THE FLORIDA STATE UNIVERSITY

COLLEGE OF SOCIAL SCIENCES AND

PUBLIC POLICY

IDENTIFYING AND ANALYZING SELECTIVITY IN NEWS MEDIA COVERAGE THROUGH HIERARCHICAL TOPIC MODELING: ISRAEL-PALESTINE AS A CASE STUDY

By

JACOB RAMPINO

A Thesis submitted to the Department of
Political Science
in partial fulfillment of the requirements for graduation with
Honors in the Major

Degree Awarded:
Spring, 2025

The members of the Defense Committee approve the thesis of Jacob Rampino defended on April 21, 2025.

Holger L. Kern

Holger L. Kern
Thesis Director

Deana A. Rohlinger

Deana A. Rohlinger
Outside Committee Member

Quintin H. Beazer

Quintin H. Beazer
Committee Member

Abstract

This study utilizes a corpus of news articles to identify broad trends in media selectivity regarding coverage of Israel-Palestine. The corpus is sourced from Nexis Uni, encompassing 8,500 articles composed of 302,564 sentences from 287 major news outlets, spanning from October 6, 2023 to October 18, 2023. By leveraging BERTopic at the sentence level, the semantic and syntactic content of articles was clustered into specific hierarchical topic groups. These groups were then aggregated across source and region level to analyze selective trends in coverage. The significance of this research lies in its potential to expand data-driven media studies by studying selectivity through semantically-oriented hierarchical topic modeling as a more objective and informative lens of analyzing the media agenda. Much prior political communication research on news-media emphasized their studies as analyses of bias, sentiment, or discourse. This study analyzes selectivity in coverage, out of the conceptualization of media simultaneously as framers and gatekeepers. Focus on selectivity reorients operationalization towards objective patterns of inclusion and exclusion at multiple scales.

Literature Review

This study seeks to build off a framework of media analysis which has recently developed utilizing topic modeling to analyze selective coverage trends within media. News analysis has historically been qualitative and hand-coded, which is why topic modeling is considered so useful, as it can analyze the content of thousands of documents at a time and identify clusters of information within them without qualitative human input. Topic modeling has been used throughout sociology, communications and political science to begin identifying the content of innumerable written sources. That includes advertisements, social media tweets, and the news itself. When topic modeling is applied to the news, it is often used for opinion mining,

sentiment analysis, or identifying partisanship to analyze the tone and purpose of news content. As applied within this study, topic modeling is employed to analyze selective trends in the media agenda, a specific framework within agenda-setting theory.

Originally introduced by McCombs and Shaw (1972), agenda-setting theory argues that the media is the fundamental decision-maker in what gets included in news coverage, defining the boundaries of public discourse. If something is not in the news agenda, it effectively is not considered newsworthy. Within this framework, gatekeeping is the process through which media outlets decide which topics or events deserve representation at all, acting as a first filter.

Framing, on the other hand, comes after this selection, and refers to how an issue and its components are presented once included. If certain components of a frame are gatekept, that alters what a frame communicates. Schaefer and Birkland (2007) define framing as "the ways in which media emphasize or de-emphasize aspects of the reality on which they are reporting." These emphases can appear through diction, story structure, sourcing choices, or visual cues which guide how audiences understand an issue. This creates issue salience by telling the audience which elements of a story are more or less important. Robert Entman (2007) offers a succinct definition: "Framing is the process of culling a few elements of perceived reality and assembling a narrative that highlights connections among them to promote a particular interpretation." Entman argues that framing depends on prior gatekeeping, meaning you can only frame something once individual elements have been selected for or against emphasizing.

This interaction between gatekeeping and framing to produce unique structures of salience was applied by Hostetter and Buss to a broader concept of selectivity, which they characterize as ubiquitous: "Any time one abstracts or generalizes, one is selecting certain details rather than others. Selectivity involves choosing and interrelating facts in some coherent way.

Implicit theories about society guide the selection process. Some facts are chosen and others omitted due to "judgments of relevance." Relevance is determined by journalists' view of the world. The audience, then, receives only a partial reconstruction and interpretation of political events." (1978) Selectivity is at its core the interaction between gatekeeping and framing as effects which produce the news agenda and its internal salience.

In this study, topic modeling is used to detect and measure these selective patterns by observing which topics are gatekept through exclusion, or selectively framed through different organizations of topics to create issue salience. This framework does not analyze sentiment or ideology directly, but rather defines the structure of selection as emphasis in the construction of the media agenda.

This is not the first to attempt to identify agenda-setting trends utilizing topic modeling. Several recent studies have applied topic modeling to examine media coverage, each with distinct emphases in method and scope. Kwon et al. (2019) use Structural Topic Modeling (STM) to analyze framing trends across both social media and news coverage, paying particular attention to proximity-based topic shifts. They analyze topic prevalence across media types, but do not explore hierarchical structure or sentence-level patterns. Similarly, Kaiser et al. (2019) employ STM to assess right-wing partisanship in news coverage through a time-series hyperlink analysis, revealing how topics evolve over time and between sources. Lee and Jang (2021) extend this temporal focus to Twitter, using STM to model issue attention over time within social media discourse. While these studies emphasize framing and agenda-setting through STM, they tend to treat topics as flat and fixed categories, without deeper semantic structure or hierarchical relationships.

Heidenreich et al. (2019) apply Latent Dirichlet Allocation (LDA) to analyze framing over time within refugee-related media coverage, focusing on longitudinal shifts. Gao et al. (2023) also use LDA to study framing and issue salience in conservation news, introducing topic salience as a measurable statistical proportion. Sonmez and Codal (2024) apply LDA to terrorism-related content on the dark web, showcasing how topic modeling can be used for content analysis across unconventional media environments. Guo (2016) employs LDA in a time-series framework to measure changes in media topic emphasis, similar in structure to Terman (2017), who uses STM to model expected topic representation across sources and regions.

Although Lee et al. (2023) begin to explore BERTopic in content analysis, their work focuses primarily on time-series and academic media comparison. While their study generates a topic hierarchy, it is used largely for descriptive purposes rather than as a tool for structural or statistical analysis. This study builds upon that foundation by developing hierarchically structured topic classes and explicitly linking them to patterns of media selectivity across regions. In doing so, it offers a new application of semantically-informed, unsupervised hierarchical modeling to the study of agenda-setting.

For their part, LDA and STMs are useful, but can be limited tools in modern natural language processing analysis. An immediate concern is their inability to formulate hierarchical relationships between topics. Both models treat topics as flat, discrete clusters without deeper semantic nesting, which restricts their ability to capture layered patterns of salience. Additionally, LDA and STM require the number of topics to be predetermined by the researcher, increasing the self-guided nature of modeling. This necessitates estimating a valid number of

topics and subtopics in advance, essentially requiring the researcher to define the size and subsequent depth of the analysis beforehand, which may introduce bias and reduce flexibility.

This rigidity stems from the methodological underpinnings of LDA and STM. LDA operates on probabilistic assumptions, using word frequency, co-occurrence, and distribution patterns to identify topic clusters (Yang et al. 2023). While STM extends this by incorporating metadata and semantics, allowing researchers to examine how topic prevalence or content varies by covariates like time or source, it remains a semi-supervised model. The number of topics must still be specified beforehand, and the structure of topic relationships cannot be dynamically inferred from the data itself.

Compared to these earlier models, BERTopic offers a more flexible and modern methodology (Egger and Yu 2022). It employs dimensionality reduction and clustering techniques to extract structure from embeddings, enabling it to model semantic relationships with far less constraint. BERTopic is relatively new and has undergone less academic testing, but it offers several key advantages: embeddings allow for contextual clustering, the model supports hierarchical topic construction, and dynamic modeling (although unexplored here) allows for longitudinal analysis of topic evolution over time.

Another notable strength of BERTopic is its use of c-TF-IDF, a variant of the bag-of-words matrix that identifies the most salient n-grams within each topic cluster. This provides an immediate and interpretable representation of each topic's core vocabulary, improving transparency and interpretability. Finally, unlike LDA or STM, BERTopic can derive a hierarchical structure not only within topics, but between them. This allows for a nesting of topics and the words that compose them, offering a clearer visual and mathematical picture of

how issues are framed and assigned importance. This nested structure enables researchers to identify and analyze selection and salience patterns across multiple scales.

Framing, in particular, requires semantic considerations which BERTopic embedding allows for. Like STM, BERTopic's dimensionality allows for semantic considerations to contribute to topic formation. This is in combination with syntax, which is the primary method of topic formation for LDA. As Dietram Scheufele finds on a conceptual level, because framing is inherently about the construction of meaning from individual words, the many ways in which a word can be used semantically must be understood in order for framing to truly be analyzed (Scheufele and Iyengar 2017). Some words have multiple use cases, something which LDA does not account for. Therefore, BERTopic's semantic and syntactic components supersede the analysis allowed for by traditional topic models like LDA when discussing framing.

A crucial aspect of this study is the actual news content being analyzed. The Israeli-Palestinian conflict has received significant attention from scholars of political science and media, with research analyzing news coverage of the conflict since at least the mid-eighties (Vallone et al. 1985). This has covered qualitative and quantitative methodologies, topic modeling included. For example, Jackson (2023) employs a custom natural language processing pipeline to analyze the conflict as it is covered by the New York Times, specifically analyzing content for bias in objectivity, tone, and violent sentiment. Antonakakil and Ioannidis (2025) employ BERTopic to social media posts related to the Israel-Hamas War, running a comparative sentiment analysis across Telegram, Reddit and Twitter. Nefriana et al. (2024) utilizes BERTopic to analyze social media engagement and general topic frequencies within Facebook groups related to anti-Israel/Semitic and anti-Palestine/Muslim sentiments during the war in a longitudinal design. Steffen (2024) used BERTopic to analyze Telegram messages along with

images pertaining to the conflict. However, analysis of BERTopic as it applies to agenda-setting trends in the news media of the Israel-Hamas War does not yet exist. Given the historically critical nature of news coverage in and about conflict, a further investigation of selective media trends pertaining to the Israel-Hamas War may elucidate previously unconfirmed or suspected trends which fundamentally shape public understanding of an ongoing conflict. By applying these specific topic modeling methods, the media agenda of the war within the time frame of the corpus may also describe a general salience map of the conflict through the topic hierarchy.

Therefore, this paper builds on existing efforts to apply topic modeling in media analysis by pushing the method toward a clearer integration with agenda-setting theory. Rather than using topic models for sentiment or bias analysis, this study focuses on patterns of selection, wherein certain frames are minimized, emphasized, and how these patterns differ by region. The Israel-Hamas War is a critical case for such analysis, not only because of the intensity and global scope of the conflict, but also because of the long-standing concerns about selective media coverage surrounding it. Applying hierarchical, semantically-informed topic modeling to a large, international news corpus allows for a closer look at the mechanics of media selectivity. In doing so, this study contributes to three overlapping areas: it extends the application of topic modeling in political communication; it offers a novel methodological approach to analyzing agenda-setting through topic hierarchies; and it provides empirical insight into the ways the Israel-Hamas War has been framed regionally.

Methods

This study draws from a corpus of 8,500 English-language newspaper articles published between October 6 and October 18, 2023, across 287 distinct sources. Articles were selected using the Nexis Uni Power Search tool, filtered by the tags “2023 Gaza-Israel Conflict” or

“Israeli-Palestinian Conflicts.” To increase transparency while remaining compliant with Nexis Uni’s institutional access policies, a complete list of sources and publication dates per article is provided in Appendix A.

Then, 302565 sentences were extracted and given a unique ID to identify its location within the corpus by article and paragraph. These sentences were cleaned of stopwords and embedded using the Sentence Transformer model ‘all-mpnet-base-v2,’ without metadata. The original sentences, including stopwords, were stored with their corresponding IDs for qualitative content analysis. This meant that topics were produced entirely based upon semantic similarity and density, and were not clustered based upon the article, source, country or time the sentence was produced.

Dimensionality reduction was done utilizing PCA (Principal Component Analysis), lowering the number of dimensions from 768 to 5. Finally, HDBSCAN was applied, with parameters ‘min_cluster_size=500,’ ‘min_samples=50,’ ‘metric= "euclidean",’ ‘cluster_selection_method="leaf",’ and ‘approx_min_span_tree=True.’ Leaf, in particular, allowed for unique topics to be identified by the model and organized in a hierarchy. The minimum cluster size was set in combination with min_samples to improve topic coherence by setting a higher barrier for topic formation. For full PCA, HDBSCAN, and model storage code, see Appendix B. In order to aggregate topics by region and country, each source was also hand-coded (Appendix C) by the source’s overarching ownership and center of publication, along with the distribution of topics per sentence by source.

This produced 62 topics, the results of which are in Appendix D. The ‘label’ column is the only qualitative component of the representations, and were hand-coded based upon the c-TF-IDF representation produced by the model, as well as an analysis of the representative

sentences assigned to the topic. Importantly, several of the topics could not be qualitatively labeled, as is expected when working with unsupervised and unguided topic models. However, the vast majority of the topics produced were coherent and semantically distinguishable.

Hierarchical organization of these topics are produced through BERTopic as well, yielding 13 general branches which each of the 62 topics falls under. The full data on Parent IDs and Right and Left Children may be found in Appendix E.

Hypothesis Testing

To analyze framing trends within the corpus based upon the topics and hierarchy produced, a hypothesis was formed in order to focus on differences in regional and source based coverage within the hierarchy. Therefore, the regions of the United States and Europe were selected, as they were the most represented within the corpus, ensuring robust and representative results. A quick overview of the U.S. and Europe's representation in the corpus demonstrates the size and variability of these two regions:

	Source Count	Article Count	Sentence Count
Europe	54	3,021	120,432
US	103	1,200	55,196

Figure 1. U.S. and E.U. Descriptive Statistics

An explanation for why Europe has more articles per source is likely attributable to the kinds of sources present within the corpus related to the U.S., several of which are university newspapers with low article counts. For this reason, along with controlling for the general size difference between the two regions, topics are analyzed as a proportion of total source coverage as opposed to their absolute representation within the region and larger corpus. This proportional approach helps normalize for structural differences in source volume and output. To analyze selective trends pertaining to "Humanitarian Aid" and the topics in the surrounding branches

regionally, the first alternative hypothesis is focused specifically with Topic 5, which was hand-coded as such. Subsequent hypotheses extend this analysis to surrounding branches of the topic hierarchy.

H_1: "Humanitarian Aid" as a topic will appear more frequently in European media outlets than in U.S. outlets.

As a preliminary look at the distribution of Topic 5 as a proportion of all topics within the U.S. and Europe, I run a Linear Mixed Effects Model on the proportion of Topic 5 out of total topics appearing in articles from U.S. and European outlets, controlling for the time of publishing. Region is treated as a categorical variable, with the U.S. as the reference category and Europe as the comparison group, with source treated as a random intercept. While I do not explicitly control for total article count or the number of articles per source, the use of a Linear Mixed Effects Model with source treated as a random intercept accounts for inter-source variation. Additionally, because the dependent variable is defined as the proportion of Topic 5 relative to all topics within each source, the analysis is normalized at this level. This means the model evaluates within-region topic salience rather than absolute frequency, reducing the need for further controls related to article or source volume. These controls are also omitted to avoid overcontrolling, which could obscure meaningful variance attributable to regional framing.

Within the MixedLM (Appendix F1), the confidence intervals and the calculated p-values, the results indicate that Europe is positively and significantly associated with Topic 5 ("Humanitarian Aid") compared to the United States. The p-values for both the U.S. intercept and the European coefficient are estimated at ($p < 0.0000000001$), reflecting extremely strong statistical significance. In contrast, the publication day control has a high p-value (~0.99),

suggesting that the timing of publication has no meaningful effect on the likelihood of Topic 5 appearing in a European versus U.S. publication sentence.

To ensure robustness, I compare these results against an Ordinary Least Squares (OLS) model with the same controls, excluding a random intercept. The OLS model (Appendix F2) shows a slightly smaller but still positive and statistically significant increase in Topic 5 coverage from Europe as compared to the United States. Unlike the MixedLM, however, the intercept's coefficient is not statistically significant, indicating that the baseline level of Topic 5 coverage in U.S. outlets cannot be reliably distinguished from zero in the OLS model.

To further validate the stability of these findings, I bootstrapped the dataset by resampling rows with replacement over 1,000 iterations (Appendix F3). For each iteration, I refitted the OLS regression and collected the coefficient estimates. The final bootstrapped means and confidence intervals were computed from the distribution of these estimates. This provides an additional robustness check and helps account for any potential sampling variability. The Bootstrapped OLS indicates that the coefficient for Europe is fairly robust and maintains its statistical significance with tighter confidence intervals. The intercept is also not statistically significant, with a lower coefficient but a larger p-value. A comparison of the MixedLM, OLS, and Bootstrapped OLS models' European coefficients as compared to the intercept are seen below.

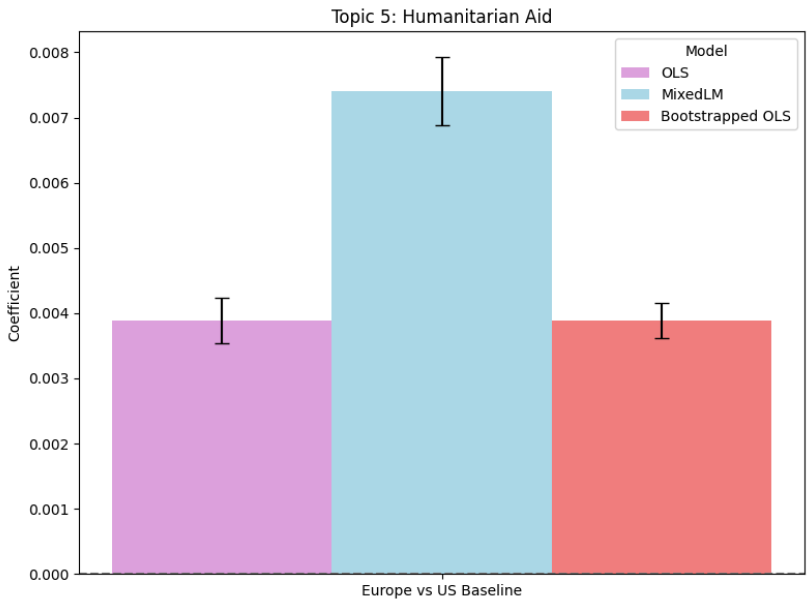


Figure 2. OLS, Bootstrapped OLS, and MixedLM Model Coefficients of European Coverage of Topic 5 ("Humanitarian Aid") Compared to U.S. as Baseline.

As a preliminary result, the OLS and MixedLM models all indicate a positive and statistically significant increase in coverage of Topic 5 ("Humanitarian Aid") from Europe as compared to the United States. However, these models are imperfect for testing topic proportions, as they presume topic proportions are unbounded and continuous, acting primarily as indicators. To ensure a robust result, I also run a logit-transformed MixedLM to account for the bounded nature of the topic proportions.

In the logit-transformed MixedLM model (Appendix F4), the magnitude of the Europe coefficient increases relative to the baseline, while retaining strong statistical significance. The publication date, as in all three prior models, remains uncorrelated with Topic 5 ("Humanitarian Aid"). Based on the results of the MixedLM, OLS, bootstrapped OLS, and logit-transformed MixedLM models, I reject the null hypothesis and find strong statistical evidence supporting the alternative hypothesis H_1. Crucially, this methodology does not attempt to control for broader

contextual factors such as economic trends, editorial priorities, or individual journalistic discretion, and should therefore not be interpreted as evidence of causality.

In order to probe the salience of topics closely related to humanitarian aid, I isolate the two branches most closely associated with Topic 5—hereafter referred to as *Class 1* and *Class 2*—from the hierarchical topic tree (Appendix E). Each class encompasses a distinct subset of topics that cluster closely around Topic 5 within the embedding space. These groupings represent structurally meaningful branches in the topic hierarchy, produced via unsupervised clustering methods and hierarchical modeling. Class 1 is seen below in Green, Class 2 in Orange:

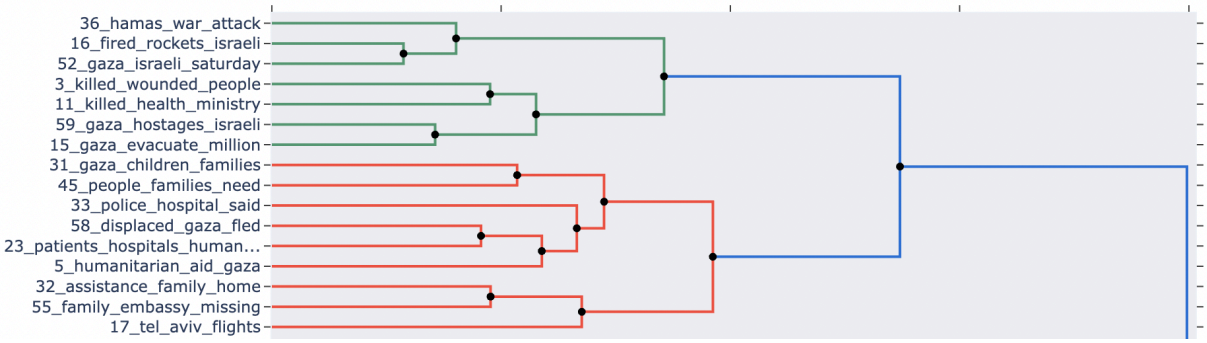


Figure 3. Class 1 and 2 Within the Hierarchy.

When analyzing the underlying Parent and ChildIDs within the hierarchy's structure (Appendix E), it is clear that Class 1's topics (36, 16, 52, 3, 11, 59, 15) cluster under ParentIDs 107 ("killed_gaza_israeli_people_hamas") and 110 ("gaza_humanitarian_aid_flights_tel"), which are children of 120 (gaza_killed_people_israeli_said). Class 2's topics (31, 45, 33, 58, 23, 5, 32, 55) cluster under ChildIDs 102 ("gaza_humanitarian_aid_people_crossing"), 98 ("tel_aviv_flights_embassy_israel"), 96 ("humanitarian_aid_gaza_crossing_egypt"), and 92 (same representation as 96), which are all part of a chain under Parent 110, and subsequently Parent 120. Based on both semantic content and structural placement within the hierarchy, I qualitatively interpret Class 1 as representing humanitarian issues in the context of violence,

while Class 2 reflects humanitarian issues in the context of logistics and aid coordination. I then produce three further alternative hypotheses.

H_2: European outlets will cover Class 2 more than the U.S.

H_{2a}: European outlets will cover Class 2 more than Class 1.

A preliminary topic distribution across the two regions reveals that Europe appears to proportionally cover both classes more than the U.S.

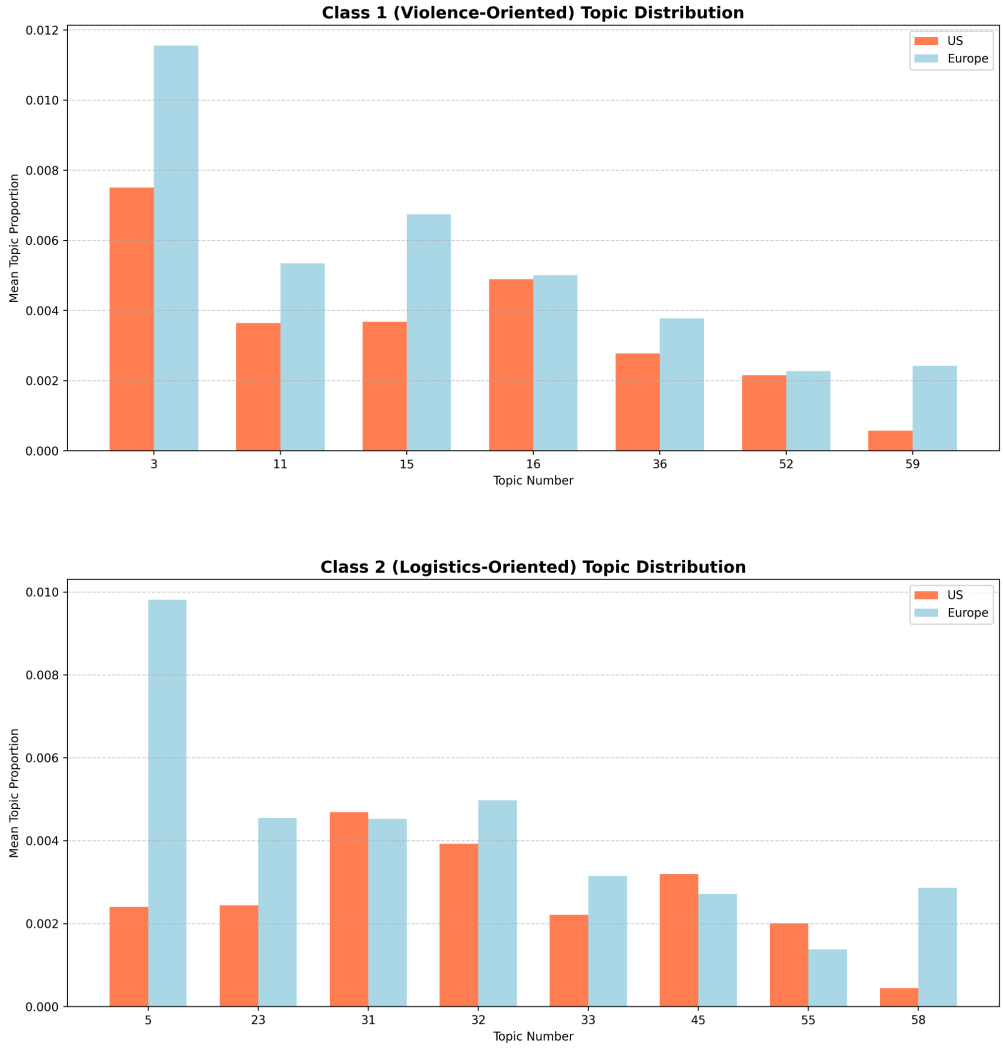


Figure 4, 5. Class 1 and 2 Topic Distribution Between U.S. and Europe.

A preliminary analysis of total class distributions between the U.S. and Europe reveals a similar top-down pattern, and indicates that Europe generally favors Classes 1 and 2, whereas the U.S. seems to favor Classes 4, 7, 9, 10 and 12.

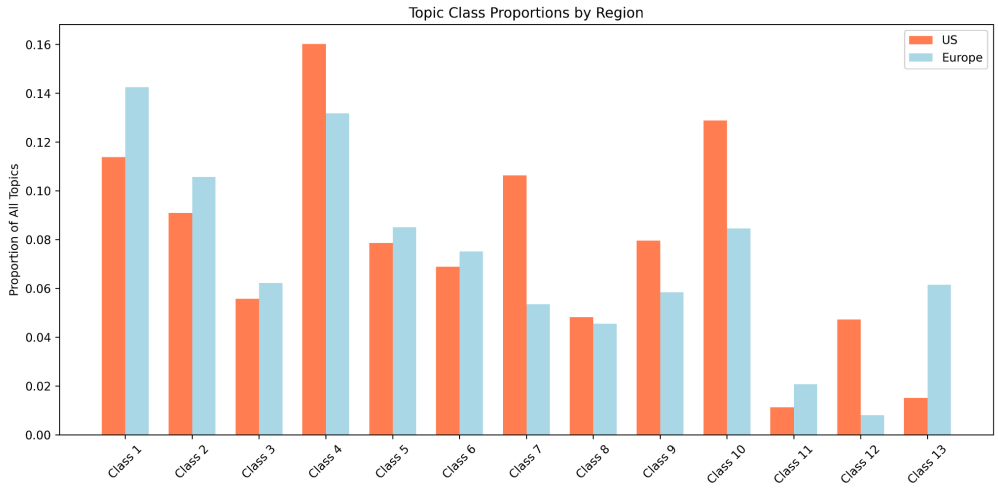


Figure 6. All-Class Distribution Between U.S. and Europe.

To test whether these observed differences in topic class salience between the U.S. and Europe are statistically significant, I run a series of logit-transformed Linear Mixed Effects Models. Given the consistency between models in the initial Topic 5 analysis, I now proceed directly to logit-transformed MixedLM for testing classes within the hierarchy, as it best accounts for the bounded and nested nature of the data.

The dependent variable is the proportion of all topics within either Class 1 or Class 2, relative to the total number of topic instances per source. Region is treated as a fixed effect, while publication date remains a control variable in all models. To test the robustness of these effects and account for variation across both outlets and topics, I run two separate logit-transformed Linear Mixed Effects Models. The first includes ‘topic’ as a random intercept to account for variation in usage between topics within each class. The second includes source as a random intercept to account for variation in coverage across media outlets. I estimate the

models separately to assess the consistency of region and class effects across both structures. This dual-model approach allows for a clearer interpretation of the main hypotheses while still accounting for nested variation in the data. While this analysis uses structurally defined classes from the topic hierarchy, future work may consider weighting topic frequencies by their embedding distance to Topic 5 in order to capture semantic gradients of humanitarian discourse.

The logit-transformed mixed effects models (Appendix F5 and F6) reveal clear and statistically significant regional differences in the coverage of Classes 1 and 2. European outlets consistently cover Class 1 topics more than U.S. outlets, with a coefficient of 2.5223 in the F6 (source as random intercept) model ($p < 0.001$) and 0.4508 in the F5 (topic as random intercept) model ($p < 0.001$). Within U.S. coverage, Class 2 topics are less emphasized than Class 1, with a statistically significant negative coefficient of -0.5382 in the F6 model ($p < 0.001$), but without significance in the F5 model ($p = 0.1641$). This suggests that while certain U.S. outlets disproportionately emphasize Class 2 in their framing, topic-level variation within these classes may dilute this effect. However, European outlets significantly increase their coverage of Class 2 topics compared to the U.S., well above the baseline differences for Class 1. This interaction effect is statistically significant in both models, with an estimated coefficient of 0.1069 in each ($p < 0.001$). Publication day is not a significant predictor in either model, suggesting that these differences are stable over time. Finally, group-level variance is substantial in the F6 model (1.1646) with a p-value of < 0.001, indicating considerable inter-source variability, while the F5 model shows a smaller but substantive variance (0.1777) with a p-value of 0.0114, both indicating statistically significant heterogeneity across topic usage within each class. Therefore, I reject the null hypothesis and find support for hypotheses H_2 and H_{2a}. Again, this methodology does not attempt to control for broader contextual factors such as economic trends, editorial

priorities, or individual journalistic discretion, and should therefore not be interpreted as evidence of causality.

Discussion

While causality cannot be determined, the statistically significant correlations found in support of H_1 indicate that European media is more likely to utilize frames related to humanitarian aid than the United States. Sources may therefore be more likely to gatekeep humanitarian aid as a salient issue within the United States as a matter of statistical significance. H_2 and H_{2a} both speak to the salience of humanitarian aid and its related topics within the agenda of European and U.S. news. By creating a topic hierarchy through which the semantic and syntactic content of the news can be categorized, and its implicit topics tested, larger trends of selection within the news as it pertains to framing, gatekeeping and salience trends may be analyzed. The application of regression models to the proportion of topics, as well as the frequency of usage within particular branches of the topic hierarchy, allows for these selective trends to be tested robustly.

That said, the study has important methodological limitations. First, it relies solely on a single unsupervised model and does not compare its results to other iterations or competing models like hLDA or STM. This limits claims of robustness and reproducibility, especially given BERTopic's known variability across runs without fixed random states. Although this study uses PCA rather than UMAP for dimensionality reduction, enabling fully replicable output, the broader concern of reproducibility still holds. Future replication of this model's topic outputs is a crucial next step for validating its findings.

A secondary limitation of this study is its corpus. While the number of sentences is over 300,000, drawn from 8,500 articles across 287 sources, the corpus as drawn from Nexis Uni may

be biased in particular ways which cannot be controlled for, or otherwise recognized as different from any public newsfeed. While a broad variety of sources across thirteen days of coverage is captured in this corpus, conclusions being drawn to the news-at-large should be done with potential sampling errors in mind. A key example is the number of articles per source, which varies a great deal within the corpus itself and may potentially skew data towards certain large or especially ideological outliers that a random intercept cannot totally account for.

Future research should continue to investigate the replicability of unsupervised hierarchical BERTopic modeling as it applies to agenda-setting trends within news coverage. Generating large corpuses and applying hierarchical BERTopic in this manner allows for a range of hypotheses to be built and tested, meaning that corpuses can be used to test innumerable hypotheses for which the corpus was not purpose-built. Any number of topics, branches, or keywords could be tested against any number of sources, regions, or individual articles, and deepen insights pertaining to framing and issue salience within the news agenda.

References

Antonakaki, Despoina, and Sotiris Ioannidis. "Israel–Hamas War through Telegram, Reddit and Twitter." *arXiv preprint* arXiv:2502.00060v1, 2025. https://arxiv.org/abs/2502.00060.

Egger, Roman and Joanne Yu. "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts." *Sociological Theory* Volume 7 (2022). https://doi.org/10.3389/fsoc.2022.886498.

Entman, Robert. "Framing Bias: Media in the Distribution of Power." *Journal of Communication 57*, no. 1 (2007): 164.

Gao, Y., Liu, Y., Luo, Y., Biggs, D., Zhao, W., & Clark, S. G. (2023). "Tracking Chinese newspaper coverage of elephant ivory through topic modeling." *Conservation Biology*, 37, e14072. https://doi.org/10.1111/cobi.14072.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). "Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling." *Journalism & Mass Communication Quarterly*, 93(2), 332-359. https://doi.org/10.1177/1077699016639231.

Heidenreich, Tobias, Fabienne Lind, Jakob-Moritz Eberl, Hajo G Boomgaarden. "Media Framing Dynamics of the 'European Refugee Crisis': A Comparative Topic Modelling Approach." *Journal of Refugee Studies*, Volume 32, December 2019, Pages i172–i182, https://doi.org/10.1093/jrs/fez025.

Hofstetter, C. Richard, and Terry F. Buss. 1978. “Bias in Television News Coverage of Political Events: A Methodological Analysis.” *Journal of Broadcasting* 22 (4): 517–30. doi:10.1080/08838157809363907.

Kaiser, J., Rauchfleisch, A., & Bourassa, N. (2019). “Connecting the (Far-)Right Dots: A Topic Modeling and Hyperlink Analysis of (Far-)Right Media Coverage during the US Elections 2016.” *Digital Journalism*, 8(3), 422–441. https://doi.org/10.1080/21670811.2019.1682629

Kwon, K. Hazel, Monica Chadha, Feng Wang. “Proximity and Networked News Public: Structural Topic Modeling of Global Twitter Conversations about the 2017 Quebec Mosque Shooting.” *International Journal of Communication*, [S.l.], v. 13, p. 24, jun. 2019. ISSN 1932-8036. https://ijoc.org/index.php/ijoc/article/view/11020.

Lee, C. S., & Jang, A. (2023). “Questing for Justice on Twitter: Topic Modeling of #StopAsianHate Discourses in the Wake of Atlanta Shooting.” *Crime & Delinquency*, 69(13-14), 2874-2900. https://doi.org/10.1177/00111287211057855.

Lee, Haein, Seon Hong Lee, Kyeo Re Lee, and Jang Hyun Kim. “ESG Discourse Analysis Through BERTopic: Comparing News Articles and Academic Papers.” *Computers, Materials & Continua* 2023, 75(3), 6023-6037. https://doi.org/10.32604/cmc.2023.039104.

McCombs, Maxwell E., and Donald L. Shaw. “The Agenda-Setting Function of Mass Media.” *The Public Opinion Quarterly 36*, no. 2 (1972): 176–87. http://www.jstor.org/stable/2747787.

Nefriana, Rr., Muheng Yan, Ahmad Diab, Wanhao Yu, Deborah L. Wheeler, Andrew Miller, Rebecca Hwa, and Yu-Ru Lin. "Shifting Patterns of Extremist Discourse on Facebook: Analyzing Trends and Developments During the Israel-Hamas Conflict." In *Proceedings of the 17th International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation*, 2024.

Schaefer, Todd M., and Thomas A. Birkland. Encyclopedia of Media and Politics. 9-89. Washington, D.C.: CQ Press. 2007.

Scheufele, Dietram A., and Shanto Iyengar. “43 The State of Framing Research: A Call for New Directions.” In *The Oxford handbook of political communication*, 619-22. 2017.

Sonmez, E., Seckin Codal, K. “Analyzing a Dark Web forum page in the context of terrorism: a topic modeling approach.” Secur J (2024). https://doi.org/10.1057/s41284-024-00421-9.

Steffen, Elisabeth. “More than Memes: A Multimodal Topic Modeling Approach to Conspiracy Theories on Telegram.” https://arxiv.org/html/2410.08642v1#Sx5 (2024).

Terman, Rochelle. “Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage.” *International Studies Quarterly*, Volume 61, Issue 3, September 2017, Pages 489–502, https://doi.org/10.1093/isq/sqx051.

Yang, Yuxuan, Yingmei Wei, Min Gao, Zanxi Ran, and Qi Wang. "Sentiment Analysis of Social Network Text Based on HDBSCAN and SO-PMI." *Journal of Physics:*

Conference Series 2504, no. 1 (05, 2023): 012055.

doi:https://doi.org/10.1088/1742-6596/2504/1/012055.

Appendix A (Article Metadata)

Appendix B (Model Code)

```
import os
import numpy as np
import pandas as pd
import pickle
import matplotlib.pyplot as plt
from bertopic import BERTopic
from umap import UMAP
from hdbscan import HDBSCAN
from sklearn.decomposition import PCA

# Paths
base_dir = os.path.expanduser("~/Documents/Rpath/")
model_store = os.path.join(base_dir, "modelstore")
os.makedirs(model_store, exist_ok=True)

sentence_path = os.path.join(base_dir, "final_sentences_fullstop.csv")
embedding_path = os.path.join(model_store, "embeddings_fullstop.npy")
reduced_embedding_path = os.path.join(model_store, "pca_model_test.npy")
cluster_path = os.path.join(model_store, "hdbscan_clusters_test.pkl")
csv_output_path = os.path.join(model_store, "topic_assignments.csv")

# Load embeddings
print("Loading sentences and embeddings...")
df = pd.read_csv(sentence_path)
sentences = df["Sentence"].astype(str).tolist()
embeddings = np.load(embedding_path)

print(f"Loaded {len(sentences)} sentences and embeddings of shape {embeddings.shape}")

# PCA instead of UMAP
print("Applying PCA for fast dimensionality reduction...")
pca = PCA(n_components=5)
reduced_embeddings = pca.fit_transform(embeddings)
np.save(reduced_embedding_path, reduced_embeddings)

print(f"PCA reduction complete. New shape: {reduced_embeddings.shape}")

# Run HDBSCAN
print("Running HDBSCAN with min_cluster_size=1000...")
hdbscan_model = HDBSCAN(min_cluster_size=1000, min_samples=50, metric="euclidean",
cluster_selection_method="leaf", approx_min_span_tree=True)
cluster_labels = hdbscan_model.fit(reduced_embeddings)

# Save HDBSCAN Clusters
with open(cluster_path, "wb") as f:
    pickle.dump(hdbscan_model, f)

print(f"HDBSCAN complete.")
```

```
# Train and save BERTopic
import os
import pickle
```

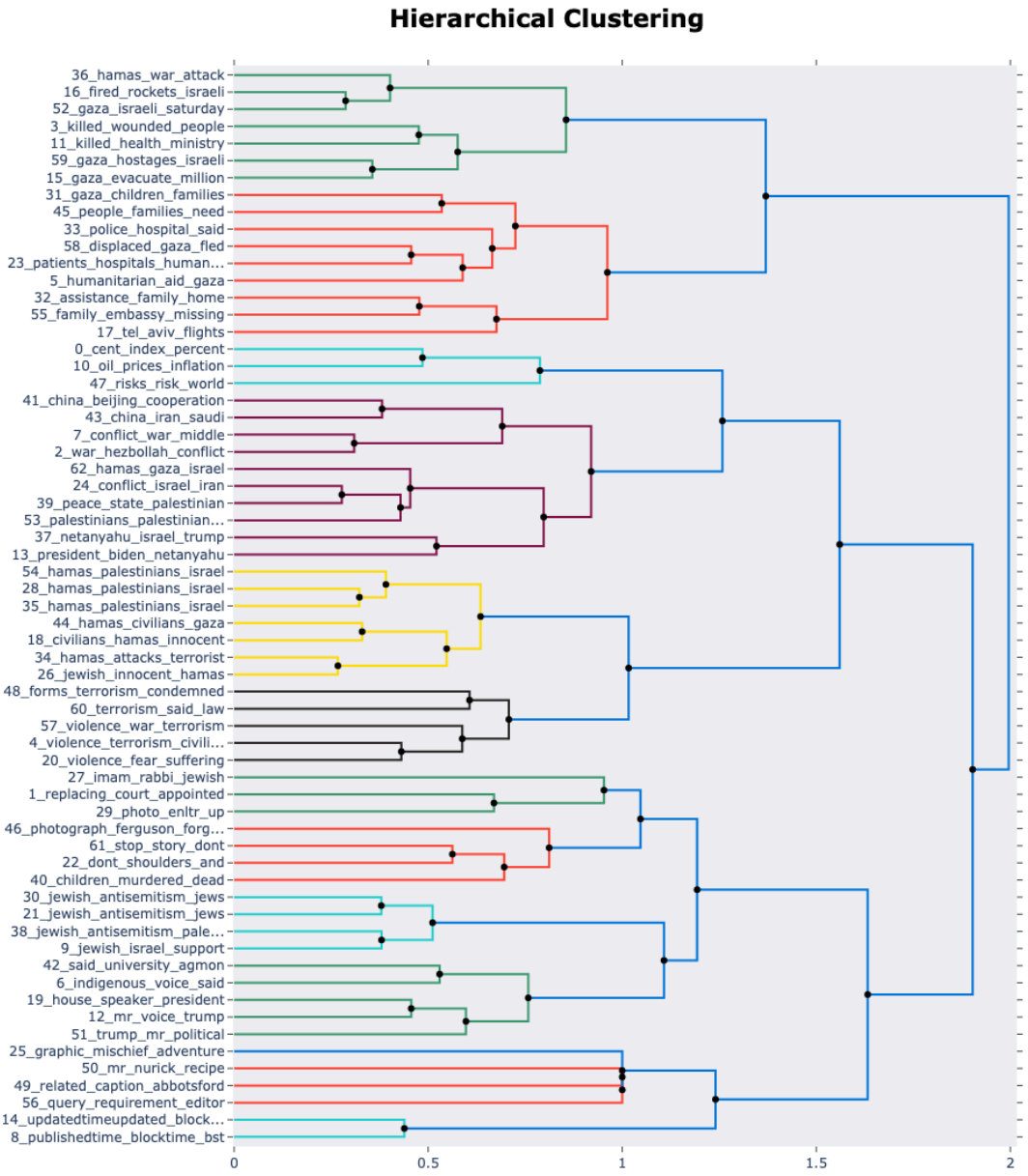
```
import numpy as np
import pandas as pd
from bertopic import BERTopic
from hdbscan import HDBSCAN
import joblib

# Paths
base_dir = os.path.expanduser("~/Documents/Rpath/")
model_store = os.path.join(base_dir, "modelstore")
os.makedirs(model_store, exist_ok=True)

sentence_path = os.path.join(base_dir, "final_sentences_fullstop.csv")
embedding_path = os.path.join(model_store, "pca_model_test.npy")
cluster_path = os.path.join(model_store, "hdbscan_clusters_test.pkl")
model_path = os.path.join(model_store, "bertopic_model_test.pkl")

# Load Data
print("Loading sentences and PCA-reduced embeddings...")
df = pd.read_csv(sentence_path)

# Ensure all sentences are strings
sentences = df["Sentence"].astype(str).tolist()

# Load PCA Embeddings
try:
    pca_embeddings = np.load(embedding_path)
    print(f"PCA embeddings loaded. Shape: {pca_embeddings.shape}")
except Exception as e:
    print(f"Error loading PCA embeddings: {e}")
    exit()

# Check PCA Embedding Shape
if len(sentences) != pca_embeddings.shape[0]:
    print(f"Mismatch: {len(sentences)} sentences vs. {pca_embeddings.shape[0]} embeddings.")
    exit()

# Load HDBSCAN Model
try:
    with open(cluster_path, "rb") as f:
        hdbscan_model = pickle.load(f)
    if not isinstance(hdbscan_model, HDBSCAN):
        raise TypeError("Loaded object is not a valid HDBSCAN model.")
    print("HDBSCAN model loaded successfully.")
except Exception as e:
    print(f"Error loading HDBSCAN model: {e}")
    exit()

# Train BERTopic Model
print("Training BERTopic model...")
try:
    topic_model = BERTopic(hdbscan_model=hdbscan_model)
    topics, probs = topic_model.fit_transform(sentences, pca_embeddings)
    print("BERTopic training completed.")
except ValueError as e:
    print(f"Error during BERTopic training: {e}")
    exit()

# Save Model
with open(model_path, "wb") as f:
```

```
    pickle.dump(topic_model, f)
print(f"BERTopic model saved.")
```

Appendix C (Hand-Coded Region)

Appendix D (Topics)

Appendix E (Hierarchy Structure)



Appendix F1-6 (Analysis Tables)

F1. H_1 MixedLM

MixedLM with Estimated Statistical Significance

Term	Coefficient	CI Low	CI High	Std. Error	z	p-value
US (Intercept)	0.0024028474	0.0020813582	0.0027243365	0.0001640251	14.6492685139	< 1e-10
Europe	0.0074026727	0.0068824816	0.0079228638	0.0002654036	27.8921315999	< 1e-10
Publication Day	-0.0	-1e-10	1e-10	1e-10	-4.7081e-06	0.9999962435

F2. H_1 OLS

OLS Regression Results

	Coef.	Std.Err.	t	P>|t|	CI Low	CI High
US (Intercept)	21.2076716501351	22.7953939263826	0.930348987107875	0.3522436633	-23.4833035732826	65.8986468735528
Europe	0.0038899870431537	0.000180353028256646	21.5687370528548	< 1e-10	0.00353640014119814	0.00424357394510925
Publication Day	-2.86986229057095e-05	3.08543437062349e-05	-0.930132339840054	0.3523557966	-8.91893832129598e-05	3.17921374015407e-05

F3. H_1 OLS Bootstrapped

Bootstrapped OLS Regression Results

	coef	ci_low	ci_high	Std. Error	z	p-value
US (Intercept)	20.2269026659	-27.6184688892	70.6121320647	25.058826774	0.8071767624	0.4195646517
Europe	0.0038858364	0.0036094134	0.0041584609	0.0001400631	27.7434639235	< 1e-10
Publication Day	-2.73711e-05	-9.55694e-05	3.73891e-05	3.3918e-05	-0.806979117	0.4196785146

F4. H_1 Logit-Transformed MixedLM

Logit-Transformed MixedLM Regression Results

	coef	ci_low	ci_high	std_err	z	pval
US (Intercept)	-10.0332988011	-10.1326587821	-9.9339388201	0.0506938679	-197.9193780899	< 1e-10
Europe	4.7086510069	4.5491872503	4.8681147635	0.0813590595	57.8749439389	< 1e-10
Publication Day	0.0	-4.52e-08	4.52e-08	2.31e-08	1.0968e-06	0.9999991249
Group Variance	16338801062.71844	15628433034.445135	17049169090.991745	362432667.4863806	45.0809282067	< 1e-10

F5. H_2 and H_{2a} Logit-Transformed LM with Topic as Random Intercept

Logit-Transformed MixedLM Results (Topic Random Intercept)

	Coefficient	CI Lower	CI Upper	Std. Error	z-score	p-value
Intercept (US, Class 1)	1897.2149	-1966.5053	5760.9351	1971.2858	0.9624	0.3358
Europe	0.4508	0.4061	0.4955	0.0228	19.752	0.0
Class 2	-0.5382	-1.2964	0.22	0.3868	-1.3913	0.1641
Europe × Class 2	0.1069	0.0457	0.1681	0.0312	3.4219	0.0006
Publication Day	-0.0026	-0.0078	0.0027	0.0027	-0.9656	0.3342
Group Variance	0.1777	0.04	0.3154	0.0703	2.5289	0.0114

F6. H_2 and H_{2a} Logit-Transformed LM with Source as Random Intercept

Logit-Transformed MixedLM Results (Source Random Intercept)

	Coefficient	CI Lower	CI Upper	Std. Error	z-score	p-value
Intercept (US, Class 1)	-45.0523	-3598.6549	3508.5502	1813.0625	-0.0248	0.9802
Europe	2.5223	1.9617	3.0828	0.286	8.8197	0.0
Class 2	-0.5382	-0.584	-0.4924	0.0234	-23.0299	0.0
Europe × Class 2	0.1069	0.0528	0.1611	0.0276	3.8702	0.0001
Publication Day	0.0	-0.0048	0.0049	0.0025	0.0197	0.9843
Group Variance	1.1646	0.9017	1.4275	0.1342	8.6811	0.0