may or may not be directly related to the pandemic, they demonstrate the method's ability to identify potentially relevant events. The consistency of the signal's growth is also notable. The automatically detected signal (blue) by BERTrend starts to decrease and becomes less stable around March 2020, not due to a loss in popularity, but because other signals discussing slightly different aspects of the pandemic begin to emerge.

6 Conclusion

In this paper, we introduced BERTrend, a novel framework for detecting and monitoring weak signals in large, evolving text corpora. BERTrend models the trends of topics over time and classifies them as weak signals, strong signals, or noise based on their popularity, which is quantified using a metric proportional to the number of documents within the topic and its update frequency, with exponential decay for long periods without updates. The classification is performed using empirically chosen thresholds based on the distribution of topic popularities over a sliding window.

The other contributions of this work include: (1) an extensive evaluation on two real-world datasets (NYT news articles and the arXiv cs.* papers) that demonstrate the effectiveness of our approach ; (2) some proposals to leverage LLMs to enhance the interpretation of topic evolution; and (3) the impact of incorporating zero-shot topic modeling into the BERTrend framework.

By the EMNLP 24 conference, we will open-source BERTrend to foster collaboration and advancement in weak signal detection. Future work includes exploring the usage of named entity recognition and knowledge graphs for further filtering and distinguishing of weak signals from noise, investigating different datasets, and developing metrics for comparing weak signal detection methods.

7 Limitations

7.1 Hyperparameter Sensitivity

BERTrend's performance is sensitive to various hyperparameters, including BERTopic parameters, merge threshold, granularity, and retrospective period. We chose BERTopic hyperparameters to produce the most fine-grained topics since larger topics will hinder the early detection process, and weak signals will get lost as the documents that should form them are assigned either to noise topics or other large, more generalized topics. To mitigate the variability of topic embeddings due to the small number of documents per topic, we selected a low merge threshold (0.6-0.7). Granularity depends on the amount of data available per time unit and the frequency of new documents. The retrospective period affects the influence of past signals on current thresholds; we found that a period of a week to a month doesn't change thresholds significantly, but bigger changes can affect classification results. Empirically fixed thresholds (10th percentile and median) balance precision and recall.

7.2 Topic Modeling Limitations

We observed that BERTopic occasionally assigns documents to existing clusters when they would be better suited as standalone topics. This can result in important documents being lost in the closest cluster, hindering early detection. A more robust approach worth investigating is training a BERTopic model on historical data to form numerous topics, then comparing each new document individually with pre-existing topics. If the similarity is high enough, the document is merged; otherwise, it forms a new cluster. This document-level operation would provide more control and bypass topic modeling mishaps at the cost of performance.

7.3 Distinguishing Between Weak Signals and Noise

There remains the challenge of distinguishing between what's considered a weak signal and what's considered noise. Relying on temporal popularity fluctuations alone isn't ideal, as both weak and noise signals behave very similarly. There's also the issue of characterizing what would be a "weak signal," since that changes from one person to another, one domain to another, etc. This is why we added the zero-shot detection to help an expert guide the detection process. We envision exploring the effect of using named entity recognition for better filtering in future work.

7.4 Limits of zero-shot method

One disadvantage of the zero-shot method is that the low similarity threshold chosen to maximize recall, combined with the incomplete description of the zero-shot topic, may capture false alarms such as articles discussing other diseases. However, this approach still serves as a powerful tool to significantly narrow down the number of documents to review based on embeddings, facilitating more targeted analysis by domain experts.

7.5 Evaluation Challenges

Evaluating the effectiveness of our weak signal detection method is challenging due to many factors:

- the subjective nature of what constitutes a weak signal, since it depends on the context, the domain, and the specific goals of the analysis, making it difficult to raise a consensus even among domain experts.
- the lack of ground truth data: unlike many other natural language processing tasks, there are no widely accepted benchmark datasets or ground truth annotations specifically designed for evaluating weak signal detection. This lack of standardized benchmarks hinders the ability to objectively compare different approaches and quantify their performance.
- dynamics over time: weak signals are often transient and can grow or dissipate over time. This dynamic nature complicates the evaluation process, as the ground truth itself may change, requiring continuous monitoring and updating of the evaluation data.

Whereas this work has focused on qualitative evaluation of trends and weak signals, future work should explore methods for quantitative evaluation, development of methodologies to keep the human in the loop, and comparison of different approaches.

References

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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cornell-University. 2023. arxiv dataset. Accessed: 2024-06-14.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Hayoung Kim Donnelly, Yoonsun Han, Juyoung Song, and Tae Min Song. 2019. Application of social big data to identify trends of school bullying forms in south korea. *International journal of environmental research and public health*, 16(14):2596.

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.

Israel Griol-Barres, Sergio Milla, Antonio Cebrián, Huaan Fan, and Jose Millet. 2020. Detecting weak signals of the future: A system implementation based on text mining and natural language processing. *Sustainability*, 12(19):7848.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Alina Irimia, P Paul, and Radu Gheorghiu. 2018. Tacit knowledge-weak signal detection. *Natural Language Processing meets Journalism III*, page 31.

Hyunuk Kim, Sang-Jin Ahn, and Woo-Sung Jung. 2019. Horizon scanning in policy research database with a probabilistic topic model. *Technological Forecasting and Social Change*, 146:588–594.

Pauliina Krigsholm and Kirsikka Riekkinen. 2019. Applying text mining for identifying future signals of land administration.

Young-Joo Lee and Ji-Young Park. 2018. Identification of future signal based on the quantitative and qualitative text mining: a case study on ethical issues in artificial intelligence. *Quality & Quantity*, 52(2):653–667.

Julien Maitre, Michel Menard, Guillaume Chiron, and Alain Bouju. 2019. Détection de signaux faibles dans des masses de données faiblement structurées. *Recherche d'Information, Document et Web Sémantique*, 3(1).

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*.

Chankook Park and Seunghyun Cho. 2017. Future sign detection in smart grids through text mining. *Energy Procedia*, 128:79–85.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Emil Rijcken, Floortje Scheepers, Kalliopi Zervanou, Marco Spruit, Pablo Mosteiro, and Uzay Kaymak. 2023. Towards interpreting topic models with chatgpt. In *The 20th World Congress of the International Fuzzy Systems Association*.

Seungkook Roh and Jae Young Choi. 2020. Exploring signals for a nuclear future using social big data. *Sustainability*, 12(14):5563.

Pauline Rousseau, Daniel Camara, and Dimitris Kotzinos. 2021. Weak signal detection and identification in large data sets: a review of methods and applications.

Matthias Rüdiger, David Antons, Amol M Joshi, and Torsten-Oliver Salge. 2022. Topic modeling revisited: New evidence on algorithm performance and quality metrics. *Plos one*, 17(4):e0266325.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986. *Biometrika*, 71:599–607.

Dirk Thorleuchter, Tobias Scheja, and Dirk Van den Poel. 2014. Semantic weak signal tracing. *Expert systems with applications*, 41(11):5009–5016.

Alexander Tumanov. 2023. New york times articles dataset. Accessed: 2024-06-14.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sun Hi Yoo and DongKyu Won. 2018. Simulation of weak signals of nanotechnology innovation in complex system. *Sustainability*, 10(2):486.

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.