PromptStream: Self-Supervised News Story Discovery Using Topic-Aware Article Representations

Arezoo Hatefi[1], **Anton Eklund**[1,2], **Mona Forsman**[2]
[1]Umeå University, Umeå, Sweden
[2]Aeterna Labs, Umeå, Sweden
{arezooh, antone}@cs.umu.se
{anton, mona}@aeternalabs.ai

Abstract

Given the importance of identifying and monitoring news stories within the continuous flow of news articles, this paper presents PromptStream, a novel method for unsupervised news story discovery. In order to identify coherent and comprehensive stories across the stream, it is crucial to create article representations that incorporate as much topic-related information from the articles as possible. PromptStream constructs these article embeddings using cloze-style prompting. These representations continually adjust to the evolving context of the news stream through self-supervised learning, employing a contrastive loss and a memory of the most confident article-story assignments from the most recent days. Extensive experiments with real news datasets highlight the notable performance of our model, establishing a new state of the art. Additionally, we delve into selected news stories to reveal how the model's structuring of the article stream aligns with story progression.

Keywords: news story discovery, online clustering, data stream, contrastive learning, cloze-style prompting, document embedding, document clustering, continual learning, unsupervised, text mining

1. Introduction

In the abundance of news being generated daily, online news story discovery streamlines individual news consumption and is invaluable for news summarization, recommendation systems, and other services reliant on structured news content understanding. The concept of recognizing and tracking topics within a stream was initially introduced in the Topic Detection and Tracking (TDT) task (Allan et al., 1998). This task revolves around techniques for the automated structuring of textual data streams into coherent topic groupings. In the context of a news stream, these topics essentially represent news stories. Individual news articles report on real-world events, and a subset of articles within the news stream that concern the same event constitutes a news story. We present a model that utilizes cloze-style prompting and self-supervised contrastive learning techniques to tackle this task.

Early efforts in news story discovery relied on sparse document representations such as keywords, and TF-IDF vectors (Laban and Hearst, 2017; Staykovski et al., 2019). However, as dense document representations encompassing richer semantic information started to emerge, researchers began exploring their potential in news story discovery. Staykovski et al. (2019) compared TF-IDF and Doc2Vec representations for news story discovery and concluded that sparse representations are better for this task. Recently, Saravanakumar et al. (2021) demonstrated that integrating contextual BERT representations alongside TF-IDF representations could enhance task performance. This improvement could be achieved through fine-tuning BERT on event similarity using a triplet network architecture (Hoffer and Ailon, 2015) and providing external entity knowledge.

Alignment and uniformity (Wang and Isola, 2020) represent fundamental characteristics inherent to any embedding space. In the context of news story discovery, alignment pertains to the proximity of articles related to the same story within the embedding space, while uniformity assesses how uniformly random articles are distributed throughout that space. One reason why document representations from pre-trained language models (PLMs) like BERT, without fine-tuning, are less effective for news event discovery is their lack of uniformity. This uniformity issue makes it challenging to differentiate between two articles that share the same theme but concern distinct events.

In recent years, contrastive learning has demonstrated its remarkable effectiveness in numerous language processing and computer vision tasks (Oord et al., 2018; Van Gansbeke et al., 2020; Radford et al., 2021). This effectiveness stems primarily from its ability to enhance the alignment and uniformity of embedding spaces, as indicated by Wang and Isola (2020). A notable example of this success in news story discovery is the work conducted by Yoon et al. (2023). They used contrastive learning for training story-indicative document representations from sentence representations in a continual learning setting over the news stream and showed that these representations are superior to sparse alternatives.

Recently, prompting has emerged as a groundbreaking technique that has significantly enhanced the performance of natural language processing tasks that require deep understanding (Le Scao and Rush, 2021; Jin et al., 2022; Qi et al., 2022). Template-based prompting methods align with the Masked Language Modeling (MLM) pre-training task of language models. In MLM, a portion of the input tokens are masked and the model is trained to predict those tokens. Similarly, in cloze-style template-based prompting, a template like "A <*mask*> event" is integrated into the input, and the prediction can be obtained by decoding the output embedding associated with <*mask*>. Thus, effectively leveraging the large-scale knowledge of PLMs, ultimately resulting in the generation of more informative representations.

In this paper, we present an approach that employs cloze-style prompting to enhance article representations with topic-related information, tailoring them to the specific needs of news story discovery in a dynamic news stream. Table 1 presents two instances that illustrate how cloze-style prompts select the most topic-related words from the text to generate the article representation. These representations undergo continuous fine-tuning via cluster-level contrastive learning, making use of a memory bank of confident article-story assignments for self-supervision, to remain relevant within the latest context. The primary objective of confidence-aware memory replay is to effectively mitigate concerns regarding data scarcity and ensure the provision of robust supervision for contrastive learning.

The main contributions of this work are:

1. To the best of our knowledge, our approach is the first in its utilization of cloze-style prompting to enhance article representations for news story discovery.

2. We continuously fine-tune article representations via contrastive learning that makes use of a memory bank of confident article-story assignments for self-supervision.

3. We make an extensive experimental comparison of PromptStream with SOTA methods for unsupervised online news story discovery. We use three real news datasets for these evaluations and establish a new state of the art.

4. Additionally, we take a closer look at selected stories to discover connections between natural story progression and how PromptStream structures the article stream.

2. Related Work

2.1. News Story Discovery

Laban and Hearst (2017) create a keyword-based graph of articles within a window spanning over N days by connecting articles that share more keywords than a specified threshold. The system then identifies local topic clusters within overlapping windows using the Louvain community detection algorithm (Blondel et al., 2008). For long-term stories, it combines topics from non-overlapping windows with a similarity above a given threshold. Staykovski et al. (2019) enhance this method by using TF-IDF vectors rather than keywords.

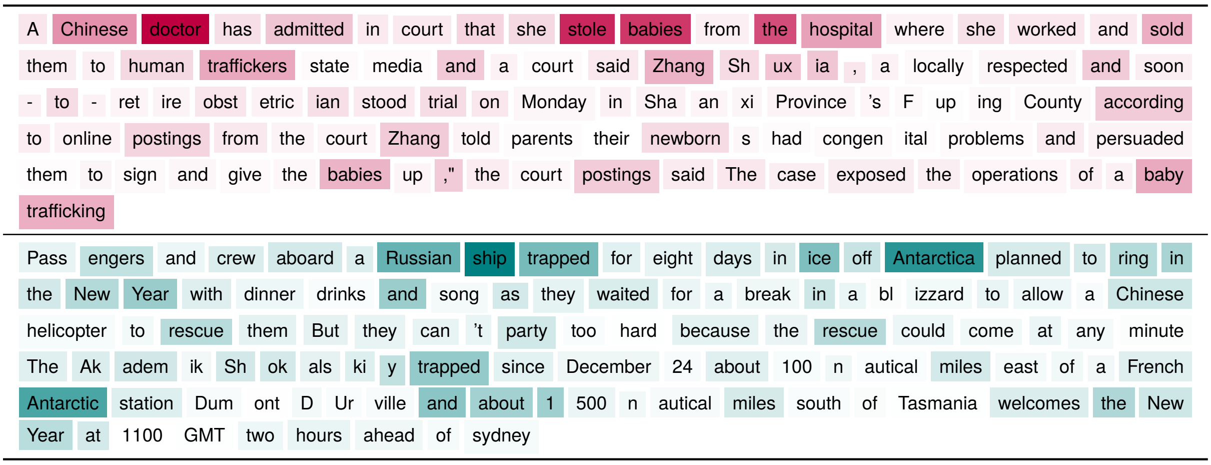
In contrast to the above batch-clustering approach, Miranda et al. (2018) employ an online clustering approach, where streaming news articles are compared against existing topic clusters to find the best match or to create a new cluster. Their method computes the similarities between an article and a cluster according to multiple sparse document representations (such as TF-IDF vectors for title, body, and concatenation of title and body) and then aggregates them using a Rank-SVM model. The decision to merge a document with a cluster or create a new cluster is again taken by an SVM classifier. Both SVM models are trained using a supervised training set. Moreover, it uses article timestamps to avoid merging recent documents with older clusters.

Saravanakumar et al. (2021) follow an approach similar to that of Miranda et al. (2018), but attune BERT embeddings for news event recognition by fine-tuning and adding external entity knowledge. Both Miranda et al. (2018) and Saravanakumar et al. (2021) exploit external knowledge and labeled datasets, which renders them less practical for real-world applications where supervised data is scarce.

In a recent study, Yoon et al. (2023) employ a hierarchical architecture to construct article representations from sentence representations derived from pre-trained sentence encoders. Sentence representations are aggregated into article representations through a one-layer transformer. These representations are then compared with the existing cluster representations within the current window to either identify the best match or establish a new cluster. Notably, the article representations are continuously refined in a self-supervised fashion, with a focus on the most confident assignments within the current window.

2.2. Prompt-Based Prediction

Template-based prompting (Brown et al., 2020; Schick and Schütze, 2021; Gao et al., 2021) approaches NLP downstream tasks as masked language modeling problems. In this approach, a lan-



A Chinese doctor has admitted in court that she stole babies from the hospital where she worked and sold them to human traffickers state media and a court said Zhang Sh ux ia , a locally respected and soon - to - ret ire obst etric ian stood trial on Monday in Sha an xi Province 's F up ing County according to online postings from the court Zhang told parents their newborn s had congen ital problems and persuaded them to sign and give the babies up ," the court postings said The case exposed the operations of a baby trafficking

Pass engers and crew aboard a Russian ship trapped for eight days in ice off Antarctica planned to ring in the New Year with dinner drinks and song as they waited for a break in a bl izzard to allow a Chinese helicopter to rescue them But they can 't party too hard because the rescue could come at any minute The Ak adem ik Sh ok als ki y trapped since December 24 about 100 n autical miles east of a French Antarctic station Dum ont D Ur ville and about 1 500 n autical miles south of Tasmania welcomes the New Year at 1100 GMT two hours ahead of sydney

Table 1: Visualization of selected samples from the News14 dataset. The color saturation indicates the attention the tokens receive from the ⟨*mask*⟩ token in the prompt. These examples are an indication that prompting results in topic-tailored representations for the articles by attending to the most important tokens in the text such as events and named entities.

guage model initially generates an output based on a predefined prompt utilizing a task-specific template that subsequently is mapped to the output space of the downstream task. This methodology allows for cost-effective knowledge extraction from pre-trained language models and maximizes the utilization of pre-trained corpora. It proves to be an ideal approach for tasks like keyword identification and topic detection since it does not rely on external tools or corpora, in contrast to several of the approaches mentioned in Section 2.1. Examples of successful uses of prompting in natural language processing tasks are: (Zhang et al., 2022; Yue et al., 2023) for zero-shot and few-shot event detection, (Jiang et al., 2022) for sentence embedding, and (Shen et al., 2023; Ashok and Lipton, 2023) for named entity recognition.

3. Preliminaries

An article d is a sequence $[w_1, w_2, \ldots, w_{|d|}]$ of words. A news story s is a set of articles, $s = \{d_1, d_2, \ldots, d_{|s|}\}$, all related to the same event. The objective of online story discovery is to incrementally assign each new article d in an unbounded news article stream $\mathbb{D} = [d_1, d_2, \ldots]$ to an existing story or create a new cluster if d does not match any existing one. This process is unsupervised.

To account for the publication time of news articles and prevent the assignment of articles to outdated, no longer relevant stories, we employ the concept of a sliding window $\mathbb{W} \subseteq \mathbb{D}$. This approach is commonly used for mining data streams (Laban and Hearst, 2017; Silva et al., 2013). The window and sliding size determine the time span of interest for ongoing stories and the frequency of updates, respectively. For example, a sliding window of 3 days, sliding by one day, addresses the articles published within the last 3 days, with daily updates.

For simplicity, we assume that each article is associated with a single story, and a story is considered alive if at least one article within the window $\mathbb{W}$ is part of that story. The set of alive stories within the window $\mathbb{W}$ is denoted by $\mathbb{S}_{\mathbb{W}}$.

4. The New Model: PromptStream

PromptStream is an online story discovery model that generates topic-aware representations by employing a cloze-style prompting technique for articles. The model architecture is illustrated in Figure 1 and the procedure is described in Algorithm 1. In summary, new articles within a sliding window are assigned to relevant stories and the article encoder is updated every N days through self-supervised learning. The encoder update process relies on prompt-based article representations and leverages a repository of confident article-story assignments. Detailed explanations of this process are elaborated in the subsequent sections.

4.1. Topic-Aware Article Representation

The representation R_d of an article d is the sum of two distinct representations: prompt-based representation and the output of the mean pooling over the last layer of the PLM:

$$R_d = R_d^{prompt} + R_d^{mean}$$

Prompt-Based representation (R_d^{prompt}) In a news article, not every word carries equal significance in identifying the described events. Some words, particularly named entities, contain a wealth of crucial information. By employing suitable and task-specific prompting templates, we can create

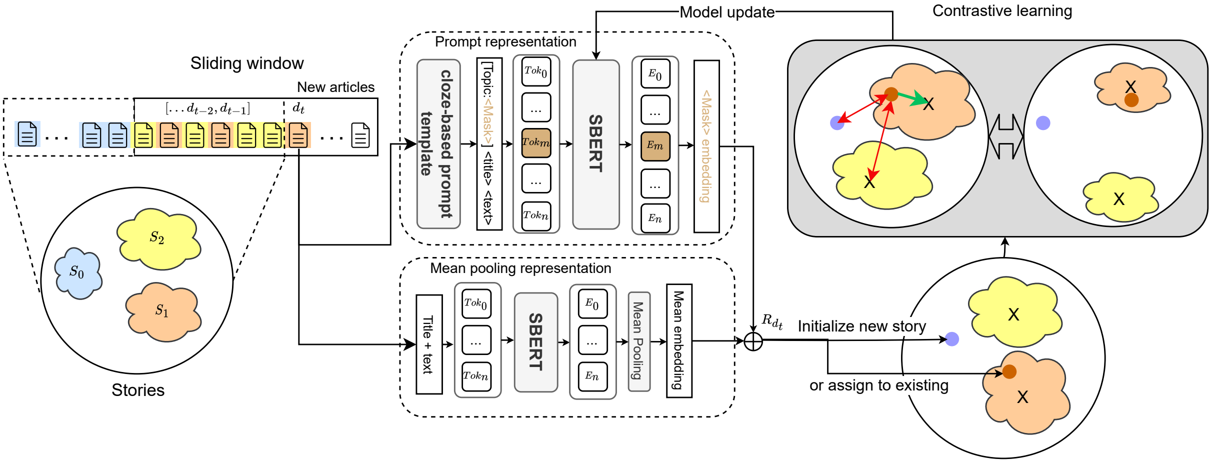


Figure 1: Architectural overview of PromptStream. The article encoder employs one SBERT. However, we use two SBERTs to illustrate the generation process of prompt-based and mean pooling-based article representations. Moreover, the encoder is updated through contrastive learning based on prompt-based article representations.

topic-aware representations that focus more on the critical aspects of the document.

To achieve this, we utilize a cloze-style prompt to extract topic-aware representations from the text. This process involves transforming the article into a cloze-style prompt using the template

```
[ topic :  <mask> ] <title> <body>
```

where `<title>` and `<body>` represent the title and body of the news article, respectively. However, unlike the case of text classification and question-answering tasks, we do not use the label tokens predicted by the PLM classification head, but we use the output of the PLM's last layer for the `<mask>` token as the topic-aware article representation.

Mean-Pooling Representation (R_d^{mean}) Cloze-style prompting focuses on specific tokens or entities within the text, making it well-suited for capturing topic-specific information. Mean pooling, on the other hand, provides a broader and more general representation of the entire document. By combining these two representations, we are effectively leveraging both the fine-grained, contextually rich information obtained from cloze-style prompting and the more holistic and global context captured by mean pooling. This combination results in a more balanced and informative view of the document that is better suited for clustering tasks. Hence,

$$R_d^{mean} = \frac{1}{n}\sum_{i=1}^{n} h_i \ ,$$

where h_i is the embedding of token i from the last layer of teh PLM encoder.

4.2. Online Story Assignment

Dynamic Story Representation The representation R_s of a story is computed as the average of the representations R_d of the articles comprising the story:

$$R_s = \frac{1}{|s|}\sum_{d\in s} R_d \ .$$

This representation is updated each time a new article is allocated to this story.

Article-Story Similarity To determine which story a new document d_i in sliding window $\mathbb{W}$ belongs to, we evaluate the similarity of document d_i with any story $s_j \in \mathbb{S}_{\mathbb{W}}$ by using the *cosine similarity* metric as follows:

$$sim(d_i, s_j) = cos\big(R_{d_i}, R_{s_j}\big)$$

If the highest similarity between d_i and the stories in the window exceeds a predefined threshold θ, we assign d_i to the story s_j that resulted in the highest similarity and update the representation of that story accordingly. Otherwise, we establish a new cluster with document d_i and set the cluster's representation to R_{d_i}. Following Yoon et al. (2023), we set the default value of threshold θ, which defines the granularity of the stories, to 0.5.

4.3. Self-Supervised Continual Learning

We update the encoder every N days using cluster-level contrastive learning, which is applied on prompt-based article representations. This loss function encourages articles to be moved closer to the center of their respective clusters while simultaneously being pushed away from other cluster

Algorithm 1: PromptStream pseudocode

```
Data: $\mathbb{D}$: a news article stream
  $prompt\_enc$: prompting-based encoder
  $mean\_enc$: mean pooling-based encoder
  $update\_freq$: updating frequency of $prompt\_enc$
  $memory$: confident article-story assignments
  $epochs$ and $iters$: hyper-params for updating encoder
  $\theta$: article-story similarity threshold
  $\delta$: confidence threshold
Result: A set $\mathbb{S}$ of stories in stream $\mathbb{D}$
$prompt\_enc \leftarrow$ fine-tune with data of initial $update\_freq$ days in a cold start    ▷ Section 4.3
$\mathbb{S} \leftarrow \emptyset$
$memory \leftarrow \emptyset$
$counter \leftarrow 0$
for every sliding window $\mathbb{W}$ in $\mathbb{D}$ do
    $\mathbb{S}_{\mathbb{W}} \leftarrow$ existing stories in $\mathbb{W}$
    for every new article $d \in \mathbb{W}$ do
        $R_d^{mean} \leftarrow mean\_enc(d)$    ▷ Section 4.1
        $R_d^{prompt} \leftarrow prompt\_enc(d)$    ▷ Section 4.2
        $R_d \leftarrow R_d^{prompt} + R_d^{mean}$
        if $max(\{sim_{d,s_j} | s_j \in \mathbb{S}_{\mathbb{W}}\}) > \theta$ then
            Assign article $d$ to corresponding $s_j$
            if $sim_{d,s_j} > \delta$ then
                $memory \leftarrow memory \cup (d, s_j)$
            end
        else
            $s_{new} \leftarrow$ make a new story with $d$
            $\mathbb{S}_{\mathbb{W}} \leftarrow \mathbb{S}_{\mathbb{W}} \cup \{s_{new}\}$
            $memory \leftarrow memory \cup (d, s_{new})$
        end
    end
    $counter \leftarrow counter + 1$
    if $mod(counter, update\_freq) == 0$ then
        $\mathbb{S}_{mem} \leftarrow$ existing stories in $memory$
        for $epoch$ in $epochs$ do
            for $iter$ in $iters$ do
                $\mathbb{B} \leftarrow$ a batch from $\mathbb{S}_{mem}$ with uniform sampling
                $\mathcal{L}_{cts} \leftarrow$ contrastive loss for $\mathbb{B}$    ▷ Section 4.3
                $prompt\_enc \leftarrow$ update with $\mathcal{L}_{cts}$
            end
        end
        $memory \leftarrow \emptyset$
    end
    $\mathbb{S} \leftarrow \mathbb{S} \cup \mathbb{S}_{\mathbb{W}}$
end
return $\mathbb{S}$
```

centers. Updating the encoder daily with data from the same day can lead to fluctuating distributions, potentially undermining the encoder's consistency. In addition, contrastive learning benefits from an abundance of negative examples, making it more effective to accumulate data over several days and then update the model with this aggregated dataset. Therefore, we integrate the *memory replay* concept from continual learning. This helps prevent catastrophic forgetting and ensures that the encoder remains temporally consistent as the article stream evolves.

Confidence-Aware Memory Replay We establish a memory bank containing data from the most recent N days, which serves as the data source for contrastive learning. In this context, we quantify the confidence of articles by their similarity with the centers of their respective stories. Only samples with confidence exceeding a predefined threshold δ are included in the memory bank.

Uniform Sampling Given the varying sizes of different stories, creating training batches with a random sampler from the memory bank can potentially lead to a trivial solution. This occurs when the vast majority of articles are consistently assigned to just a few stories, causing the encoder to become biased toward those clusters and predict them for all subsequent articles. A strategy to address this issue is to sample articles using a uniform distribution across the clusters. This is equivalent to weighting the contribution of an input to the loss function by the inverse of the size of its assigned cluster. Therefore, for training the encoder, we construct batches using a uniform sampler from the memory bank.

Contrastive Loss Given a batch $\mathbb{B}$ of positive article-story pairs $(d, s) \in \mathbb{B}$ the following contrastive loss function is utilized for fine-tuning the encoder using prompt-based article representations:

$$L_{cts} = - \sum_{(d,s)\in\mathbb{B}} \log \left(\frac{e^{\cos\left(R_d^{prompt}, R_s\right)/\tau}}{\sum_{s' \in \mathbb{S}_{\mathbb{W}}} e^{\cos\left(R_d^{prompt}, R_{s'}\right)/\tau}} \right)$$

Here, τ is the temperature parameter. This loss function encourages articles to be moved closer to the center of their respective clusters while simultaneously being pushed away from other cluster centers. This enhances the uniformity and alignment of the embedding space for prompt-based representations.

To initially fine-tune R_d^{prompt} during the early stages of stream processing, within the initial N days, we utilize R_d^{mean} for clustering articles and computing story representations. Subsequently, we fine-tune R_d^{prompt} using the aforementioned contrastive loss.

5. Experiments

We evaluate the performance of PromptStream on three labeled news datasets in Section 6.1 with common extrinsic clustering evaluation metrics. An ablation study is performed to investigate the impact of different components of the model in Section 6.2. Finally, we make a qualitative analysis to investigate the performance of the model beyond the metrics in Section 6.3.

5.1. Datasets

We conduct experiments on three labeled datasets that were constructed by Yoon et al. (2023) from real news datasets:

NEWS14: This dataset consists of 16,136 articles categorized into 788 unique stories from the year 2014, sourced from the dataset introduced in (Miranda et al., 2018).

WCEP18: This dataset was created by curating 828 news events published in 2018. It comprises 59,073 articles and has been sourced from the WCEP dataset (Gholipour Ghalandari et al., 2020).

WCEP19: This dataset was assembled by selecting 519 events from the year 2019, gathered from the WCEP dataset (Gholipour Ghalandari et al., 2020). It encompasses a total of 37,637 articles.

5.2. Baselines

We compared PromptStream with five state-of-the-art algorithms that can be used for unsupervised and online story discovery: ConStream (Aggarwal and Yu, 2010), NewsLens (Laban and Hearst, 2017), BatClus (Miranda et al., 2018), DenSps (Staykovski et al., 2019), and SCStory (Yoon et al., 2023). ConStream is a widely recognized streaming document clustering algorithm frequently used for story discovery. It relies on keyword-count statistics and employs incremental clustering through micro-clusters. The other three algorithms were discussed in Section 2.1. We adopted the same naming convention for these baseline methods as Yoon et al. (2023). Following Yoon et al. (2023), in the case of BatClus, we employed an unsupervised setting to ensure a fair comparison. For a better assessment, we also compared our model with advanced variants of the existing algorithms that incorporate PLMs.

5.3. Evaluation Metrics

The evaluation metrics used were the score by Bagga and Baldwin (1998), denoted by B^3-F1 here, the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), and the Adjusted Mutual Information (AMI) Vinh et al. (2010). B^3-F1 focuses on the precision and recall of individual data points within clusters, rather than pairwise comparisons, and is considered one of the best metrics to evaluate text clustering algorithms (Amigó et al., 2009). Staykovski et al. (2019) provide an extensive explanation of how it is computed. ARI is a symmetric measure that provides an overall assessment of clustering quality, considering both pairwise agreements and disagreements. AMI also favors clusterings where data points that are similar are placed into the same cluster, but it is less sensitive to clusterings that divide ground truth classes into multiple clusters. ARI and AMI are both adjusted for chance agreement. All three metrics have a maximum score of 1.

Following (Yoon et al., 2023), Table 2 reports the average scores for each metric across each sliding window over the data streams. Table 3 presents these metrics over the complete data streams, an approach we also employ in our analysis detailed in Section 6.2.

5.4. Experiment Settings

We implemented our model in PyTorch (Paszke et al., 2019) with Transformer Library (Wolf et al., 2020) and chose sentence-transformers/all-roberta-large-v1[1] as the PLM for the encoder. This is a roberta-large (Liu et al., 2019) model well-suited for tasks such as clustering and semantic search. For fine-tuning the encoder using prompt-based representations, we used the AdamW (Loshchilov and Hutter, 2019) optimizer with a batch size of 64, and a learning rate of 5e-6. We set the max sequence length for the tokenizer to 128. This choice may be advantageous because, in news articles, the most informative content is typically found in the title and the introductory section of the text. The θ and δ thresholds were both set to 0.5. The window and sliding sizes were 3 and 1, respectively, in both our model and our runs for SCStory. The temperature for contrastive loss for all datasets was 0.2. We updated the encoder every 10 days. Regarding the parameters for SCStory, with the exception of the window size, which we adjusted to 3, we maintained the default values as reported in the paper.

6. Results and Discussions

Here we present the performance evaluation, the ablation study, and the qualitative cluster analysis.

6.1. Overall Performance

Tables 2 and 3 provide a comparison between the baseline models and PromptStream for the online story discovery task. As shown in Table 2, PromptStream exhibits superior performance compared to SCStory achieving a 3.7% higher average B^3-F1 score over sliding windows of the entire data

[1] https://huggingface.co/sentence-transformers/all-roberta-large-v1

	NEWS14			WCEP18			WCEP19		
	B^3-F1	AMI	ARI	B^3-F1	AMI	ARI	B^3-F1	AMI	ARI
ConStream†	0.314	0.128	0.069	0.408	0.444	0.222	0.400	0.497	0.292
NewsLeans†	0.481	0.309	0.077	0.527	0.490	0.117	0.554	0.529	0.141
BatClus†	0.706	0.726	0.572	0.694	0.786	0.571	0.698	0.791	0.574
DenSps†	0.669	0.602	0.358	0.697	0.759	0.487	0.701	0.765	0.487
ConStream+SBERT†	0.434	0.413	0.276	0.701	0.784	0.657	0.704	0.795	0.667
NewsLeans+SBERT†	0.749	0.718	0.564	0.767	0.823	0.631	0.784	0.887	0.664
BatClus+SBERT†	0.764	0.785	0.648	0.751	0.835	0.656	0.759	0.837	0.657
DenSps+SBERT†	0.750	0.720	0.567	0.754	0.824	0.624	0.762	0.830	0.660
SCSTory+SBERT	0.895	**0.873**	**0.837**	0.867	0.876	0.809	0.873	0.89	0.83
PromptStream	**0.915**	0.845	0.835	**0.913**	**0.885**	**0.863**	**0.919**	**0.904**	**0.887**

Table 2: The average B^3-F1, AMI, and ARI over **each sliding window** in the article streams. For PromptStream and SCStory, the scores are the average of five different runs with different random seeds. Scores marked with † are included from Yoon et al. (2023) to self-contain this paper.

	NEWS14			WCEP18			WCEP19		
	B^3-F1	AMI	ARI	B^3-F1	AMI	ARI	B^3-F1	AMI	ARI
SCSTory+SBERT	0.806	0.862	0.294	0.799	0.904	0.628	0.820	0.917	0.718
PromptStream	**0.843**	**0.898**	**0.610**	**0.825**	**0.916**	**0.644**	**0.852**	**0.931**	**0.766**

Table 3: B^3-F1, ARI, and AMI over the **entire** article streams. For PromptStream and SCStory, the results are the average scores of five different runs with different random seeds.

stream across all datasets. Furthermore, with respect to other metrics, it outperforms SCStory in most cases. When assessing metrics across the entire data stream, as presented in Table 3, it becomes evident that PromptStream surpasses SCStory in terms of B^3-F1, AMI, and ARI by an average of 3.1%, 2%, and 12.7%, respectively, across all three datasets.

Both PromptStream and SCStory perform superior to the other baselines, which we attribute to the fact that they both use attention to compute representations that emphasize the relevant parts of each article.

6.2. Ablation Study

Table 4 presents the results of the ablation study. In most cases, the scores closely resemble those of the default model. However, two notable outliers are worth mentioning. Firstly, when the prompting-based representation is removed, there is a substantial drop in the B^3-F1 score. Secondly, the omission of updates to the encoder leads to a dramatic reduction in this score. These results strongly suggest that the main technical contributions of our proposal, the fine-tuned prompting-based representation, and the self-supervised continual learning, indeed play a central role in achieving superior performance.

The results indicate that continual training results in representations that outperform those trained only for the initial 10 days. Additionally, as previously discussed in Section 4.3, the effectiveness of contrastive loss is highly dependent on the number of negative examples. In our model, these negative examples correspond to the centers of clusters other than the cluster an article belongs to. Therefore, it is advantageous to maintain a larger memory bank with more clusters for contrastive learning. Reducing the update frequency, which effectively retains data for more days in the memory bank, increases the likelihood of having more clusters in the memory bank. As seen in Table 4, updating the encoder less frequently, e.g. every 15 days instead of every 5 days, increases the performance.

A surprising finding was the relatively minor impact of the prompting templates on the final scores. It appears that merely having a prompt-based representation in place is sufficient to improve performance. Notably, the scores achieved by the prompt-based representation on its own are only slightly lower than those of the default model.

Furthermore, the results reveal that mean-pooling representations and the uniform sampler significantly contribute to the overall performance of the model.

In Figure 2, we compare the performance of PromptStream and SCStory with regard to B^3-F1 over the entire data stream while varying the window sizes. As the figure illustrates, our model consistently outperforms SCStory and demonstrates greater stability. However, it is noteworthy that the performance of both models tends to decline as the window size increases. This makes sense because

	NEWS14	WCEP18	WCEP19
PromptStream (default)	0.843	0.825	0.853
w/o prompt-based rep.	0.813	0.767	0.793
w/o mean rep.	0.831	0.814	0.843
w/o uniform sampler	0.842	0.807	0.842
Updating encoder			
No update	0.564	0.493	0.525
Only with first 10 days	0.833	0.818	0.842
Every 5 days	0.836	0.822	0.846
Every 15 days	0.843	0.824	0.854
Prompting templates			
(This news is about: <mask>) [title] [body]	0.845	0.823	0.851
<mask> [title] [body]	0.843	0.822	0.851
Keywords: <mask> \n [title] \n [body]	0.845	0.823	0.854
[title] \n [body] \n Keywords: <mask>	0.844	0.817	0.855

Table 4: B^3-F1 results from PromptStream ablation study with various configurations. Since the variation of the results for different seeds is very low, we report the results only for one run.

the window size signifies the time period of our interest in the stories, and if it becomes excessively large, the models may merge events with similar themes rather than detecting fine-grained events.

6.3. Qualitative Analysis

To investigate the quality of the clustering into stories we made a basic qualitative analysis of the WCEP18 dataset. The size distribution of the gold labels is rather balanced with the largest story containing 82 articles. In contrast, the largest cluster found by PromptStream comprised 603 articles, suggesting that improvements can be made to the model. The cluster contains news about North and South Korean progress on peace and denuclearization spanning over a month in April–May 2018. This could rightfully be considered a coherent story from a worldwide perspective but one could also argue that the gold labels ($n = 15$) contain individual stories within this larger event.

To further the analysis, we took the five stories that PromptStream had split into the largest number of sub-stories. These have the most negative impact on the model performance and are therefore interesting to investigate for model improvement. In Table 5 we see a summary of the stories and the number of predicted labels by PromptStream. It is fair to say that the clusters in the table do not describe coherent stories, even though the gold labels suggest so. E.g. the story with id 67432 contains general reporting on news in Calgary and British Colombia concerning e.g. local politics, accidents, and sports events. PromptStream has made a finer division of this story into such themes. The same is evident when analyzing the articles in the story with id 64606 centered around various opinions about Trump as a president from different perspectives. Its sub-stories revolve around Trump mocking the accent of the president of India, and Haitians protesting Trump's derogatory comments, which are individually captured by the model. This indicates that better datasets are needed to test the capacity of extracting granular stories. For this study, we conclude that PromptStream is performing adequately even though the scores were reduced due to limitations of the gold labeling.

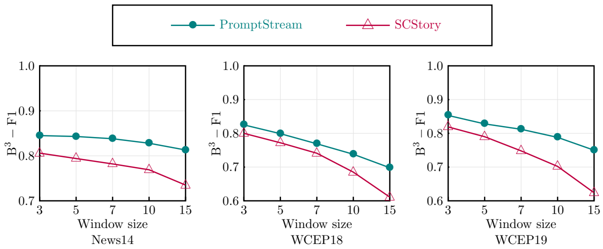


Figure 2: Comparison of PromptStream performance with that of SCStory for different window sizes, using the B^3-F1 score for the entire stream.

7. Conclusion

We introduced PromptStream as a novel approach to unsupervised online story discovery. PromptStream combines a cloze-style prompt representation with mean-pooling representation from SBERT to embed articles, ensuring a balance between the article's topic-specific information and a more general representation of the entire document. Prompt-based representations are continuously updated throughout the stream with contrastive learning using a memory of the recent confident article-story assignments. This process refines these representations and aligns them with the latest context within the news stream. In the evaluation of three labeled

story_id	n_articles	n_pred_labels	Theme
64606	54	28	Opinion about Donald Trump and his expressions **Keywords:** *Trump, president, Haiti, mocks, Modi*
67432	72	22	General reporting related to Calgary and British Colombia **Keywords:** *Calgary, family, accident, Stampeders (sports team), wildfires*
64773	51	20	American foreign policy and international politics **Keywords:** *Trump, election, Venezuela, Jerusalem, U.S.*
66490	68	18	Financial markets related to health and tech **Keywords:** *global, market, treatment, Vodafone, U.S.*
65030	66	17	U.S. foreign policy dominated by trade agreements and North Korea **Keywords:** *Trump, north, Korea, tariffs, NAFTA*

Table 5: Five stories from WCEP18 that PromptStream divided into multiple smaller stories, resulting in low performance on the evaluation. The story_id, n_articles, and n_pred_labels are the story's gold label, the story size, and the number of clusters generated by PromptStream respectively. The theme is a general description of the articles in the story which all could be argued to contain multiple sub-stories.

datasets, our model demonstrated performance improvements over the previous state of the art. Further, the subsequent ablation study highlighted the efficacy of prompt-based representation and continual training.

Data and Code Availability

The datasets are publicly available[2]. The code for PromptStream is available at https://github.com/Aha6988/PromptStream.

Acknowledgments

We would like to extend our sincere gratitude to *Frank Drewes*. His insightful comments and feedback greatly aided us in improving the quality and clarity of our work. We also express our gratitude to the reviewers for their thorough reading of the text, which has not only helped in clarifying ambiguities but also enhanced the language and conciseness of the paper.

References

Charu Aggarwal and Philip Yu. 2010. On clustering massive text and categorical data streams. *Knowledge and Information Systems*, 24:171–196.

J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. 1998. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and M. Verdejo. 2009. Amigó e, gonzalo j, artiles j et ala comparison of extrinsic clustering evaluation metrics based on formal constraints. inform retriev 12:461-486. *Information Retrieval*, 12:461–486.

Dhananjay Ashok and Zachary Chase Lipton. 2023. Promptner: Prompting for named entity recognition. *ArXiv*, abs/2305.15444.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85. Association for Computational Linguistics.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume*

[2] https://github.com/cliveyn/SCStory

1: Long Papers), pages 3816–3830. Association for Computational Linguistics.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308. Association for Computational Linguistics.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. PromptBERT: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837. Association for Computational Linguistics.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2022. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775. Association for Computational Linguistics.

Philippe Laban and Marti Hearst. 2017. newsLens: building and visualizing long-ranging news stories. In *Proceedings of the Events and Stories in the News Workshop*, pages 1–9. Association for Computational Linguistics.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Sebastiao Miranda, Arturs Znotinv, Shay B. Cohen, and Guntis Barzdins. 2018. Multilingual clustering of streaming news. pages 4535–4544. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763.

Kailash Karthik Saravanakumar, Miguel Ballesteros, Muthu Kumar Chandrasekaran, and Kathleen McKeown. 2021. Event-driven news stream clustering using entity-aware contextual embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2330–2340. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269. Association for Computational Linguistics.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12507. Association for Computational Linguistics.

Jonathan A. Silva, Elaine R. Faria, Rodrigo C. Barros, Eduardo R. Hruschka, André C. P. L. F. de Carvalho, and João Gama. 2013. Data stream clustering: A survey. *ACM Comput. Surv.*, 46(1).

Todor Staykovski, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. 2019. Dense vs. sparse representations for news stream clustering. In *Proceedings of Text2Story - 2nd Workshop on Narrative Extraction From Texts, co-located with the 41st European Conference on Information Retrieval, Text2Story@ECIR 2019, Cologne, Germany, April 14th, 2019*, volume 2342, pages 47–52. CEUR-WS.org.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. 2020. Scan: Learning to classify images without labels. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X*, pages 268–285. Springer-Verlag.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Susik Yoon, Yu Meng, Dongha Lee, and Jiawei Han. 2023. Scstory: Self-supervised and continual online story discovery. In *Proceedings of the ACM Web Conference 2023*, pages 1853–1864, New York, NY, USA. Association for Computing Machinery.

Zhenrui Yue, Huimin Zeng, Mengfei Lan, Heng Ji, and Dong Wang. 2023. Zero- and few-shot event detection via prompt-based meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7928–7943. Association for Computational Linguistics.

Senhui Zhang, Tao Ji, Wendi Ji, and Xiaoling Wang. 2022. Zero-shot event detection based on ordered contrastive learning and prompt-based prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2572–2580. Association for Computational Linguistics.