Topic discovery for news articles

Anna Wróblewska (Supervisor)
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Daniel Tytkowski
Warsaw University of Technology
01161612@pw.edu.pl
@pw.edu.pl

Abstract

In this work, we study embedding-based topic discovery on full-length news articles using a carefully scoped and balanced subset of the All the News dataset. Publisher-specific section labels are first normalized into a unified set of high-level categories to reduce heterogeneity and guide sampling.

Articles are represented using sentence-level transformer embeddings and clustered using unsupervised methods, including k-means and density-based algorithms, to identify coarse topical structure. Low-dimensional UMAP projections are employed to qualitatively assess embedding organization and cluster separability. Building on the resulting macro-topics, we incorporate discourse-level features such as sentiment, ideological bias, and framing to explore narrative variation within topical clusters.

1 Introduction

Alongside recent information overload large-scale analysis of news content has become a crucial task in natural language processing and computational social science. Topic discovery methods are commonly used to organize and explore large news corpora, typically relying on semantic similarity between documents to identify clusters corresponding to broad subjects such as politics, business, or sports. While effective at capturing high-level topical structure, such approaches often overlook narrative and framing differences that emerge within the same topic, including variations in sentiment, ideological perspective, and emphasis.

Moreover, news articles are produced by diverse publishers and follow heterogeneous editorial conventions, which further complicates unsupervised analysis. Publisher-specific metadata, such as section labels, is often noisy or inconsistent across sources, limiting its direct usefulness for large-scale analysis. As a result, purely semantic clustering may conflate articles that share topical content but differ substantially in how events are framed and interpreted.

To address these challenges, we adopt an embedding-based topic discovery framework that emphasizes careful data scoping and metadata normalization as a foundation for narrative analysis. Rather than treating discovered clusters as final outputs, we view them as coarse macro-topics that provide a structured context for examining discourse-level variation. By incorporating sentiment, bias, and framing signals within topical clusters, our approach aims to uncover narrative patterns that are not captured by topic structure alone.

2 Related Work

2.1 Topic Discovery and Topic Modeling

Topic discovery has been extensively studied in the context of probabilistic and neural topic models. Classical approaches such as Latent Dirichlet Allocation and its extensions have been widely applied to news corpora, while more recent neural topic models aim to incorporate contextualized representations and deep generative frameworks. Wu et al. (1) provide a comprehensive survey of neural topic modeling methods, highlighting advances in representation learning as well as persistent challenges related to interpretability, evaluation, and scalability in large and heterogeneous datasets. These limitations motivate alternative approaches to topic discovery that rely on pretrained embeddings and clustering rather than explicit topic model training.

Recent work has explored guided and semi-

supervised topic discovery as a way to improve topic coherence and interpretability. Zhang et al. (2) propose a seed-guided framework that integrates multiple types of contextual information, including word-level and document-level signals, to steer topic formation. While effective, such methods require model training and predefined supervision. In contrast, our work focuses on fully unsupervised, embedding-based clustering and explores lightweight forms of guidance at the analysis stage rather than during model optimization.

Recent advances have further shifted topic discovery toward embedding-based representations derived from large pretrained language models, enabling clustering-driven approaches such as BERTopic (3) and Top2Vec (4). While large language models also support prompt-based and generative topic modeling, embedding-based methods remain widely adopted due to their scalability, interpretability, and suitability for fully unsupervised analysis of large text corpora. Our work follows this line of research by leveraging semantic embeddings as a foundation for topic discovery, while extending their use toward narrative analysis through discourse-level feature distributions rather than topic keyword extraction alone.

2.2 Discourse-Level Signals in News Analysis

Beyond topical structure, prior research has emphasized the importance of discourse-level features such as sentiment, subjectivity, and framing in understanding news content. Augustyniak et al. (5) introduce a large multilingual benchmark demonstrating that sentiment in news is multifaceted and cannot be adequately captured by simple polarity labels. Their findings motivate the incorporation of richer discourse signals when analyzing news articles, particularly in the context of narrative variation within shared topics.

Understanding narrative and framing differences has also become increasingly relevant for downstream applications such as misinformation detection and claim verification. Recent shared tasks, including SemEval 2025 on fact-checked claim retrieval (6), highlight the need for structured representations of news content that go beyond topical similarity. While our work does not address fact-checking directly, it aims to provide an exploratory foundation for analyzing narrative structure within topical clusters, which may support such downstream tasks.

2.3 Positioning of This Work

In contrast to neural topic models and guided topic discovery approaches, we adopt a lightweight, embedding-based clustering framework that emphasizes data curation, metadata normalization, and qualitative analysis. By combining semantic embeddings with sentiment, bias, and framing features, our approach bridges topic discovery and narrative analysis without introducing additional model complexity. This positions our work as complementary to existing topic modeling methods, focusing on exploratory analysis and interpretability in large-scale news corpora.

3 Data

We conduct our experiments using articles from the All the News corpus, a large-scale collection of English-language news articles gathered from a diverse set of publishers. The dataset contains full-length articles along with associated metadata such as publication source and section labels. To preserve discourse-level information necessary for narrative analysis, we focus exclusively on full article texts rather than short summaries or headlines.

Due to the size and heterogeneity of the corpus, we apply a series of preprocessing and sampling steps. Articles with insufficient textual content are filtered out using a minimum length threshold, ensuring that retained documents contain enough material for semantic and discourse-level analysis. Publisher-specific section labels, which are often inconsistent or noisy, are normalized into a small set of unified categories using keyword-based matching. This normalization enables controlled sampling and facilitates comparison between semantic clusters and editorial metadata.

From the filtered corpus, we construct a balanced subset of 2,000 articles by stratified sampling across unified section categories, excluding a heterogeneous “Other” group. This sampling strategy reduces topical imbalance while retaining diversity across major news domains such as business, politics, world news, sports, and technology. The resulting dataset provides a manageable yet representative foundation for embedding-based topic discovery and subsequent narrative analysis.

Publication	# Articles
Reuters	840,094
The New York Times	252,259
CNBC	238,096
The Hill	208,411
People	136,488
CNN	127,602
Refinery 29	111,433
Vice	101,137
Mashable	94,107
Business Insider	57,953

Table 1: Distribution of articles across major publications in the All the News corpus. Top 10 contributors.

3.1 Dataset creation

- We restrict the corpus to articles published in 2019 to ensure temporal consistency and reduce long-term distributional shifts in news coverage.
- Articles with insufficient textual content are filtered out by retaining only documents with a minimum length of 1,500 characters, ensuring the presence of discourse-level signals necessary for narrative analysis.
- We analyze publisher metadata to identify sources that provide section labels and examine the consistency of these labels across publications.
- Publisher-specific section labels are normalized into a unified set of high-level categories using keyword-based mapping, reducing noise and heterogeneity in the metadata.
- From the filtered corpus, we construct a balanced subset of 2,000 articles by stratified sampling across unified section categories, specifically Business, Politics, World, Sports, Tech, and Movies.

To represent the semantic content of news articles, we compute dense vector embeddings using a pretrained sentence-level transformer model. Each article is encoded as a fixed-length embedding capturing its overall semantic meaning, which serves as the basis for subsequent clustering and visualization. The embedding model is applied to the article text after preprocessing, and the resulting representations are L2-normalized to ensure stable distance-based comparisons. These embeddings enable efficient unsupervised topic discovery using clustering algorithms such as k-means and density-based methods, while preserving semantic similarity between documents.

In addition to semantic embeddings used for topic discovery, we extract a set of discourse-level features to capture narrative variation within topical clusters. These features include sentiment polarity, subjectivity, ideological bias, and framing. Sentiment is used to characterize the overall emotional tone of articles, while subjectivity captures the extent to which content is presented as opinionated versus factual. Ideological bias is included as a proxy signal for editorial perspective, and framing is estimated using a zero-shot classification approach with a predefined set of high-level news frames. All discourse-level features are computed on truncated article excerpts to ensure compatibility with pretrained transformer models. These signals are not treated as ground-truth annotations, but are analyzed comparatively to identify narrative patterns within and across topical clusters.

Task	Model	Purpose
Semantic Embedding	Sentence Transformer	Document representation for clustering
Sentiment Analysis	twitter-roberta-base-sentiment	Estimate article-level sentiment polarity
Subjectivity Detection	mDeBERTa-v3 Subjectivity	Identify subjective vs. objective content
Bias Detection	UnBIAS Classifier	Capture proxy signals of ideological bias
Framing Classification	DeBERTa-v3 NLI (zero-shot)	Assign high-level news frames
Dimensionality Reduction	UMAP	Visualization of embedding space

Table 2: Overview of models and methods used in the study.

3.2 EDA

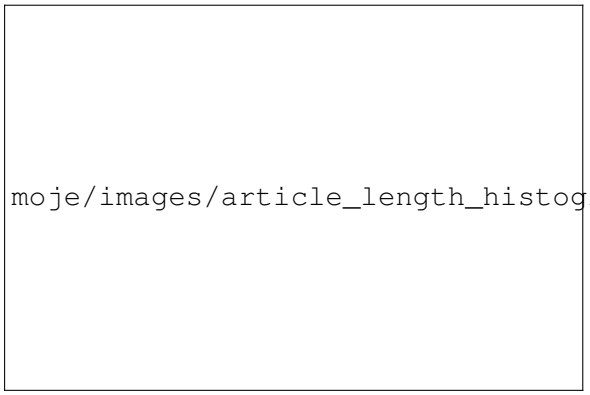

Figure 1: Lengths distribution on prepared dataset

Unified Section	# Articles
Business	2,000
Movies	2,000
Politics	2,000
Sports	2,000
Tech	2,000
World	2,000

Table 3: Distribution of articles in the final balanced dataset after stratified sampling by unified section.

Publication	# Articles
Reuters	3,879
The New York Times	2,358
CNN	2,302
People	1,775
CNBC	1,509
Economist	160
Wired	10
Gizmodo	4
The Verge	3

Table 4: Publication distribution in the final sampled dataset.

Table 3 summarizes the balanced composition of the final dataset across unified section categories.

Table 4 shows the distribution of articles across publishers in the final sampled dataset. The sampled dataset remains dominated by a small number of large publishers, reflecting the underlying distribution of the original corpus.

Figure 2: UMAP visualisation of semantic embeddings in 2D space

Sentiment Label	# Articles
Positive (LABEL_1)	6,787
Neutral (LABEL_0)	2,697
Negative (LABEL_2)	2,516

Table 5: Distribution of sentiment labels in the final sampled dataset.

Sentiment polarity is used to characterize the overall emotional tone of news articles and is estimated using a pretrained transformer-based classifier. Table 5 shows the distribution of sentiment labels in the sampled dataset, with a majority of articles assigned a positive label, followed by neutral and negative categories. This distribution reflects the prevalence of neutral-to-positive reporting styles in mainstream news coverage, as well as the model's sensitivity to mild evaluative language. In this study, sentiment is treated as a relative signal and analyzed comparatively within topical clusters to identify narrative differences rather than as an absolute measure of affect.

Bias Label	# Articles
Highly Biased	11,930
Slightly Biased	70

Table 6: Distribution of bias labels in the final sampled dataset.

Ideological bias is estimated using a pretrained media bias classifier and treated as a proxy signal for editorial perspective rather than a ground-truth annotation. As shown in Table 6, the majority of articles are classified as highly biased, reflecting the broad definition of bias adopted by the model,

which captures subtle linguistic and framing cues commonly present in news reporting. The small number of articles labeled as slightly biased suggests that the classifier is conservative in assigning low-bias labels. Consequently, bias scores are analyzed comparatively within topical clusters rather than interpreted as absolute measures of ideological bias.

Subjectivity Label	# Articles
Objective (LABEL_0)	10,957
Subjective (LABEL_1)	1,043

Table 7: Distribution of subjectivity labels in the final sampled dataset.

Subjectivity is estimated using a pretrained transformer-based classifier and reflects whether an article is presented in a primarily objective or subjective manner. As shown in Table 7, the majority of articles are classified as objective, which is consistent with the journalistic norm of factual reporting in mainstream news outlets. Articles labeled as subjective typically correspond to opinionated language, evaluative statements, or interpretive framing. In this work, subjectivity is used as a comparative signal to examine differences in narrative style within and across topical clusters, rather than as a definitive indicator of journalistic quality.

Frame	# Articles
Conflict	6,984
Economic consequences	2,103
Human interest	1,967
Neutral factual reporting	700
Policy and governance	215
Morality	31

Table 8: Distribution of framing labels assigned by zero-shot classification in the final sampled dataset.

Framing is estimated using a zero-shot natural language inference model with a predefined set of high-level news frames. As shown in Table 8, conflict-oriented framing is the most prevalent, followed by economic consequences and human-interest frames. This distribution reflects the tendency of news reporting to emphasize tension, disagreement, and impact on individuals, particularly in coverage of politics, business, and international affairs. Frames such as policy and governance and morality occur less frequently, suggesting that explicit normative or institutional framing is comparatively rare in the sampled corpus. In this study, framing labels are interpreted as relative narrative signals rather than definitive categorizations.

3.3 Time/memory estimation

Task	Batches	Runtime
Semantic Embedding	188	∼15 s
Sentiment Analysis	188	∼2 min 16 s
Bias Detection	188	∼2 min 17 s
Subjectivity Detection	188	∼6 min 14 s
Framing Classification	1500	∼18 min 40 s

Table 9: Approximate runtime for feature extraction tasks on the final dataset using a single NVIDIA RTX 3070 GPU.

Processing Stage	Estimated GPU Memory
Semantic Embedding	∼1.5–2.0 GB
Sentiment Analysis	∼1.2–1.5 GB
Bias Detection	∼1.2–1.5 GB
Subjectivity Detection	∼2.0–2.5 GB
Framing Classification (zero-shot)	∼3.5–4.5 GB

Table 10: Estimated GPU memory usage per processing stage on a single NVIDIA RTX 3070 (8 GB VRAM).

All experiments were conducted on a single NVIDIA RTX 3070 GPU with 8,GB of VRAM. Feature extraction was performed in batches to balance computational efficiency and GPU memory constraints. Semantic embedding computation completed in under 20 seconds, while sentiment, bias, and subjectivity classification required several minutes per feature. Zero-shot framing classification was the most computationally expensive and memory-intensive stage, reflecting the cost of natural language inference over multiple candidate labels. Estimated GPU memory usage ranged from approximately 1–2,GB for lightweight classifiers to up to 4,GB for framing analysis, and no out-of-memory errors were encountered during execution.

3.4 Preprocessed dataset

The preprocessed dataset used in this study, containing semantic embeddings and numerically encoded discourse features without raw article text, is publicly available on the Hugging Face Hub at: `https://huggingface.co/datasets/Tytanito/`

`news-narratives-embeddings`. The dataset is released in an experiment-ready format to facilitate reproducibility and downstream analysis.

3.5 Limitations

A limitation of this study is the reliance on pretrained classifiers for sentiment, bias, subjectivity, and framing, which are not specifically trained on the target news corpus. As a result, these discourse-level features should be interpreted as proxy signals rather than ground-truth annotations. In addition, zero-shot framing classification is computationally expensive and limits scalability to larger datasets without additional optimization. Finally, while embedding-based clustering captures coarse topical structure, it does not explicitly model temporal dynamics or fine-grained narrative evolution, which we leave for future work.

4 Planned experiments

Experiment	Description
Embedding Space Analysis	Analyze semantic embeddings using UMAP to inspect global structure and relationships between articles.
Topic Discovery	Apply unsupervised clustering methods to identify coarse-grained topics based on semantic similarity.
Section-Level Comparison	Compare embedding distributions and discourse features across unified news sections.
Narrative Feature Analysis	Analyze sentiment, subjectivity, bias, and framing distributions within and across topics.
Cluster Profiling	Aggregate numeric discourse features to construct narrative profiles for discovered clusters.
Qualitative Inspection	Perform manual inspection of representative articles to assess narrative coherence and framing patterns.

Table 11: Overview of planned experiments conducted in this study.

5 Experiments and results

This section presents a series of experiments designed to analyze semantic topic structure and narrative variation in a large news corpus. We first evaluate whether semantic embeddings recover high-level editorial sections using unsupervised clustering. We then analyze discourse-level features within the discovered structure.

5.1 Evaluation Metrics

To evaluate the quality of unsupervised topic discovery, we employ clustering evaluation metrics that compare the discovered clusters with ground-truth section labels in a permutation-invariant manner. Importantly, the section labels are used exclusively for evaluation and are not incorporated into the clustering process.

Adjusted Rand Index (ARI). The Adjusted Rand Index measures the similarity between two partitions by considering all pairs of samples and counting pairs that are assigned consistently in both clusterings. Unlike the Rand Index, ARI corrects for chance agreement. It is defined as:

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}, \tag{1}$$

where n_{ij} denotes the number of samples assigned to cluster i and ground-truth class j, $a_i = \sum_j n_{ij}$, $b_j = \sum_i n_{ij}$, and n is the total number of samples. ARI values range from 0 for random agreement to 1 for perfect alignment.

Normalized Mutual Information (NMI). Normalized Mutual Information measures the amount of shared information between cluster assignments and ground-truth labels. It is defined as:

$$\mathrm{NMI}(U, V) = \frac{2\, I(U;V)}{H(U) + H(V)}, \tag{2}$$

where $I(U;V)$ denotes the mutual information between the clustering U and the ground-truth labeling V, and $H(\cdot)$ denotes entropy. NMI takes values in the range $[0, 1]$, with higher values indicating stronger agreement between the two partitions.

Cluster Purity. Cluster purity is a descriptive metric that measures the extent to which each cluster contains samples from a single dominant class. It is computed as:

$$\mathrm{Purity} = \frac{1}{n}\sum_k \max_j |C_k \cap L_j|, \tag{3}$$

where C_k denotes the set of samples in cluster k and L_j denotes the set of samples belonging to

ground-truth class j. While purity is intuitive, it is sensitive to the number of clusters and is therefore reported only as a complementary measure.

Silhouette Score. To assess the intrinsic quality of the clustering independent of ground-truth labels, we additionally report the silhouette score, which measures cluster compactness and separation. For a given sample i, the silhouette coefficient is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{4}$$

where $a(i)$ is the mean distance between i and all other points in the same cluster, and $b(i)$ is the minimum mean distance between i and points in the nearest neighboring cluster. The overall silhouette score is obtained by averaging $s(i)$ over all samples.

Jennsen-Shannon divergence To compare narrative feature distributions across clusters, we employ Jensen–Shannon (JS) divergence, a symmetric and bounded measure of dissimilarity between probability distributions. Given two discrete distributions P and Q, the JS divergence is defined as

$$\mathrm{JS}(P \,\|\, Q) = \frac{1}{2}\mathrm{KL}(P \,\|\, M) + \frac{1}{2}\mathrm{KL}(Q \,\|\, M), \quad \text{where} \quad M = \frac{1}{2}(P+Q),$$

and $\mathrm{KL}(\cdot \,\|\, \cdot)$ denotes the Kullback–Leibler divergence. Unlike KL divergence, JS divergence is symmetric and always finite, making it suitable for comparing empirical distributions. In this work, JS divergence is used to quantify differences in framing, sentiment, subjectivity, and bias distributions between clusters, with higher values indicating greater narrative separation.

5.2 Density-Based Clustering with HDBSCAN

Goal. The goal of this experiment is to assess whether density-based clustering can identify highly coherent local structures in the semantic embedding space without imposing a predefined number of clusters. Unlike k-means, which enforces a global partitioning, this experiment explores whether dense narrative pockets emerge naturally in the embedding space of news articles.

Setup. We apply HDBSCAN to the L2-normalized semantic embeddings using a fixed minimum cluster size to control the granularity of detected structures. HDBSCAN groups articles based on local density and labels low-density regions as noise or part of a large diffuse cluster. The resulting cluster assignments are analyzed qualitatively using UMAP visualizations to examine the presence of compact, high-density regions and to contrast density-based clustering behavior with centroid-based methods.

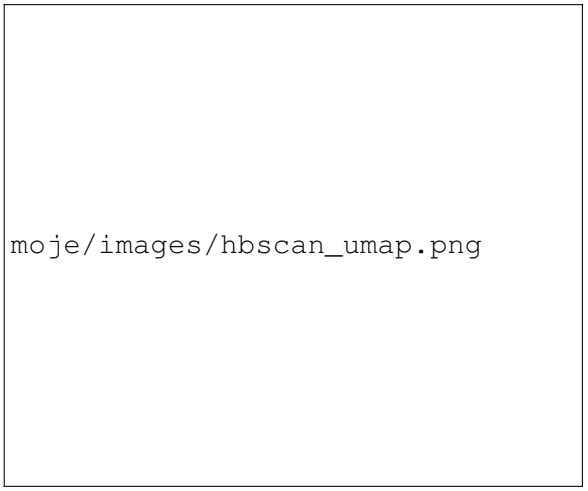

Figure 3: UMAP projection of semantic embeddings colored by HDBSCAN cluster assignments. HDBSCAN identifies dense local regions while assigning most points to a large diffuse cluster.

Figure 3 presents a UMAP visualization of semantic embeddings colored by HDBSCAN cluster assignments. In contrast to k-means, HDBSCAN is designed to identify dense local regions in the embedding space rather than enforce a global partitioning into a fixed number of clusters. As a result, the algorithm isolates a small number of compact clusters while grouping the majority of articles into a large, low-density region. This behavior reflects the continuous and overlapping nature of news semantics, where only a subset of articles forms highly coherent narrative pockets. Consequently, HDBSCAN is used here as a diagnostic tool to highlight locally dense structures rather than as a primary topic discovery method.

5.3 Semantic Topic Discovery with k-Means

Goal. The goal of this experiment is to evaluate whether semantic embeddings of news articles capture high-level topical structure corresponding to editorial sections. Specifically, we assess whether unsupervised clustering recovers the unified section categories without using label information during training.

Setup. Each article is represented by a dense semantic embedding computed using a pretrained sentence-level transformer model. Semantic embeddings are L2-normalized to preserve cosine similarity structure. We apply k-means clustering to the embeddings with the number of clusters set to $k = 6$, matching the number of unified section categories in the dataset. Although the clustering procedure is fully unsupervised, the section labels are used post hoc for evaluation purposes only.

Figure 4: UMAP visualisation of semantic embeddings in 2D space, coloured by groundtruth section

Figure 5: UMAP visualisation of semantic embeddings in 2D space, coloured by k-means cluster

Figures 4 and 5 show UMAP projections colored by section labels and k-means clusters, respectively. The similar spatial organization of clusters across the two plots indicates substantial alignment between semantic clustering and editorial sections.

Clustering quality is evaluated using Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), which quantify the alignment between cluster assignments and section labels while remaining invariant to label permutations.

Metric	**Value**
Adjusted Rand Index (ARI)	0.643
Normalized Mutual Information (NMI)	0.648
Cluster Purity	0.822
Silhouette Score	0.048

Table 12: Clustering evaluation metrics for Experiment 1 (k-means, $k = 6$).

Table 12 summarizes the quantitative evaluation of k-means clustering with $k = 6$. The clustering achieves an Adjusted Rand Index of 0.643 and a Normalized Mutual Information score of 0.648, indicating substantial alignment between the discovered clusters and the unified editorial section labels. The high cluster purity of 0.822 suggests that most clusters are dominated by a single section, while still allowing for cross-sectional overlap. In contrast, the relatively low silhouette score of 0.048 reflects the gradual transitions and semantic overlap inherent in real-world news content, rather than sharply separated clusters. Together, these results indicate that semantic embeddings capture broad topical structure while preserving narrative continuity across sections.

Figure 6: Silhouette scores for k-means clustering across different values of k, computed on semantic embeddings.

Figure 6 shows silhouette scores for k-means

clustering across a range of values of k. The scores remain uniformly low, reflecting the overlapping and continuous semantic structure of news articles. Although the highest silhouette value is observed at $k = 2$, this corresponds to an overly coarse partitioning that fails to capture editorial structure. A local maximum is observed at $k = 6$, which is consistent with the number of unified section categories and aligns with the strong external evaluation results. Consequently, silhouette analysis is used as a supporting diagnostic rather than a primary criterion for selecting k.

Figure 7: UMAP visualisation of semantic embeddings in 2D space, coloured by k-means cluster

5.4 Narrative-Aware Topic Discovery

Goal. The goal of this experiment is to examine whether incorporating narrative-level features alters or refines the topical structure obtained from semantic embeddings alone. In particular, we investigate whether sentiment, subjectivity, bias, and framing information introduce cross-sectional relationships that are not captured by purely semantic similarity.

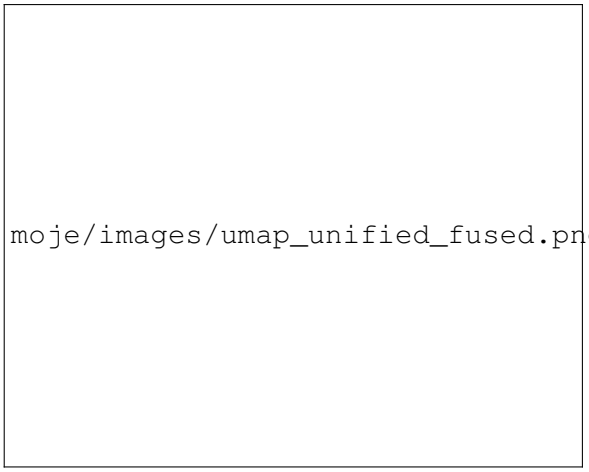

Figure 8: UMAP visualisation of semantic embeddings in 2D space, coloured by groundtruth section

Setup. Each article is represented using a fused feature representation that combines its semantic embedding with numerically encoded narrative features. Semantic embeddings are L2-normalized to preserve cosine similarity structure, while narrative features are standardized to zero mean and unit variance. The two representations are concatenated using a controlled feature fusion strategy, where narrative features are down-weighted by a scaling factor $\alpha = 0.2$ to prevent them from dominating the semantic space. K-means clustering with $k = 6$ is then applied to the fused representation, and the resulting clusters are analyzed and compared to those obtained using semantic embeddings alone.

As seen in Figure 7 2D projections are changed, however the overall structure remains more or less similar. The Business-Politics-World-Tech topics still remain close to each other, while Movies and Sports distant away.

Metric	Value
Adjusted Rand Index (ARI)	0.690
Normalized Mutual Information (NMI)	0.680

Table 13: Structural similarity between semantic clustering and narrative-aware clustering.

Metric	Value
Adjusted Rand Index (ARI)	0.517
Normalized Mutual Information (NMI)	0.527
Cluster Purity	0.822
Silhouette Score	0.048

Table 14: Evaluation of narrative-aware clustering with respect to unified section labels.

The narrative-aware clustering exhibits substantial structural consistency with the embedding-based clustering, achieving an Adjusted Rand Index of 0.690 (Table 13) between the two cluster assignments. There is strong similarity, however also significant impact is observed. This indicates that the incorporation of narrative features refines the existing semantic topics rather than replacing them entirely. At the same time, the alignment with editorial section labels decreases, with the Adjusted Rand Index dropping to 0.517 (Table 14) when comparing the fused clustering to the unified sections. This reduction suggests that narrative-aware clustering introduces cross-sectional groupings driven by shared narrative characteristics such as sentiment, bias, and framing, thereby revealing patterns that are not captured by semantic similarity alone.

5.5 Experiment 4: Intra-Topic Narrative Discovery

Goal. The goal of this experiment is to investigate narrative variation within a single topical category, independent of broad semantic differences between sections. By focusing on one unified section exhibiting high narrative diversity, we aim to assess whether multiple distinct narratives coexist within the same topic and to characterize these narratives using discourse-level features.

Setup. We first compute the variance of narrative features (sentiment, subjectivity, bias, and framing) across unified sections and select the section with the highest average narrative variance. Restricting the analysis to this section controls for topical variation and isolates narrative effects. Articles within the selected section are represented using their semantic embeddings, and k-means clustering is applied with a small number of clusters to identify latent subgroups. Narrative features are then analyzed within each sub-cluster to construct narrative profiles and assess intra-topic narrative separation.

Section	Mean Var.	Frame Entropy
Business	0.197	1.096
Movies	0.372	1.069
Politics	0.150	0.813
Sports	0.286	0.999
Tech	0.267	1.278
World	0.117	0.818

Table 15: Variance of numeric narrative features and framing diversity across unified sections.

Based on the narrative statistics reported in Table 15, we select the Movies section for intra-topic narrative analysis. This section exhibits the highest mean variance across numeric narrative features, indicating substantial variation in sentiment and subjectivity, while also maintaining high framing entropy. The combination of elevated numeric variance and diverse framing strategies suggests the presence of multiple coexisting narratives within the same topical category, making Movies a suitable candidate for detailed narrative analysis.

Cluster Pair	JS Divergence
0 – 1	0.089
0 – 2	0.116
1 – 2	0.173

Table 16: Pairwise Jensen–Shannon divergence between framing distributions for intra-topic clusters in the Movies section.

Cluster Pair	JS Divergence
0 – 1	0.051
0 – 2	0.089
1 – 2	0.117

Table 17: Pairwise Jensen–Shannon divergence between sentiment distributions for intra-topic clusters in the Movies section.

Cluster Pair	JS Divergence
0 – 1	0.061
0 – 2	0.105
1 – 2	0.165

Table 18: Pairwise Jensen–Shannon divergence between subjectivity distributions for intra-topic clusters in the Movies section.

Cluster Pair	JS Divergence
0 – 1	0.025
0 – 2	0.007
1 – 2	0.018

Table 19: Pairwise Jensen–Shannon divergence between bias distributions for intra-topic clusters in the Movies section.

Tables 16–19 summarize pairwise Jensen–Shannon divergence between narrative feature distributions across intra-topic clusters in the Movies section. Framing exhibits the strongest separation between clusters, with the largest divergence observed between clusters 1 and 2, indicating distinct framing strategies within a shared topical context. Sentiment and subjectivity show moderate distributional differences, suggesting secondary narrative variation in tone and discourse style, while bias displays minimal divergence across clusters. Overall, these results demonstrate that narrative differentiation within this topic is driven primarily by framing rather than sentiment polarity or bias.

6 Reproducibility

To ensure reproducibility of the experimental results, all stochastic components, including clustering initialization and dimensionality reduction, are executed with fixed random seeds. While certain operations (e.g., GPU-based computation and approximate nearest neighbor search) may introduce minor non-determinism, these effects do not materially impact the qualitative or quantitative conclusions of the study. All experiments were repeated under consistent settings, yielding stable clustering structures and narrative patterns.

7 Conclusion

In this work, we investigated embedding-based topic discovery for large-scale news articles, with a particular focus on uncovering narrative variation beyond coarse topical structure. Using semantic embeddings and unsupervised clustering, we showed that high-level editorial sections can be recovered without supervision, indicating that pretrained language models capture meaningful topical organization in news content. At the same time, intrinsic clustering metrics revealed a continuous and overlapping semantic structure, reflecting the gradual transitions characteristic of real-world news discourse.

Building on this macro-level structure, we incorporated discourse-level features such as sentiment, subjectivity, bias, and framing to explore narrative patterns within and across topics. Our results demonstrate that augmenting semantic representations with narrative signals refines topical clusters and introduces cross-sectional relationships that are not captured by semantic similarity alone. In particular, intra-topic analysis within the Movies section revealed the coexistence of multiple narratives distinguished primarily by framing strategies rather than sentiment or bias. Distributional comparisons using Jensen–Shannon divergence confirmed that framing is the dominant axis of narrative separation within this topic.

Overall, this study highlights the importance of treating topic discovery as a foundation for narrative analysis rather than an end goal. By combining careful data curation, embedding-based clustering, and comparative discourse analysis, our approach provides an interpretable and scalable framework for exploring narrative structure in large news corpora.

8 Future Work

Several directions for future research emerge from this study. First, extending the analysis to additional topical sections and longer temporal spans would enable the study of narrative dynamics over time, including the emergence, evolution, and decay of narratives in response to real-world events. Incorporating temporal modeling could further reveal how framing strategies shift across news cycles.

Second, while this work relies on pretrained classifiers to extract discourse-level features, future work could explore domain-adapted or jointly trained models to improve the reliability and granularity of narrative signals, particularly for framing and ideological bias. Human-in-the-loop annotation or targeted evaluation on curated subsets could further validate these signals.

Finally, narrative-aware clustering could be integrated more tightly with downstream tasks such as misinformation detection, claim verification, or media bias analysis. By explicitly modeling narrative structure within topics, future systems may better capture subtle differences in perspective and emphasis that are critical for understanding information ecosystems at scale.

9 Rebuttal

Overall literature review is sparse and lacks more detailed presentation.

The literature review has been expanded and refined. An additional paragraph has been added to explicitly position this work with respect to existing approaches in topic discovery and narrative analysis, clarifying how the proposed method relates to and extends prior research.

The dataset is not clearly defined. FEVER is mentioned, but it is unclear whether it will be used. It is stated only in the evaluation that it will be used.

The dataset used in this study is now clearly defined in the Dataset section. Although FEVER was mentioned in an earlier version, it is not used in the experiments reported in this work. All analyses are conducted exclusively on the described news article dataset, and references to FEVER have been removed to avoid confusion.

The evaluation metrics should be explained (formula, word explanation, etc.), e.g., "topic coherence", "factuality", "topic sta- bility". The claim "recent advancements have shifted toward embedding-based meth- ods like BERTopic and Top2Vec" is rela- tively outdated given the advances in LLMs.

The evaluation metrics used in this study are now explicitly defined in the Evaluation section, including formal definitions and intuitive explanations where applicable. In particular, metrics used to assess clustering quality and narrative separation are described in detail. Additionally, the claim regarding embedding-based topic models has been revised to reflect recent advances in large language models. The discussion now clarifies that while LLMs enable new approaches to topic modeling, this work focuses on embedding-based methods due to their scalability and suitability for unsupervised analysis of large news corpora.

10 Work contribution

Name	Contribution	Workload
Daniel	100%	30h

Table 20: Author contribution and workload.

References

[1] X. Wu, T. H. Nguyen, and A. T. Luu, "A survey on neural topic models: Methods, applications, and challenges," *Artificial Intelligence Review*, vol. 57, no. 18, 2024.

[2] Y. Zhang, Y. Zhang, M. Michalski, Y. Jiang, Y. Meng, and J. Han, "Effective seed-guided topic discovery by integrating multiple types of contexts," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 429–437.

[3] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[4] D. Angelov, "Top2vec: Distributed representations of topics," *arXiv preprint arXiv:2008.09470*, 2020.

[5] L. Augustyniak, S. Woźniak, M. Gruza, P. Gramacki, K. Rajda, M. Morzy, and T. Kajdanowicz, "Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark," *Advances in Neural Information Processing Systems*, vol. 36, pp. 38 586–38 610, 2023.

[6] SemEval Organizers, "Semeval-2025 task: Fact-checked claim retrieval," https://disai.eu/semeval-2025/, 2025, accessed: 2026.