TrendRep: A Long Context Embedding-Based Trend Representation for Weak Signal Detection

Anonymous ACL submission

Abstract

Weak signal detection traditionally relies on counting-based representations of the data, tracking feature frequencies, such as keywords or topics, over time. However, these methods struggle with adaptability and often fail to detect trends at an early stage. In this work, we propose TrendRep, a novel embedding-based trend representation that leverages long context embeddings to encode richer semantics within time windows, providing a more robust and adaptable approach to weak signal detection. To evaluate TrendRep, we construct a new dataset and introduce a quantitative evaluation framework with defined ground truth and key performance metrics. Experimental results show that TrendRep outperforms conventional approaches, demonstrating the effectiveness of embedding-based representations and highlighting the potential of long context embeddings for weak signal detection.[1]

1 Introduction

In today's fast-paced world driven by constant change and big data, detecting emerging trends at an early stage is crucial for informed decision-making and strategic planning. Since Ansoff (1975) coined the term "weak signal", it has been widely used to describe subtle, emerging trends that were not significant in the past but predicted to rise in the future (Holopainen and Toivonen, 2012; van Veen and Ortt, 2021; Ha et al., 2023). Weak signal detection from text data has been extensively studied across various data sources, including news articles (Yoon, 2012; El Akrouchi et al., 2021), academic research papers (Boutaleb et al., 2024), and social media data (Nazir et al., 2019). Insights from these studies have had significant impacts on domains such as technology (Ebadi et al., 2022), politics and economy (Baumeister and Kilian, 2016), business and finance (Mühlroth and Grottke, 2018), and healthcare (Nicolaidou et al., 2021). On the road to more effective weak signal detection, there are mainly two challenges: detecting signals early enough to act on them and distinguishing meaningful signals from random fluctuations.

Weak signal detection relies on a trend representation of text corpora, which defines how signals are extracted and analyzed over time. A common approach is counting-based representations, where weak signals are detected by tracking the raw frequency of specific keywords (Yoon, 2012) or topics (Park and Kim, 2021). This method assumes that a rise in frequency reflects the emergence of a meaningful trend. However, it suffers from two critical limitations. First, counting-based methods treat signals as isolated frequency shifts, ignoring the semantic relationships between keywords or topics. Just as the Bag-of-Words model disregards word order and context (Salton et al., 1975), these methods fail to capture how weak signals interact, evolve, or influence one another. This lack of contextual understanding leaves considerable room for developing better trend representations of text data. Second, counting-based representations are highly sensitive to dataset characteristics. Since weak signals are identified based on raw frequency changes, the same threshold that works well in one dataset may completely fail in another. As a result, traditional methods require extensive threshold tuning for each dataset to balance sensitivity and precision. This lack of adaptability makes counting-based approaches impractical for real-world applications where emerging trends must be detected dynamically and across diverse data sources. These limitations motivate the development of a more context-aware and adaptable trend representation for weak signal detection.

On the other hand, text embeddings produced by modern transformer-based models (e.g., Devlin

[1]The implementation of TrendRep and the Trends2025 dataset are available at https://anonymous.4open.science/r/TrendRep-EE16. We will make the repository public upon acceptance.

et al., 2019; Touvron et al., 2023) have demonstrated strong performance across a wide range of NLP tasks that require deep natural language understanding. To overcome the constraints of input length, researchers have explored training long context embedding models from scratch (Chen et al., 2024) and extending context windows of existing models (Zhu et al., 2024). However, benchmarks such as Bai et al. (2024) primarily evaluate long context embeddings on a narrow set of tasks, leaving their broader potential in information retrieval and text mining unexplored. In particular, the effectiveness of long context embeddings in weak signal detection remains uncertain, as it is unclear how well they can represent chronological document sets and capture underlying trends. In this work, we investigate whether long context embeddings offer a more effective trend representation than traditional counting-based methods.

Challenges also lie in the quantitative evaluation of weak signal detection. Most prior studies have conducted qualitative analysis (El Akrouchi et al., 2021; Boutaleb et al., 2024) due to the lack of well-defined ground truth data, as discussed in BERTrend (Boutaleb et al., 2024). To the best of our knowledge, no widely accepted metrics or benchmarks currently exist for this task. As a result, there is also a lack of objective comparisons between different approaches, making it difficult to measure progress in the field. In this work, we take an initial step toward establishing a quantitative evaluation framework for weak signal detection. We construct datasets specifically designed for this task, define ground truth timestamps indicating trend emergence, and systematically compare our approach with existing methods.

To summarize, our contributions are as follows:

- We introduce a novel trend representation in contrast to the conventional counting-based representation, and reveal that our approach detects weak signals more effectively and robustly across diverse datasets.
- We explore the potential of long context embeddings beyond conventional benchmarks, and investigate their capabilities in representing chronological document sets and capturing trends.
- We provide a paradigm for quantitative evaluation of weak signal detection, laying the groundwork for approach comparisons and future advancements in the field.

2 Related Work

2.1 Topic Modeling

Topic modeling has played a central role in weak signal detection by uncovering latent topics within document collections. Conventional approaches, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non-Negative Matrix Factorization (NMF) (Févotte and Idier, 2010), rely on Bag-of-Words representations to model documents as mixtures of latent topics.

Recent advancements in topic modeling have leveraged text embeddings for better representations of the text. Sia et al. (2020) applied centroid-based clustering on word embeddings, while Angelov (2020) introduced joint document and word semantic embeddings to derive topic vectors. Grootendorst (2022) further extended these methods by employing Sentence-BERT (Reimers and Gurevych, 2019) to generate document embeddings, and Wu et al. (2024) proposed a novel Embedding Transport Plan to map document embeddings into topic embeddings.

These embedding-based approaches enable the computation of topic embeddings, typically represented as centroids of document embedding clusters. Collectively, they have established a strong foundation for the embedding-based weak signal detection in our work.

2.2 Weak Signal Detection

Early weak signal detection methods were keyword-based, detecting trends by tracking keyword occurrence over time. Yoon (2012) introduced keyword portfolio maps, which were later adapted for signal detection across various domains (Sheng et al., 2017, 2019; Goria, 2022). More recently, topics have largely replaced keywords to better capture underlying trends, as first explored by Park and Kim (2021) and further demonstrated in Ebadi et al. (2024).

Both keyword-based and topic-based weak signal detection rely on counting-based representations, extracting signals by measuring keyword or document counts within a corpus. Figure 1(a) uses hypothetical topics to illustrate a counting-based representation, where document count changes over time serve as signals for weak trend emergence. The significance and dynamics of these sig-

nals are typically evaluated using metrics such as frequency, velocity (growth rate), and acceleration. However, these methods often fail to detect trends early, as they depend on frequency shifts rather than deeper semantic patterns. They also struggle to generalize across different datasets, requiring extensive threshold tuning for optimal performance.

In contrast, we introduce TrendRep, a trend representation based on long context embeddings, which captures topic evolution through semantic relationships rather than relying solely on frequency-based metrics. TrendRep operates within an embedding space and is sensitive to dynamic changes like velocity and acceleration, allowing for more robust and adaptable weak signal detection.

2.3 Topic Detection and Tracking (TDT)

TDT and weak signal detection both analyze evolving topics within text corpora, but they differ in objectives, methodologies, and outputs. The term TDT was introduced by Aiello et al. (2013) and is defined as the task of extracting topics from a stream of textual information and quantifying their trend over time. TDT primarily focuses on describing past events, clustering or classifying documents into coherent topic groups to track the evolution of discussions. In contrast, weak signal detection aims to identify emerging trends as early as possible, often before they gain widespread recognition. Rather than focusing on topic structure, weak signal detection emphasizes dynamic patterns within the data.

Despite these differences, weak signal detection often shares techniques with TDT, such as topic modeling. Additionally, datasets originally designed for TDT can be repurposed for trend detection with proper annotation of trend starting points along the timeline.

2.4 Long Context Embedding

Most existing approaches for long context embedding models rely on backbone models that are native to handling long context inputs. More recently, Zhu et al. (2024) proposed adapting context window extension techniques, originally developed for LLMs, to improve text embedding models. A simple technique is Parallel Context Windows (PCW) (Ratner et al., 2022), where a long input is first segmented into shorter chunks, each processed individually by a text embedding model. The final long context embedding is obtained by aggregating the embeddings of all chunks. To mitigate the loss of interactions between chunks, more advanced techniques have been introduced, such as SelfExtend (Jin et al., 2024), DCA (An et al., 2024), PI (Chen et al., 2023), and YaRN (Peng et al., 2023), which incorporate refined position embedding techniques to better preserve the underlying relationships between chunks.

In our work, individual documents are relatively short, and we treat all documents within a time range as a single long context. Since our focus is on trend representation rather than preserving interactions between documents, cross-document interactions are less of a concern. Given its simplicity and effectiveness, we adopt PCW as the basis for our approach.

3 Methodology

In this section, we introduce TrendRep, a novel trend representation designed to capture trends from arbitrary text corpora. We detail the process of extracting TrendRep from a corpus and demonstrate its use in weak signal detection.

3.1 TrendRep Extraction

Extracting TrendRep involves two key components: topic embeddings and temporal embeddings.

3.1.1 Topic Embeddings

Topic embeddings are calculated based on text embeddings, as described in Section 2.1. For datasets without predefined topic labels, embedding-based topic modeling is applied to discover underlying topics in the corpus and compute the embedding of each topic. For datasets with predefined topic labels, topic embeddings are defined as the cluster centroids of the text embeddings for all documents within each topic. The embedding for topic k is calculated as follows:

$$CE_k = \frac{\sum_{i=1}^{n} E_i}{n} \quad (1)$$

where E_i is the ith document embedding of topic k, and n is the total number of documents sharing the topic label k. The notation *CE* stands for cluster embeddings to avoid confusion with *TE* for temporal embeddings below.

Figure 1(b) illustrates the topic embeddings in the embedding space, where three hypothetical topics (CE_1, CE_2, CE_3) are represented as dots.

3.1.2 Temporal Embeddings

Temporal embeddings represent all documents within a specific time window. To compute these

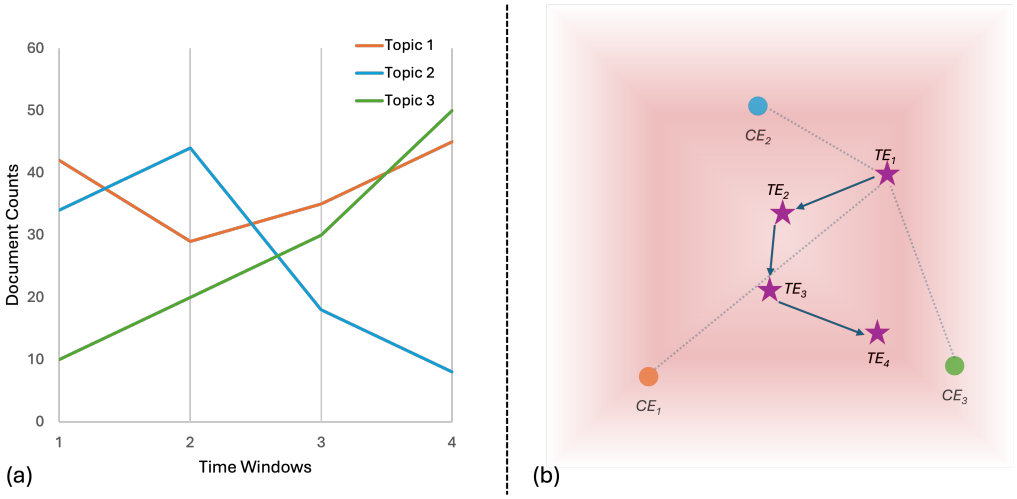

Figure 1: Comparison of Counting-Based and Embedding-Based Weak Signal Detection. **(a) Counting-based weak signal detection** tracks trends by analyzing document count changes over time. **(b) Embedding-based weak signal detection** represents topics and time windows in a shared embedding space. Topic embeddings (dots) define the semantic structure, while temporal embeddings (stars) capture the distribution of topics at different time steps. The movement of temporal embeddings over time reflects trend evolution.

embeddings, the corpus is segmented into time windows (e.g., weeks, days, or hours) based on the dataset's characteristics. A long context consists of all documents within a time window, and its text embedding serves as the temporal embedding. Among approaches for computing long context embeddings described in Section 2.3, we adopt PCW for its simplicity and effectiveness. Following the PCW experiment setup in Zhu et al. (2024), temporal embeddings are computed by averaging document embeddings within each time window, with no overlap between adjacent documents. The calculation of temporal embedding at time step t is as follows:

$$TE_t = \frac{\sum_{j=1}^{m} E_j}{m} \tag{2}$$

where E_j is the jth document embedding in time window t, and m is the total number of documents within time window t.

In Figure 1(b), temporal embeddings are shown as stars (TE_1, TE_2, TE_3, TE_4), and their movement in the embedding space illustrates the evolution of trends over time.

3.1.3 Similarity Matrix

Inspired by prior work in topic modeling (Grootendorst, 2022; Wu et al., 2024), which computes pairwise similarity scores between document embeddings and topic embeddings to assign topic labels to new documents, we extend this approach to reveal the relationship between time windows and topics.

Specifically, we compute pairwise similarity scores between each temporal embedding and each topic embedding, producing a $T \times K$ similarity matrix, where T is the number of time windows, and K is the number of topics.

As formulated in Eq. (3), the similarity score between temporal embedding TE_t and topic embedding CE_k is computed as the Euclidean distance, followed by an exponential transformation to map the distance to a 0-1 range.

$$sim_{tk} = \exp(-\|TE_t - CE_k\|^2) \tag{3}$$

In Figure 1(b), the dotted lines between temporal embedding TE_1 and topic embeddings (CE_1, CE_2, CE_3) visually represent these similarity relationships, indicating how the temporal embedding is positioned based on its proximity to different topic embeddings.

3.1.4 Normalization

Normalization is applied sequentially for the similarity matrix, first along the topic axis and then the temporal axis. For the topic axis, we adopt parameterized softmax normalization as described in Jang et al. (2017). At each temporal step t, we generate a normalized vector $y_t \in \Delta^{K-1}$, in which

$$y_{tk} = \frac{\exp(-\|TE_t - CE_k\|^2/\tau)}{\sum_{k'=1}^{K} \exp(-\|TE_t - CE_{k'}\|^2/\tau)} \tag{4}$$

for $k = 1, \ldots, K$, where the differentiable parameter $\tau = 0.1$.

For normalization along the temporal axis, we first experiment with the full normalization across all temporal slices as formulated in Eq. (5). However, it risks leaking future data into past trend detection.

$$z_{tk} = \frac{y_{tk}}{\sum_{t'=1}^{T} y_{t'k}} \tag{5}$$

To address the risks in full normalization, we also experiment with step-wise normalization restricted to temporal slices preceding and including the current time window, as formulated in Eq. (6).

$$z_{tk} = \frac{y_{tk}}{\sum_{t'=1}^{t} y_{t'k}} \tag{6}$$

This normalized similarity matrix reflects the topic distribution within each time window and serves as the basis for analyzing topic evolution over time. We refer to this matrix as TrendRep.

3.2 Dynamic Metrics

We calculate two key dynamics of TrendRep as the main indicators for weak signals:

- Velocity: The difference in normalized similarity scores between consecutive time windows.
- Acceleration: The difference in velocity between consecutive time windows.

Weak signals are detected by applying simple thresholds on velocity and acceleration, such as velocity > 0 and acceleration > 0. These thresholds effectively highlight topics experiencing rapid changes over time.

3.3 Practical Considerations

When computing TrendRep, it is crucial to address potential biases that can arise from overlap between topic and temporal embeddings. Specifically, if the same document embeddings are used to compute both topic embeddings and temporal embeddings, the similarity scores between them will tend to be disproportionately high compared to other pairs. This inflated similarity arises because the shared embeddings introduce artificial alignment, undermining the objectivity of the signal detection process.

To mitigate this bias, we propose using a subset of document embeddings, or sub-clusters, for calculating topic embeddings. These sub-clusters should include only a portion of the document embeddings within each topic. Importantly, the documents included in these sub-clusters should be excluded from the calculation of temporal embeddings. By ensuring no overlap between the embeddings used for topic and temporal representations, we can maintain unbiased similarity scores.

Dataset	#Docs	#Topics	#Trends	#TWs
News2013	8,726	222	180	144
Trends2025	1,668,934	2,663	100	216

Table 1: Dataset Statistics

4 Quantitative Experiments

In this section, we present our datasets, evaluation metrics, and experiment setup for the quantitative evaluation of weak signal detection. A valid weak signal should indicate the onset of an emerging trend, and our evaluation is designed based on this premise.

4.1 Datasets

We use two datasets for quantitative experiments: News2013 and Trends2025, whose statistics are summarized in Table 1. Each dataset is required to contain: (1) documents with timestamps indicating their release times, (2) a topic label for each document, and (3) ground truth timestamps, marking the start of each emerging trend. We describe below how these datasets were collected or modified to satisfy these requirements.

4.1.1 News2013

News2013 originates from the topic detection and tracking (TDT) study in Jiang et al. (2024). It was derived as the English subset from the multilingual news dataset in Miranda et al. (2018), which itself is based on Rupnik et al. (2015).[2] We adopt this dataset because each document is accompanied by (1) a release timestamp and (2) a topic event label. These labels have served as ground truth in prior TDT experiments (Jiang et al., 2024), and we use them as the predefined topic labels to avoid introducing bias from topic modeling. Our experiments are conducted on the test set, split by hours.

However, to repurpose this TDT dataset for weak signal detection, we must address two key issues:

1. **Scarcity of Topics**: News2013 test set contains 222 topics scattered among 10 months, averaging less than one active topic per day. Sparsely distributed topics make the experiment results less convincing, even when

[2] https://github.com/Priberam/news-clustering/, distributed under The 3-Clause BSD License

trends are successfully detected. To address this, we compressed the timeline by parallelly relocating the 205 short-lived topics and their associated news articles into a six-day span. Each topic was randomly assigned a starting point within the span. The remaining 17 longer-duration topics were excluded from our experiments.

2. **Lack of Trend Start Labels**: News2013 lacks explicit annotations for when topics start trending. To generate these labels, we employed a statistical trend detection method. First, we counted the number of news articles per topic per hour. Next, we applied the Pettitt Test (Pettitt, 1979) to identify change points in each trend, followed by the Mann-Kendall Test (Mann, 1945; Kendall, 1975) to detect increasing trends. This combined approach of Pettitt Test and Mann-Kendall Test, originally proposed in Helsel and Hirsch (1993), has been widely used in various fields such as environmental science (Slater et al., 2021), finance (Yoo et al., 2021), healthcare (Chen et al., 2022), and academic research (Curiac et al., 2022). The detected change points preceding increasing trends were annotated as the ground truth for trend start times.

4.1.2 Trends2025

Trends2025 is an original dataset created to provide diverse weak signal distributions and more accurate trend start annotations.

We collected 100 topics that began trending between January 2 and January 9, 2025, using Google Trends (Google Trends, 2025). For each topic, we retrieved its name, a list of associated search keywords in English, and the timestamp marking when it started trending. We searched Reddit by these keywords, and gathered the top 250 posts for each keyword within the same date range, resulting in $250 \times$ *number of keywords* posts per topic.

Although the Reddit API limited us from retrieving more posts via keyword searches, we identified 15,779 seed subreddits where these top posts originated. To enrich the dataset and simulate real-world scenarios where numerous topics are simultaneously active, we collected all new posts created in these seed subreddits between January 1 and January 9, 2025. The title and the body of each post were stacked together to form one document.

One missing component in this dataset was document-level topic labels. Since not all posts were related to one of the 100 trending topics, we performed topic modeling from scratch to allow for the presence of other unrelated topics. We first computed topic embeddings for the 100 trending topics using the top posts retrieved via keyword searches, where each document embedding was computed using a Sentence-BERT model[3]. Following Boutaleb et al. (2024), the corpus was split by hours, and a BERTopic model was trained for each hour. Topics with cosine similarity below 0.7 were merged. This resulted in a total of 2,663 topics, of which 100 corresponded to the trending topics of interest, while the remaining topics were treated as noise during evaluation.

4.1.3 Datasets Comparison

Table 1 compares the sizes of the two datasets. Beyond dataset size, we also highlight the differences in weak signal distributions that stem from varying noise levels. Figure 2(a) shows the ratio distribution of emerging topics among all active topics with positive document counts. A higher ratio indicates fewer noise signals. Figure 2(b) illustrates the percentile rank distribution of the document count for emerging trends at their start time. A higher percentile rank suggests that noise signals have lower activeness. These metrics significantly influence conventional weak signal detection approaches that rely on raw feature frequencies, requiring careful threshold tuning for optimal performance.

4.2 Evaluation Metrics

In our experiments, the outcome for each topic is a set of timestamps marking the detected starting points of the trend. To evaluate the effectiveness of weak signals as early indicators of emerging trends, we compute recall, precision, and F1-score by comparing the detected timestamps against the ground truth.

However, requiring detected timestamps to match the exact ground truth does not align with the goal of weak signal detection, which is often to detect trends before they gain widespread attention. To address this, we expand the acceptable detection window to include the 24 hours preceding the ground truth timestamp. Recall, precision, and F1-score are then calculated based on this expanded

[3]Distributed under Apache License 2.0: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, distributed under Apache License 2.0

Figure 2: Comparison of weak signal distributions in the two datasets. (a) The ratio distribution of emerging topics among active topics. (b) Document count percentile rank distribution at the trend start time.

24-hour range as the evaluation criteria.

In addition to these metrics, we calculate the root mean square error (RMSE) for the Trends2025 dataset to measure the average time deviation between detected timestamps and the ground truth. RMSE provides insight into how early or late the detected trends are relative to the actual trend start times. Smaller RMSE values indicate higher detection accuracy. Notably, RMSE is calculated only for the Trends2025 dataset and not for News2013, where topics in the ground truth may have no trends or multiple trends. In such cases, RMSE becomes unreliable and therefore is excluded from the evaluation.

4.3 Experiment Setup and Results

In our experiments, we focus on comparing trend representations rather than signal detection algorithms. Therefore, weak signals are detected by applying simple thresholds to each representation, and an emerging trend is detected when a topic transitions from having no weak signals to carrying a weak signal. Since different trend representations are derived from various dataset characteristics, the thresholds in use are selected to fit each representation.

We evaluate TrendRep against two baseline representations: document counts and topic popularity. Document counts represent a widely used frequency-based method, while topic popularity follows the state-of-the-art approach described in Boutaleb et al. (2024). The experiments are conducted on both News2013 and Trends2025, using hourly time granularity for signal detection.

For TrendRep, we first compute document embeddings using a Sentence-BERT model[4], followed by the computation of topic and temporal embeddings. For News2013 dataset, despite the bias mentioned in Section 3.3, all documents are used to calculate both embeddings due to data scarcity. For Trends2025 dataset, topic embeddings are derived from the top posts retrieved via keyword searches, while temporal embeddings are computed from the remaining documents. Using these embeddings, we construct a similarity matrix and apply normalization. As mentioned in section 3.2.1, normalizing across all time windows could inadvertently introduce future information into weak signal detection. To address this, we report results for two variants in Table 2: TrendRep-full, which is fully normalized across all time windows, and TrendRep-step, which applies step-wise normalization where only temporal embeddings up to the current time window are used. Finally, velocity and acceleration are calculated for each topic at every hour, and a weak signal is detected when both exceed zero.

For the document counts baseline, we adopt the method proposed by Park and Kim (2021), where a topic is considered to carry a weak signal when its proportion among all topics is below the average, and its growth rate is positive. For the topic popularity approach, we follow the method proposed by Boutaleb et al. (2024). The same decay factor and thresholds are adopted, where a topic is considered to carry a weak signal if its popularity falls between the 10th and 50th percentiles.

The experiment results are shown in Table 2.

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, distributed under Apache License 2.0

Approach	News2013			Trends2025			
	Precision	Recall	F1	Precision	Recall	F1	RMSE
Doc Counts	None Detected			12.57	86.87	21.97	87.55
Popularity	5.83	7.22	6.45	3.45	5.05	4.10	131.94
TrendRep-full	**7.67**	**56.67**	**13.51**	**14.38**	**100**	**25.14**	**77.24**
TrendRep-step	6.77	55.0	12.06	12.86	**100**	22.80	83.46

Table 2: Experiment Results

4.4 Results Analysis

As discussed in Section 4.1.3, News2013 and Trends2025 exhibit different weak signal distributions due to varying noise levels. Despite these differences, TrendRep demonstrates superior performance on both datasets compared to conventional trend representations, with the fully normalized TrendRep slightly outperforming its step-wise normalized variant.

Using document counts as a trend representation can yield decent performance on data with a certain noise level, as evidenced by the results on Trends2025. However, this approach lacks robustness across diverse datasets and fails to detect any weak signals in News2013, which has a different noise signal distribution. This limitation arises because, in News2013, noise topics are relatively inactive, while trending topics inherently have document counts above the mean value. As a result, effectively detecting emerging trends with document counts requires careful threshold tuning for each dataset or application.

On the other hand, the topic popularity approach yields consistently low performance across both datasets in our experiments. We observed a persistent lag in detection, with most trends identified within 48 hours after their ground truth starting points. While this delayed detection might be useful for general trend analysis, it is unsuitable to detect emerging trends, which requires high sensitivity at the earliest stages of trend development.

5 Conclusions

In this work, we introduced TrendRep, a novel long context embedding-based trend representation that contrasts with conventional counting-based methods. Our experimental results demonstrate that TrendRep detects weak signals as emerging trends more effectively and robustly across diverse datasets. Beyond weak signal detection, our study extends the potential of long context embeddings beyond conventional benchmarks. By leveraging long context embeddings to represent chronological document sets, we explored their potential for capturing semantic structures in evolving textual data. Additionally, we addressed the long-standing challenge of quantitatively evaluating weak signal detection. We introduced a new evaluation framework, defining ground truth trend indicators and key metrics, enabling systematic comparisons between different weak signal detection approaches.

6 Limitations

While TrendRep offers significant improvements over counting-based methods, several limitations remain.

First, our approach relies on Parallel Context Windows for long context embeddings, which is effective when dealing with individually short documents. However, when longer documents are involved, simply aggregating chunk embeddings may not be sufficient to preserve cross-chunk interactions. Future work could explore alternative long context embedding techniques for signal detection.

Second, while TrendRep consistently achieves higher recall, its precision remains modest. This suggests that relying solely on threshold-based detection may not fully address false-positive cases. To further improve precision, we plan to explore more advanced signal detection algorithms that could better model temporal variations and outlier behaviors in emerging trends.

Finally, while our evaluation framework introduces quantitative metrics for weak signal detection, the concept of trend often involves subjective human interpretation. Future research could explore human-in-the-loop evaluation methods or hybrid approaches that integrate qualitative insights with quantitative metrics for a more comprehensive assessment of emerging trends and weak signals.

References

Luca Maria Aiello, Georgios Petkos, Carlos J. Martín-Dancausa, David P. A. Corney, Symeon Papadopou-

los, Ryan Skraba, Ayse Göker, Yiannis Kompatsiaris, and Alejandro Jaimes. 2013. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15:1268–1282.

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. 2024. Training-free long-context scaling of large language models. *arXiv preprint arXiv:2402.17463*.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *Preprint*, arXiv:2008.09470.

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

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Christiane Baumeister and Lutz Kilian. 2016. Forty years of oil price fluctuations: Why the price of oil may still surprise us. *Journal of Economic Perspectives*, 30(1):139–160.

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

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

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Xiang Chen, Hui Wang, Weixuan Lyu, and Ran Xu. 2022. The mann-kendall-sneyers test to identify the change points of covid-19 time series in the united states. *BMC Medical Research Methodology*, 22(1):233.

Christian-Daniel Curiac, Ovidiu Banias, and Mihai Micea. 2022. Evaluating research trends from journal paper metadata, considering the research publication latency. *Mathematics*, 10(2).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashkan Ebadi, Alain Auger, and Yvan Gauthier. 2022. Detecting emerging technologies and their evolution using deep learning and weak signal analysis. *Journal of Informetrics*, 16(4):101344.

Ashkan Ebadi, Alain Auger, and Yvan Gauthier. 2024. WISDOM: an ai-powered framework for emerging research detection using weak signal analysis and advanced topic modeling. *CoRR*, abs/2409.15340.

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.

Cédric Févotte and Jérôme Idier. 2010. Algorithms for nonnegative matrix factorization with the beta-divergence. *CoRR*, abs/1010.1763.

Google Trends. 2025. Google Trends. Accessed: 2025-01-09.

Stéphane Goria. 2022. A deck of cards to help track design trends to assist the creation of new products. *International Journal of Technology, Innovation and Management (IJTIM)*, 2(2):1–17.

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

Taehyun Ha, Heyoung Yang, and Sungwha Hong. 2023. Automated weak signal detection and prediction using keyword network clustering and graph convolutional network. *Futures*, 152:103202.

Dennis R Helsel and Robert M Hirsch. 1993. *Statistical methods in water resources*. Elsevier.

Mari Holopainen and Marja Toivonen. 2012. Weak signals: Ansoff today. *Futures*, 44(3):198–205. Special Issue: Weak Signals.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Hang Jiang, Doug Beeferman, Weiquan Mao, and Deb Roy. 2024. Topic detection and tracking with time-aware document embeddings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16293–16303, Torino, Italia. ELRA and ICCL.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*.

Maurice G. Kendall. 1975. *Rank Correlation Methods*, 4th edition. Charles Griffin.

Henry B Mann. 1945. Nonparametric tests against trend. *Econometrica: Journal of the econometric society*, pages 245–259.

Sebastião Miranda, Artūrs Znotinš, Shay B. Cohen, and Guntis Barzdins. 2018. Multilingual clustering of streaming news. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4535–4544, Brussels, Belgium. Association for Computational Linguistics.

Christian Mühlroth and Michael Grottke. 2018. A systematic literature review of mining weak signals and trends for corporate foresight. *Journal of Business Economics*, 88(5):643–687.

Faria Nazir, Mustansar Ali Ghazanfar, Muazzam Maqsood, Farhan Aadil, Seungmin Rho, and Irfan Mehmood. 2019. Social media signal detection using tweets volume, hashtag, and sentiment analysis. *Multimedia Tools and Applications*, 78:3553–3586.

Olga Nicolaidou, Christos Dimopoulos, Cleo Varianou-Mikellidou, Georgios Boustras, and Neophytos Mikellides. 2021. The use of weak signals in occupational safety and health: An investigation. *Safety science*, 139:105253.

Chankook Park and Minkyu Kim. 2021. A study on the characteristics of academic topics related to renewable energy using the structural topic modeling and the weak signal concept. *Energies*, 14(5).

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Anthony N Pettitt. 1979. A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2):126–135.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2022. Parallel context windows for large language models. *arXiv preprint arXiv:2212.10947*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jan Rupnik, Andrej Muhic, Gregor Leban, Primoz Skraba, Blaz Fortuna, and Marko Grobelnik. 2015. News across languages - cross-lingual document similarity and event tracking. *CoRR*, abs/1512.07046.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Jie Sheng, Joseph Amankwah-Amoah, and Xiaojun Wang. 2017. A multidisciplinary perspective of big data in management research. *International Journal of Production Economics*, 191:97–112.

Jie Sheng, Joseph Amankwah-Amoah, and Xiaojun Wang. 2019. Technology in the 21st century: New challenges and opportunities. *Technological Forecasting and Social Change*, 143:321–335.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

L. J. Slater, B. Anderson, M. Buechel, S. Dadson, S. Han, S. Harrigan, T. Kelder, K. Kowal, T. Lees, T. Matthews, C. Murphy, and R. L. Wilby. 2021. Nonstationary weather and water extremes: a review of methods for their detection, attribution, and management. *Hydrology and Earth System Sciences*, 25(7):3897–3935.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Barbara L. van Veen and J.Roland Ortt. 2021. Unifying weak signals definitions to improve construct understanding. *Futures*, 134:102837.

Xiaobao Wu, Thong Thanh Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024. Fastopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sanghyuk Yoo, Sangyong Jeon, Seunghwan Jeong, Heesoo Lee, Hosun Ryou, Taehyun Park, Yeonji Choi, and Kyongjoo Oh. 2021. Prediction of the change points in stock markets using dae-lstm. *Sustainability*, 13(21).

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.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. LongEmbed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816, Miami, Florida, USA. Association for Computational Linguistics.