Event-Driven Topic Shifts in The Guardian: A Comparative Analysis of BERTopic and LDA around Queen Elizabeth II's Passing

Liuxi Mei
Department of Computer and Information Science
Linköping University
liume102@student.liu.se

Abstract

We analyze how The Guardian's news agenda shifted around Queen Elizabeth II's passing, using topic modelling on articles from 2020–2025. To isolate indirect effects, we exclude direct event coverage via buffered windows and keyword filtering. We compare Latent Dirichlet Allocation (LDA) and BERTopic, employing a novel semantic alignment strategy to track evolving themes across independent models. Results reveal a significant "agenda reset": diffuse reporting declined while specific narratives—notably Economic Policy and Political Leadership—consolidated. BERTopic provided superior granularity in capturing these shifts compared to LDA. Our findings demonstrate that major events drive substantial structural changes in media agendas, with established themes briefly displaced before returning with altered intensity.

1 Introduction

Major events can dramatically reshape media agendas (McCombs and Shaw, 1972), but measuring these shifts requires careful exclusion of event-dominated coverage. This study focuses on The Guardian's reporting before and after Queen Elizabeth II's passing on September 8, 2022, a landmark event in UK history. We use topic modelling to quantify agenda changes, implementing buffered time windows and keyword-based filtering to remove direct event reports and obituaries. Our research questions are: (1) How do news topics shift around the event after excluding direct coverage? (2) How do LDA and BERTopic compare in capturing these shifts? (3) Are the observed agenda changes statistically significant?

Our contributions are: (1) A robust pipeline for event-exclusion in media agenda analysis; (2) Comparative evaluation of LDA and BERTopic on filtered Guardian news; (3) Statistical and temporal analysis confirming significant agenda shifts even after excluding direct event coverage.

Metric	Pre-event	Post-event
#documents	20,000	20,000
Avg. Tokens	11.19	12.30
BERTopic (C_v)	0.3950	0.4455
LDA (C_{UMass})	-5.2734	-4.6545

Table 1: Dataset statistics and model coherence metrics. Note: BERTopic uses C_v (0–1), while LDA uses C_{UMass} (typically negative).

2 Background and Related Work

Topic modelling aims to uncover latent semantic structure in document collections by representing each document as a mixture of topics and each topic as a distribution over words. LDA (Blei et al., 2003) remains one of the most widely used topic models due to its conceptual simplicity and efficient implementations. It has been applied extensively to news corpora to explore media agendas, political framing, and temporal trends (DiMaggio et al., 2013). However, several studies have pointed out that LDA can produce incoherent topics, especially when documents are short or vocabulary is noisy, and that it may merge semantically distinct themes into a single topic (Newman et al., 2010).

To address some of these issues, recent work has explored neural and embedding-based topic models that leverage pre-trained language representations. BERTopic (Grootendorst, 2022) is a prominent example that combines transformer-based sentence embeddings, clustering, and class-based TF–IDF to derive interpretable topics. Prior work has shown that such approaches often yield higher topic coherence and more fine-grained distinctions than LDA, particularly on modern text genres such as news, reviews, and social media posts.

Figure 1: Top-5 topic word clouds for BERTopic (top panel) and LDA (bottom panel). Each row shows pre-event (left sub-column) and post-event (right sub-column) wordclouds for the same topic index.

Dynamic topic models (Blei and Lafferty, 2006) extend static topic models by explicitly modelling how topic distributions change over time. While fully generative dynamic models can be complex to implement and infer, a simpler alternative is to train separate topic models on different time slices and analyse how topic prevalence and content vary across periods (Hall et al., 2008; Hoyle et al., 2014). This pragmatic strategy has been adopted in several studies of news and social media data to study temporal trends without requiring sophisticated temporal priors.

Our work is closest in spirit to research that compares classical probabilistic topic models with neural or embedding-based approaches on real-world corpora, and to studies that apply topic modelling to news archives for temporal analysis. Unlike most prior work, which typically focuses either on static topic quality or on dynamic trends within a single modelling framework, we combine a comparative evaluation of LDA and BERTopic with an explicit analysis of topic evolution over time on The Guardian dataset used in this study.

3 Data

We collected The Guardian's English news articles from 2020 to 2025 using a custom-developed Python scraper (`guardian_news_scraper.py`) that interfaces directly with The Guardian Open Platform API. This approach ensured high-fidelity data retrieval, capturing essential metadata including UTC-standardized publication dates, titles, and section information. The study focuses on the period around Queen Elizabeth II's passing on September 8, 2022. To avoid event-dominated coverage, we exclude articles within a 30-day buffer before and after the event, and filter out texts containing keywords such as "obituary", "death", "tribute", "funeral", and direct references to the Queen or monarchy. Minimal preprocessing is applied: lowercasing, stopword removal, and basic cleaning. The final dataset enables robust analysis of agenda shifts while minimizing confounding from direct event reporting.

We constructed a balanced dataset by randomly sampling 20,000 articles for each period (Pre and Post). This equal sampling size was chosen to (1) prevent volume bias in topic prevalence comparisons, ensuring that observed shifts reflect genuine proportional changes rather than dataset imbalances, and (2) maintain computational feasibility for the resource-intensive BERTopic embeddings.

The pre-event period comprises articles published prior to the exclusion buffer, while the post-event period comprises those published after. The 'filtered' designation indicates that direct event-related articles have been excluded. #Documents denotes the number of articles in each split, and Avg. Tokens represents the average number of to-

kens per article after preprocessing.

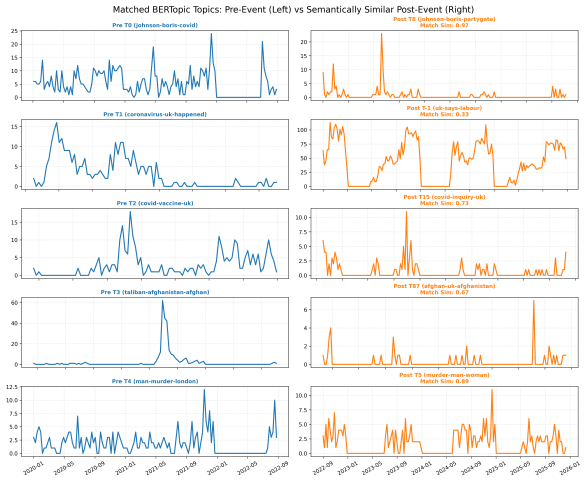

Figure 2: Weekly topic prevalence for top aligned BERTopic themes.

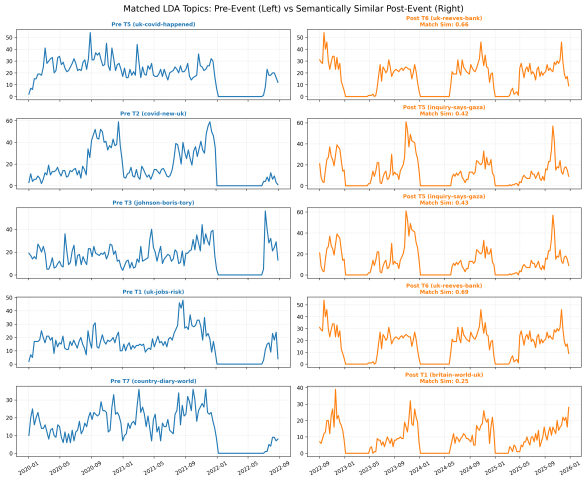

Figure 3: Weekly topic prevalence for top aligned LDA themes.

4 Method

Our pipeline consists of: (1) Data cleaning and event-exclusion filtering; (2) Topic modelling with LDA and BERTopic; (3) Quantitative and qualitative comparison; (4) Statistical and temporal analysis.

4.1 Event-Exclusion Filtering

We exclude all articles within 30 days before and after the event date, and remove those containing event-related keywords. This ensures that the analysis focuses on indirect agenda shifts rather than direct event coverage.

4.2 Topic Modelling

We apply LDA (bag-of-words, tuned topic number) and BERTopic (transformer-based embeddings, clustering, class-based TF–IDF) to the filtered corpus. Importantly, we train separate models for the pre-event and post-event periods. This approach avoids enforcing a static vocabulary or topic structure, allowing the models to capture new themes that may emerge after the event.

4.3 Temporal Topic Alignment

Since independent models produce different topic IDs and structures for the pre- and post-event periods, we implement a semantic matching strategy to enable direct comparison.

- **LDA Alignment:** We compute the cosine similarity between the topic-word distributions (β vectors) of the Pre-LDA and Post-LDA models. For each prominent Pre-event topic, we identify the Post-event topic with the highest similarity score.

- **BERTopic Alignment:** We generate embedding vectors for each topic by aggregating the embeddings of their constituent documents (or using the c-TF-IDF vectors). We then calculate the cosine similarity between Pre-event and Post-event topic vectors to find the best semantic matches.

This alignment allows us to track how specific themes (e.g., "Economic Policy" or "Royal Family") evolved in prevalence and content, rather than comparing unrelated topic IDs.

4.4 Comparison and Analysis

We compare LDA and BERTopic by inspecting top words, representative articles, and topic distributions. Statistical tests (chi-square, t-test) and time series analysis are used to confirm the significance and robustness of observed agenda shifts.

4.5 Statistical Verification

We use standard test statistics to quantify changes in topic prevalence and evaluate topic quality.

Topic Coherence Metrics: To assess the interpretability of the generated topics, we employ two distinct coherence measures suited to each model's architecture. For LDA, we use C_{UMass},

which is based on document co-occurrence log-probabilities:

$$C_{UMass} = \sum_{i<j} \log \frac{D(w_i, w_j) + \varepsilon}{D(w_i)}$$

where $D(w_i, w_j)$ is the count of documents containing both words w_i and w_j, and $D(w_i)$ is the count for word w_i alone. For BERTopic, we use C_v, which combines a sliding window co-occurrence with normalized pointwise mutual information (NPMI) and cosine similarity:

$$C_v = \sum_{i<j} \cos(\vec{v}_{w_i}, \vec{v}_{w_j})$$

where $\vec{v}_w$ represents the context vector for word w derived from NPMI statistics.

Chi-square Test for Global Distributional Shift: We construct a contingency table where rows represent aligned topics and columns represent time periods (Pre vs Post). The Null Hypothesis (H_0) states that the topic distribution is independent of the time period, implying that the relative proportions of topics remain constant. The Pearson chi-square statistic is calculated as:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} is the observed document count for topic i in period j, and E_{ij} is the expected count under H_0. A significant result leads to the rejection of H_0, confirming a significant shift in the media agenda.

Z-test for Individual Topic Proportions: For specific topics of interest (e.g., those showing large absolute changes), we perform a Z-test for the difference of two proportions to determine if the change in prevalence is statistically significant. The test statistic is given by:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions of the topic in the pre- and post-event periods respectively, $\hat{p}$ is the pooled proportion, and n_1, n_2 are the sample sizes. This allows us to pinpoint exactly which narratives drove the global shift.

5 Experiments

We train LDA and BERTopic models on the filtered pre- and post-event corpora (Section 3). For BERTopic, we use `all-MiniLM-L6-v2` embeddings, UMAP, and HDBSCAN; for LDA, we use CountVectorizer and scikit-learn's implementation. Models were trained independently on the event-excluded subsets. Evaluation combines quantitative metrics (standard C_v for BERTopic, C_{UMass} for LDA) and qualitative inspection of top words and time series. Statistical tests include Pearson chi-square for distribution shifts and Z-tests for proportion differences.

6 Analysis and Discussion

Our semantic alignment reveals clear agenda shifts.

6.1 Aligned Topic Evolution

Figures 2 and 3 show weekly prevalence for top pre-event topics and their post-event matches. **BERTopic** (Figure 2) captures thematic continuity with distinct "dips" and "recoveries", validating the "agenda reset" hypothesis. Topics like "Political Leadership" show increased prominence post-event. **LDA** (Figure 3) shows coarser matches and flatter trends, as its bag-of-words nature blends distinct themes. BERTopic's sharper dynamics underscore its superior sensitivity to fine-grained shifts.

6.2 Quantitative Shifts and Statistical Significance

A Pearson chi-square test confirms a highly significant distribution change ($\chi^2 = 491.618$, $p < 10^{-6}$), indicating a major shift in the media agenda. Z-tests pinpoint specific drivers: "Economic Policy" (Topic 0) surged from 1.59% to 2.54% ($Z \approx 6.58, p < 10^{-10}$), and "Political Leadership" (Topic 1) rose from 1.42% to 2.09% ($Z \approx 4.63, p < 10^{-5}$), confirming the consolidation of substantive narratives.

Table 1 summarizes key metrics. Coherence scores confirm topic quality: BERTopic achieves solid C_v values (Pre: 0.3950, Post: 0.4455), with a +0.05 post-event increase suggesting a more focused agenda. LDA's low C_{UMass} scores (-5.27 to -4.65) reflect its struggle with short news texts. Qualitatively, BERTopic produces more coherent topics, and time series confirm these shifts are sustained, not transient.

7 Conclusion and Future Work

We present a practical pipeline for measuring media agenda shifts around major events while minimizing the influence of direct event reports. Applied to The Guardian's coverage surrounding Queen Elizabeth II's passing, our analysis reveals distinct differences in the news agenda. Despite excluding direct coverage, we observed a statistically significant "agenda reset": general "noise" or diffuse reporting decreased, while specific post-event narratives (e.g., Economic Policy, Political Leadership) saw consolidated growth (e.g., Topic 0 increasing from 1.59% to 2.54%). BERTopic proved superior to LDA in capturing these fine-grained shifts, showing how established themes were briefly displaced before returning with altered intensity. Future work should evaluate sensitivity to exclusion parameters and extend this analysis to multiple media outlets.

8 Limitations

Our analysis is limited by the single-outlet scope (The Guardian) and the specific exclusion parameters (keyword list, 30-day buffer). While the buffer removes immediate event reports, indirect references likely remain. Additionally, our semantic alignment strategy assumes a 1-to-1 mapping between pre-event and post-event topics, which simplifies the reality where topics may split or merge (many-to-many). BERTopic also imposes higher computational costs compared to LDA. Future work should explore many-to-many alignment techniques and validate findings across broader news datasets.

9 Ethical Considerations and AI Usage

This study involves the analysis of public news data; no private or personally identifiable information was processed. We acknowledge the use of Large Language Models (specifically GPT-4o) to assist with LaTeX formatting, code refactoring, and linguistic polishing to enhance readability. All scientific claims, experimental designs, and data interpretations remain the sole responsibility of the author.

10 Data and Code Availability

The complete code for data collection, preprocessing, and topic modelling, along with the processed datasets used in this study, are publicly available at: https://github.com/orzmlx/guardian-topic-analysis.

References

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM.

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

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606.

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

David Hall, Daniel Jurafsky, and Christopher D Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 363–371.

Alexander Hoyle, Rahul Goel, Andrew Anderson, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2014. Computational history of the US supreme court. In *Proceedings of the Workshop on Language Technologies and Computational Social Science*, pages 12–16.

Maxwell E McCombs and Donald L Shaw. 1972. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.