RESEARCH NOTE

Exploring the Editorial Topics in The Guardian (Jan 1, 2024 – Nov 30, 2024)

Abstract: This study investigates the thematic structure and narrative depth of The Guardian's editorial content published between January 1, 2024, and November 30, 2024. The research employs two advanced natural language processing (NLP) techniques – Latent Dirichlet Allocation (LDA) and BERT embeddings with clustering – to identify and analyze key themes. The findings reveal 15 distinct topics across the editorial corpus, including governance, healthcare, climate change, political dynamics, international conflicts, and social justice. LDA's results highlight overarching themes such as fiscal policies, electoral strategies, and geopolitical conflicts, demonstrating The Guardian's emphasis on systemic issues. BERT complements these insights by uncovering detailed narratives, such as patient experiences in healthcare, cultural reflections, and the complexities of global diplomacy.

Natural Language Processing – F 2024

(GRAD-E1282)

Submitted By: Aditya Narayan Rai (235843)

Introduction:

The editorial sections of newspapers play a crucial role in shaping public discourse by providing essential insights into the socio-political and cultural perspectives of the time. They serve as a platform for newspapers to articulate their stances on pressing issues, allowing readers to engage with the opinions that influence societal norms and policies. As a longstanding voice in British journalism, *The Guardian* has established itself as a key platform for expressing diverse perspectives on both global and domestic matters. Its editorials, written by a team of experienced journalists and editors, delve into important topics ranging from climate change and social justice to economic policy and international relations, reflecting the institution's commitment to informed commentary. In contrast to standard news articles, which focus on presenting factual information, editorials are crafted to advocate for particular viewpoints, often pushing for specific actions or changes in public opinion. This makes them a valuable resource for understanding the themes and narratives prioritized by influential media outlets over time. The year 2024 saw an array of pressing global and local challenges, ranging from geopolitical tensions to environmental crises and social equity debates. This study aims to systematically analyze editorials published between January 1, 2024, and November 30, 2024, to uncover thematic patterns and shifts over time. The significance of this analysis lies in its potential to provide insights into how editorial content aligns with societal concerns and its role as an agenda-setter in public discourse.

This study uses two topic modeling approaches – Latent Dirichlet Allocation (LDA) and BERT embeddings combined with clustering algorithms – to achieve a dual-layered understanding of editorial themes. LDA, a probabilistic model, identifies topics based on word co-occurrence patterns, providing a macroscopic view of the dataset. And, BERT embeddings, which encode semantic and contextual nuances, allow for a deeper exploration of the editorial narratives through clustering. By combining these two methodologies, this study addresses the following key questions about *The Guardian*'s editorial focus during this period:

1. What were the dominant themes across the corpus?
2. How do the results from a traditional probabilistic model compare with those from a contextualized, neural-network-based approach?

In addition to identifying the dominant topics, this research aims to interpret how these themes reflect broader socio-political contexts. For instance, topics like climate change, healthcare, and global conflicts – frequent in *The Guardian*'s editorials – offer a lens into the concerns prioritized in 2024's public discourse. Furthermore, this study examines how well the thematic structure of the editorials aligns with significant global events of the year, such as the Russia-Ukraine war, economic instability, and social justice movements.

Through this exploration, the research highlights the intersection of journalism, public opinion, and technology. The editorials not only mirror societal concerns but also shape them, serving as a critical node in the feedback loop between media and public sentiment. The findings of this study thus provide valuable insights into the narratives that dominated one of the most turbulent years in recent history, offering a foundation for further inquiry into media influence, agenda-setting, and the evolution of public discourse in the digital age.

Methodology[1]:

To explore the thematic structure of editorials published by *The Guardian* from January 2024 to November 2024, this study used the advanced natural language processing (NLP) techniques. It combined Latent Dirichlet Allocation (LDA) and contextualized BERT embeddings, to disentangle the underlying topics and narratives within the editorials published. This section provides a step-by-step explanation of the methodology, including preprocessing, feature representation, topic modeling, and comparison of results.

Dataset and Preprocessing

The editorial articles for the study were collected using the Guardian Open Platform API[2], which provides access to the metadata and full text of published articles. The dataset consists of 4,503 textual entries, each with attributes for the publication date, headline, and content published between January 1, 2024, and November 30, 2024. Once the raw data collection was over, it was pre-processed to refine the raw textual data into a structured format suitable for computational analysis. The following steps were undertaken:

1. Tokenization and Lowercasing: The text data was split into individual words (tokens), and all words were converted to lowercase to ensure uniformity in representation.

[1] **Here is the GitHub Repo where you can access all the code chunks. For getting access to the data, get in touch with the author: https://github.com/adityanarayan-rai/editorials_topic_modeling**

[2] https://open-platform.theguardian.com/

2. Stopword Removal: Commonly used English stop words, such as “the,” “is,” “and,” which do not carry thematic meaning, were removed using SpaCy’s stopword list. This reduced the noise in the dataset and focused the analysis on contextually significant terms.
3. Punctuation and Non-Alphanumeric Removal: Special characters, symbols, and non-alphanumeric elements were eliminated to ensure the corpus consisted only of words relevant to thematic analysis.
4. Lemmatization: Each token was reduced to its base form or lemma (e.g., “running” to “run”), minimizing redundancy and ensuring semantic consistency across variants of the same word.
5. Frequency Filtering: Terms that appeared either too frequently or too rarely across the dataset were filtered out. The threshold was set using a minimum document frequency (min_df = 2), and the vocabulary size was limited to 5,000 terms.
6. Bag-of-Words (BoW) Representation: The cleaned dataset was converted into a BoW matrix using the CountVectorizer. This matrix represented the frequency of terms in each document, forming the foundation for LDA analysis.

The preprocessing steps ensured that the corpus was semantically rich, lacking irrelevant noise, and computationally efficient for downstream modeling.

Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model widely used in topic modeling. It assumes that each document is a mixture of topics and that each topic is a distribution of words. The following steps outline the application of LDA:

1. Input Transformation: The BoW matrix served as the input for LDA, where each document was represented as a row vector of term frequencies.
2. Determining the Number of Topics: Perplexity scores were calculated across different topic numbers (ranging from 5 to 20) to determine the optimal number of topics. A configuration of 15 topics was chosen, balancing interpretability and thematic granularity.
3. Model Fitting: The LDA model was trained on the BoW matrix. During training, the algorithm identified word co-occurrence patterns to infer latent topics.
4. Topic Representation: The model output consisted of:

    a. A document-topic matrix, where each document was represented as a probability distribution over topics.
    b. A term-topic matrix, which indicated the probability of each word belonging to a particular topic.
5. Topic Labeling: For each topic, the top ten most probable words were extracted and analyzed to assign thematic labels. These labels encapsulated the overarching themes captured by the topics.

BERT Embeddings and K-Means Clustering

The contextualized embeddings from BERT were employed to capture the semantic nuances of the corpus, complementing the LDA that was done first. The BERT embeddings condense the meaning of words in context, enabling a richer representation of the editorials. The steps included:

1. Embedding Generation: SentenceTransformer, a BERT-based pre-trained model, was used to transform each article into a 384-dimensional embedding vector. These embeddings captured both syntactic and semantic relationships.
2. Dimensionality Reduction: Principal Component Analysis (PCA) was applied to reduce the embeddings to 50 dimensions, balancing computational efficiency with information retention.
3. Clustering: K-Means clustering was performed on the reduced embeddings to group articles into semantically coherent clusters. The number of clusters was set to 15 to align with the number of LDA topics.
4. Cluster Labeling: The representative documents from each cluster were analyzed to infer topic labels. The top documents closest to the cluster centroids were identified as exemplars for each topic.

Results:

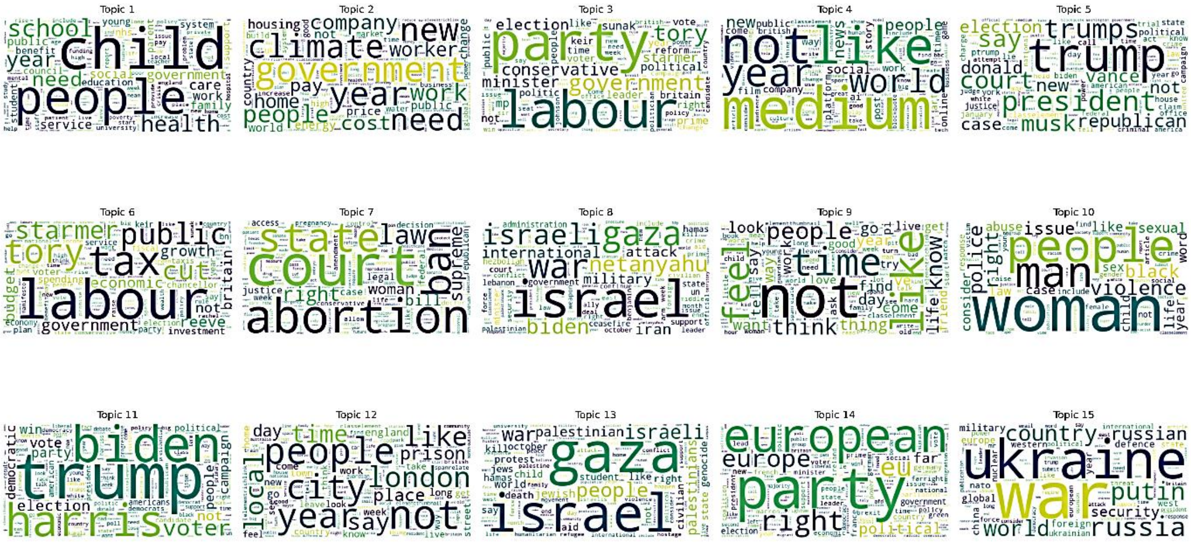
The dataset contains 4,503 entries, with unique headlines and content. The headline lengths ranged from 17 to 134 characters, while content lengths varied from 1,880 to 48,873 characters. This diversity reflects the dataset’s capacity to capture a wide range of textual expressions, from brief summaries to detailed narratives. The frequent words such as “government,” “people,” “climate,” and “year” indicated the editorial focus on governance, societal trends, and environmental concerns. The BoW matrix exhibited high sparsity (94.94%), which reflects the diversity of terms and the distinctiveness of the editorial articles. The frequent word analysis revealed meaningful insights. The terms like “not,” “people,” and “year” dominate, reflecting the editorial tone’s focus on collective challenges and temporal reflections. However, terms like “new” and “labour” underscore the emphasis on policy developments and political shifts.

Interpretation of the LDA Results:

The LDA model clusters the editorial corpus into distinct topics based on word co-occurrence patterns. Each topic is represented by the ten most salient terms, offering a high-level overview of thematic concentrations in the editorial data. Below are the 15 topics that I found from the LDA model:

- Topic 0 captures discussions around public services and social welfare, with terms such as “child,” “school,” and “health.” These terms suggest a focus on systemic issues within education and healthcare, areas frequently debated in the context of government policies and societal well-being. The inclusion of “government” highlights its central role in shaping these sectors.
- Topic 1 is centered on climate and economic concerns. The words like “climate,” “work,” and “cost” indicate the editorial emphasis on the economic implications of climate policies and the need for sustainable labor practices. This topic aligns with global conversations on balancing economic growth and environmental responsibility.
- Topic 2 focuses on politics, particularly within the United Kingdom. The terms “party,” “labour,” “tory,” and “conservative” suggest a detailed examination of political dynamics, elections, and leadership, with particular attention to figures like Keir Starmer and Rishi Sunak.
- Topic 3 is more abstract, characterized by terms like “medium,” “social,” and “public.” This topic seems to delve into the role of media and public discourse, which reflects how information and social narratives shape societal understanding.
- Topic 4 shifts to U.S. politics, dominated by terms like “trump,” “president,” and “court.” This indicates a focus on Donald Trump's presidency, legal controversies, and broader implications for American political structures.

- Topic 5 revisits the UK political landscape but emphasizes fiscal policies. Terms such as “labour,” “tax,” and “growth” point to debates on economic management, austerity, and public spending under Labour and Conservative leaderships.
- Topic 6 highlights judicial and reproductive rights, with terms like “court,” “abortion,” and “law.” This topic underscores contentious debates around state laws, Supreme Court decisions, and women's rights.
- Topic 7 addresses the Israel-Palestine conflict, with “israel,” “gaza,” and “war” leading the discussion. This topic reflects on military escalations, political leadership, and international diplomatic interventions.
- Topic 8 seems more conversational, with terms like “feel,” “know,” and “think” dominating. This topic might represent editorials exploring societal attitudes, personal reflections, and cultural dynamics.
- Topic 9 examines gender and social justice issues, indicated by terms such as “woman,” “violence,” and “police.” This suggests coverage of systemic inequalities, gender-based violence, and policing policies.
- Topic 10 returns to U.S. politics, particularly election processes, with terms like “trump,” “biden,” and “election.” This aligns with ongoing debates about voter behavior, democratic integrity, and campaign strategies.
- Topic 11 focuses on urban issues, with “city,” “london,” and “local” suggesting discussions around urban policy, infrastructure, and community-level governance.
- Topic 12 revisits the Israel-Palestine topic but expands to include global implications, with terms like “gaza,” “war,” and “jewish.” This broader framing highlights geopolitical ramifications and humanitarian concerns.
- Topic 13 deals with European politics, with “party,” “european,” and “france” indicating themes around EU governance, political alignments, and regional challenges.
- Topic 14 captures discussions around global conflicts, particularly the Russia-Ukraine war, with terms like “war,” “ukraine,” and “russia” reflecting the editorial focus on international security and geopolitical stability.



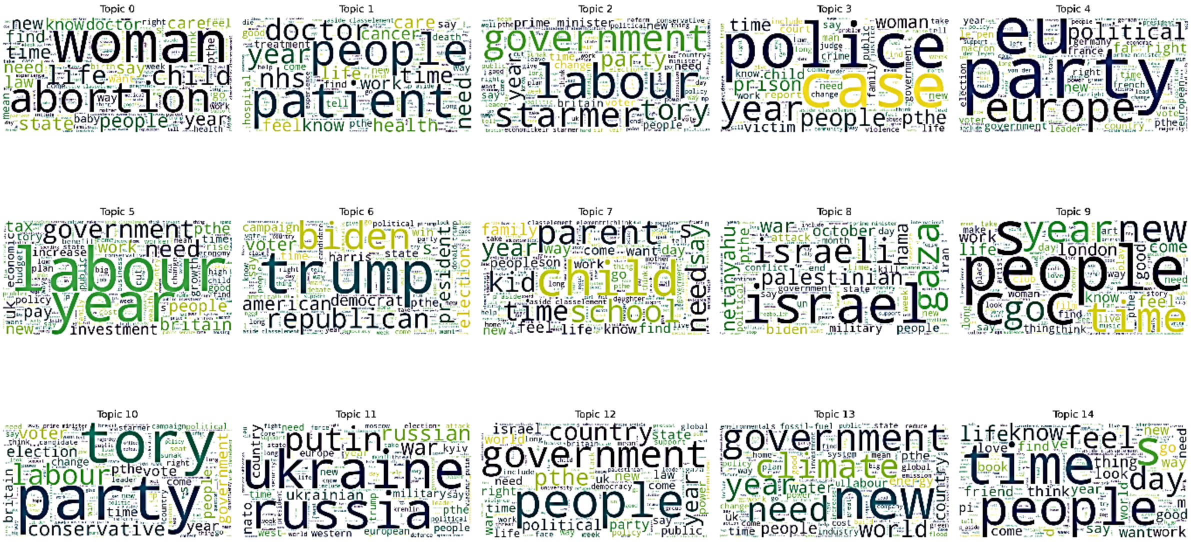
Interpretation of BERT Results:

The BERT-derived topics provide richer semantic narratives due to their contextual embeddings. These topics not only group words but also emphasize phrases and document-level coherence, offering deeper insights. Below are the 15 topics that I found from the BERT model:

- Topic 0 explores healthcare reforms, with detailed references to pregnancy loss funding, endometriosis treatment, and ethical dilemmas in abortion care. Unlike LDA's broad categorization, BERT highlights systemic barriers, patient narratives, and federal policy implications.
- Topic 1 delves into the UK healthcare crisis, emphasizing the plight of patients with chronic conditions like myalgic encephalomyelitis (ME/CFS). It critiques the psychological burden placed on patients by systemic inefficiencies and advocates for personalized care.
- Topic 2 focuses on Keir Starmer's leadership and Labour's political strategy, providing nuanced narratives about speeches, donor controversies, and party dynamics. This thematic richness surpasses LDA's more generic political clustering.
- Topic 3 examines criminal justice, with discussions ranging from systemic safeguards to high-profile cases like Sarah Everard's murder. The BERT captures the emotional and ethical dimensions of justice, offering a granular view of editorial perspectives.

- Topic 4 addresses European politics, particularly the rise of far-right parties. The BERT emphasizes election campaigns, policy shifts, and the broader implications of political polarization, enriching LDA’s thematic outline.
- Topic 5 critiques the UK economy, highlighting structural flaws, fiscal policies, and Labour’s proposed economic strategies. This topic reveals deeper policy critiques and economic analyses compared to LDA’s high-level terms.
- Topic 6 focuses on U.S. politics, with narratives about Biden’s State of the Union address, Republican dynamics, and Robert F. Kennedy Jr.’s campaign. The BERT’s ability to capture political nuances enhances the understanding of American political discourse.
- Topic 7 examines parenting and societal norms, discussing stereotypes, guilt, and the role of play in child development. This thematic focus is less prominent in LDA, showcasing BERT’s ability to identify unique angles.
- Topic 8 revisits the Israel-Palestine conflict but emphasizes international accountability and humanitarian crises. The BERT provides a more detailed account of diplomatic failures and political escalations.
- Topic 9 explores British culture, from iconic films to contemporary art. This topic highlights narratives around cultural identity and artistic evolution, which are less evident in LDA.
- Topic 10 critiques UK politics, focusing on electoral shifts, Tory leadership, and Labour’s potential governance. The BERT captures the intricacies of political transitions and public sentiment.
- Topic 11 addresses the Russia-Ukraine war, emphasizing humanitarian atrocities, cross-border incursions, and geopolitical strategies. This topic aligns with LDA but provides richer narratives.
- Topic 12 critiques systemic racism and immigration policies, discussing public protests, racial history, and governmental actions. BERT’s focus on marginalized voices and historical contexts adds depth to LDA’s thematic scope.
- Topic 13 focuses on climate change, discussing carbon capture, extreme weather, and global summits. The BERT emphasizes policy failures and scientific urgency, enhancing LDA’s high-level terms.

- Topic 14 explores personal narratives, from mental resilience to menopause. This unique focus on individual experiences and societal perceptions is largely absent in LDA.



Comparison of LDA and BERT Models:

The editorial content of *The Guardian* revealed a wide array of themes that align with the global and domestic challenges of 2024. The LDA model provided an overarching view of these themes, highlighting topics such as governance, social justice, economic policies, and international relations. For instance, topics like “child, school, health” (Topic 0) and “climate, work, cost” (Topic 1) reflect editorial emphasis on systemic challenges in public services and the economy, respectively. These themes resonate with ongoing debates about the impact of austerity, public healthcare reforms, and the economic implications of climate policies. Similarly, topics related to U.S. and U.K. politics (e.g., “trump, president, court” in Topic 4 and “party, labour, tory” in Topic 2) illustrate how editorials engage with leadership dynamics, electoral strategies, and political accountability.

The BERT embeddings, in contrast, provided a more nuanced exploration of these themes, capturing semantic and contextual subtleties. For example, while LDA broadly categorized healthcare into systemic discussions, BERT highlighted specific narratives, such as the ethical dilemmas in abortion care, the psychological burden on patients with chronic illnesses, and the implications of federal healthcare policies. These granular insights reveal how editorials weave individual experiences into broader systemic critiques, making them accessible and relatable to readers. Similarly, the Israel-Palestine conflict emerged as a significant theme in both models, but the BERT results delved deeper into the humanitarian crises, diplomatic failures, and political escalations. The LDA model categorized this under broad terms such as “israel, gaza, war” (Topic 7), whereas BERT emphasized international accountability, public backlash against military actions, and nuanced geopolitical implications. This difference highlights how contextual embeddings can enrich our understanding of editorial narratives by uncovering layers of meaning that probabilistic models might overlook.

Conclusion:

This study demonstrates the value of using advanced NLP techniques to analyze editorial content, offering insights into the thematic structure, narrative depth, and societal impact of *The Guardian*’s editorials in 2024. The combination of LDA and BERT provided a comprehensive understanding of the corpus, revealing both broad thematic patterns and detailed contextual narratives. The findings underscore the importance of journalism in framing public discourse, highlighting its role as an agenda-setter and a mirror of societal concerns. However, several areas warrant further exploration. First, the integration of dynamic topic modeling could provide insights into how themes evolve over time, capturing shifts in editorial focus in response to global events. Second, incorporating sentiment analysis could add another layer of interpretation, revealing the emotional tone of editorials and their potential influence on public opinion. Third, expanding the analysis to include other media outlets could facilitate comparative studies, uncovering differences in editorial priorities and narrative strategies across publications. Also, the use of more advanced contextual models, such as GPT-based embeddings, could further enhance the analysis, capturing even finer nuances in editorial narratives. Future research could also explore the interplay between editorials and audience engagement metrics, such as reader comments and social media shares, to understand the impact of editorial content on public discourse.

In conclusion, this study highlights the potential of hybrid methodologies in media analysis, providing a robust framework for understanding the thematic and narrative structure of editorial content. By combining LDA’s macroscopic perspective with BERT’s contextual depth, it offers a comprehensive lens into the editorial priorities of *The Guardian*, contributing to a deeper understanding of journalism's role in shaping public discourse in a rapidly evolving socio-political landscape.

Note to the Prof: I have used Grammarly to edit and structure the sentences written in this research note.