Energy Policy in Flux: An NLP Approach to Media Discourse Before and After the Russia-Ukraine Conflict

Priscila Rebeca Stisman

April 29th, 2025

Abstract

This paper explores shifts in the discourse surrounding energy policy by applying natural language processing (NLP) techniques to New York Times articles published between 2021 and 2023. Using topic modeling approaches such as Latent Dirichlet Allocation (LDA) and Structural Topic Modeling (STM), alongside regression analysis, I examine whether the prominence and framing of energy-related issues changed following the beginning of the Russia-Ukraine war. My findings show that following the onset of the war, energy-related topics became significantly more prominent. However, their salience appears to decline gradually over time, suggesting a normalization or shift in media focus.

Keywords: *Energy Policy, NLP, Topic Modeling*

1. Introduction

In recent years, discussions around energy policy have become increasingly common. While topics like climate change, renewable energy, fossil fuels, and natural gas were once reserved for more technical or niche debates, they have now entered the mainstream. As climate change becomes an increasingly urgent global concern, energy security has emerged as a critical issue on the policy agenda.

However, this shift in focus is not without controversy. Some individuals remain skeptical of global warming, while others argue that pressing challenges such as poverty, inequality, economic instability, and geopolitical tensions should take precedence over investing in innovation for energy security. This divergence of opinion makes it particularly interesting to analyze whether public discourse around energy has shifted in recent years.

One valuable way to trace changes in public sentiment is through media reporting [3]. Media outlets play a crucial role in the policymaking process by acting as a bridge between politicians and the public. In this context, examining media coverage offers a useful proxy for understanding societal attention and concern around energy issues.

In this paper, I aim to examine whether there has been a significant shift in the media discourse surrounding energy since the onset of the Russia-Ukraine war. This conflict has had profound implications for global energy markets and security, potentially reshaping both the way energy issues are discussed in the media and the level of attention society pays to them. For this analysis, I will focus on The New York Times, one of the most influential newspapers in the world. With its extensive international readership and established role in shaping the global news agenda, The New York Times provides a valuable lens through which to assess changing narratives and shifting public attention on energy-related matters.

2. Data

I use New York Times articles from 2021 to 2023, filtered using a set of energy-related keywords. The New York Times API provides access to article metadata as soon as content is published on their website, while the Times Newswire API offers a real-time stream of newly released articles. The original dataset comprises over two million articles spanning from 2000 to 2025, each including metadata such as the publication date and time, headline, abstract, lead paragraph, associated tags, responsible department, section, authors, and word count.

For the purposes of my research, I focus on the years 2021 to 2023—surrounding the onset of the Russia-Ukraine war, which officially began on February 24, 2022. From this period, I retain 126,109 articles and focus on four key features: the abstract, headline, lead paragraph, and publication date.

To investigate whether there was a shift in the narrative around energy policy in the context of the Russia-Ukraine conflict, I further filter the dataset using a set of energy-related keywords. This results in a final subset of 18,293 energy-related articles.

3. Methods and Results

I begin by filtering the dataset to include only articles published between 2021 and 2023 and then further narrow it down to energy-related content using a carefully selected set of energy-specific keywords.

As a first step, I convert the text into a corpus and pre-process it by removing punctuation, numbers, symbols, and common stopwords, as well as applying stemming to normalize word forms. I retain hyphenated words and apply additional custom filtering to exclude high-frequency but semantically neutral terms such as "said," "article," "first," and "new.", among others. From this cleaned corpus, I construct a document-feature matrix (DFM), removing extremely rare and overly common terms to reduce noise and focus on more meaningful vocabulary.

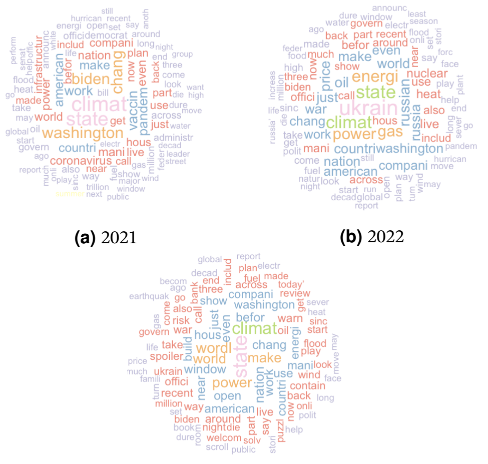
Using this pre-processed text, I generate word clouds for each year—2021, 2022, and 2023—to visually capture the most prominent terms in the energy discourse (see figure 1). In addition, I extract the top 10 most frequent words from each year to identify salient themes and possible changes in the language used around energy topics over time.

Year	Top 10 Words
2021	climat, state, chang, washington, biden, vaccin, work, pandem, american, make
2022	ukrain, state, climat, energi, power, gas, russian, oil, price, russia
2023	state, climat, wordl, power, make, world, nation, even, work, window

The most frequently used terms in energy-related articles reveal how global events have shaped the media narrative over time. In 2021, dominant words include reflect attention to U.S. politics and the COVID-19 recovery.

In 2022, there is a marked shift, with words like "Ukraine," "energy," "climate," "power," "gas," "oil," "price," and "Russian" emerging at the top. This surge in energy-related vocabulary coincides with the outbreak of the Russia-Ukraine war, which triggered widespread concern over energy security, price volatility, and geopolitical tensions.

By 2023, however, only two energy-related terms—"climate" and "power"—remain among the most frequent. This suggests that the heightened media focus on energy policy seen in 2022 began to taper



(a) 2021 **(b)** 2022 **(c)** 2023

Figure 1. Word Clouds

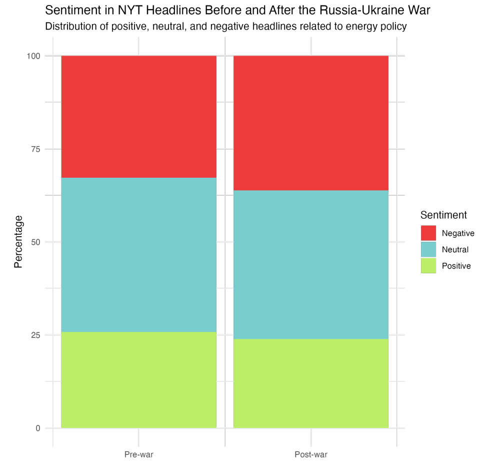


Figure 2. Sentiment Analysis of Energy Policy Discourse Pre and Post War

off, possibly as the initial shock of the crisis settled and coverage diversified into other areas.

3.1. Dictionary-Based Analysis

To examine how the narrative on energy policy evolved in response to the Russia-Ukraine conflict, I filter the energy corpus to retain only those articles that include references to policy-related terms. This results in an energy-policy subcorpus consisting of 6,664 articles that explicitly engage with energy policy.

I pre-process the data and then conduct sentiment analysis on the headlines of these articles using the VADER (Valence Aware Dictionary and sEntiment Reasoner) algorithm. VADER provides a compound sentiment score for each headline, ranging from -1 (most negative) to +1 (most positive). Based on these scores, I classify each headline as positive, neutral, or negative.

Next, I compare the distribution of sentiment before and after the onset of the Russia-Ukraine war (February 24, 2022) to detect shifts in the tone of energy policy coverage.

As shown in Figure 2, there is an increase in negative sentiment following the start of the war: the share of negative headlines rises from approximately 32% pre-war to around 36% post-war. This shift warrants further reflection. One possible explanation is that the war triggered a global energy crisis, marked by rising oil and gas prices and concerns about supply disruptions. Such developments may have led to more frequent use of negative language in headlines, not necessarily indicating dissatisfaction with energy policy itself, but rather reflecting the broader challenges and anxieties surrounding energy in the context of war.

3.2. Latent Dirichlet Allocation

To explore how the thematic focus within energy-related news coverage shifted following the onset of the Russia-Ukraine war, I perform topic modeling on the full energy corpus using the Latent Dirichlet Allocation (LDA) algorithm [1]. LDA is an unsupervised machine learning method that uncovers hidden thematic structures in a large collection of documents. It works by identifying groups of words that frequently appear together across articles and assigning these groups to "topics." Each article is then treated as a mixture of several topics, with different topics weighted differently depending on the content of the article.

The goal of this analysis is to compare the dominant topics in the energy corpus before and after the war, and to identify the most representative words within each topic. In order to determine the optimal number of topics (K) for each period, I rely on two evaluation metrics: the coherence score and the perplexity score. The coherence score measures the semantic similarity among the top words in each topic; in other words, it evaluates how logically coherent and interpretable the topics are. Higher coherence scores indicate more meaningful and human-readable topics. In the coherence results (see figure 6), the first local maximum is observed at K = 5, with further increases at K = 10, 20, and 25. While coherence continues to rise beyond 5 topics, a simpler model with fewer topics is more desirable for interpretability and clarity.

In addition, I examine the perplexity score, which measures how well the model predicts unseen data. Lower perplexity values reflect a better statistical fit of the model. In my analysis, perplexity decreases until it reaches a minimum around K = 22 (see figure 7). Although this suggests that a more granular model with 22 topics may better fit the data, it would also introduce complexity and reduce the ease of interpretation.

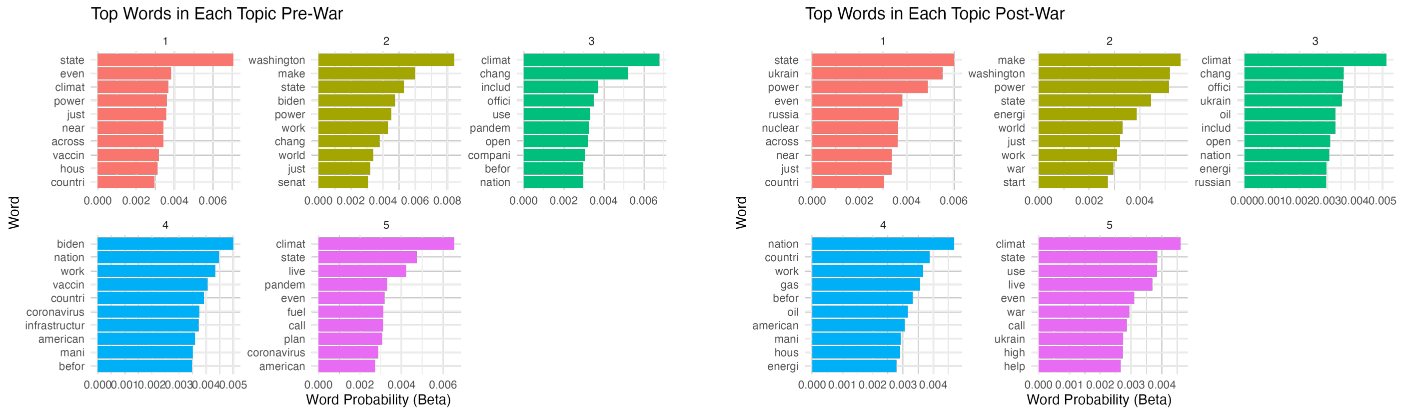
Given these considerations, I choose to work with five topics for both the pre-war and post-war energy corpora. This number strikes a balance between model performance and interpretability, allowing for a meaningful comparison across time periods without becoming overly complex or fragmented.

The two figures below present the results of topic modeling before and after the onset of the war. Each figure displays five distinct topics, along with the top terms most representative of each topic, as identified through the modeling process.

In the pre-war period, the dominant themes appear more diverse and less centered on conflict. In contrast, the post-war topic model reveals a shift toward themes directly connected to the conflict and energy geopolitics. In Topic 1, terms like “ukrain,” “russia,” “nuclear,” and “power” indicate that international conflict and energy security have become more salient. Topic 2 continues the theme of power and state actors, with words such as “washington,” “war,” and “start,” underscoring the geopolitical context. Topic 3 still retains some focus on climate issues but includes war-related terms like “ukrain,” “oil,” and “russian,” suggesting the intersection of environmental topics with the conflict. Topic 4 appears to focus more on national policies and energy concerns, with “gas,” “oil,”. Finally, Topic 5 reflects humanitarian and political aspects, including “war,” “call,” “help,” and “ukrain,” pointing toward international responses and civilian impact.

Based on the comparative analysis, there has been a discernible thematic shift in the discourse after the war began. While pre-war conversations leaned more toward climate, pandemic, and domestic politics, the post-war period is marked by the rise of topics related to conflict, energy, and geopolitical tensions.

In the appendix, you can find the results of the LDA topic modeling broken down by year, offering valuable insight into the dominant themes for each period. In 2021, the topics still heavily reflect the impact of the pandemic. By 2022, the discourse shifts significantly,



(a) Pre-War **(b)** Post-War

Figure 3. Topic Modeling using Latent-Dirichet Allocation

with all five topics centering on energy security, different energy sources, gas prices, and the war. However, this focus appears to fade in 2023—energy is only prominent in Topic 3, suggesting a decline in media attention to energy issues.

3.3. Interrupted Time Series Regression

To better understand how the Russia–Ukraine conflict shaped public discourse on energy policy, I construct an energy salience score that quantifies the prominence of energy-related language in news articles. Specifically, this score is calculated as the proportion of energy-related words—such as "oil," "gas," "fuel," "energy," and "nuclear"—that appear in the headline, abstract, and lead paragraph of each article, relative to the total number of words in those sections. The goal is to capture how central energy topics are in the framing of the news over time.

I begin the analysis by estimating a baseline regression that evaluates whether the war had a statistically significant effect on the energy salience of news articles. This initial model includes a simple binary indicator for whether an article was published after the start of the war. By comparing the average energy scores before and after the war, this approach allows me to detect whether there was a discrete shift in the importance of energy topics following the invasion.

$$EnergyScore = \beta_0 + \beta_1 War \qquad (1)$$

Variable	Energy Score
Intercept	3.2066 ***
	(0.039)
War	-0.05329
	(0.0535)
F(1,9295)= 0.9937 p-value: 0.3189	

Table 1. Regression Results for Model 1

To explore whether the effect of the war evolved over time, I extend the analysis by including a variable that measures the number of days since the beginning of the war. This variable is negative for pre-war articles and positive afterward. Including this continuous time measure enables me to assess whether the salience of energy topics increased, decreased, or remained stable in the weeks and months that followed. It moves beyond the binary war indicator to capture more nuanced dynamics in public discourse.

Lastly, I introduce an interaction term between the war indicator and the time elapsed variable. This specification allows the time trend to differ before and after the war, helping to identify whether the rate of change in energy salience accelerated, decelerated, or shifted direction due to the conflict. For instance, it can reveal whether the post-war rise in energy attention was sustained or simply a temporary spike. Together, these regressions offer a clearer picture of how much the war influenced the prominence of energy policy in media narratives, both in the immediate aftermath and over time.

$$EnergyScore = \beta_0 + \beta_1 War + \beta_2 TimeTrend + \beta_3 Time * War \quad (2)$$

Variable	Energy Score
Intercept	3.334***
	(0.0832)
War	0.0009
	(0.1052)
Time Trend	0.0006
	(0.0003)
War*TimeTrend	-0.0014***
	(0.0004)
F(3,9295)= 5.162 p-value: 0.0015	

Table 2. Regression Results for Model 2

The effect of the war is not statistically significant in either Model 1 or Model 2. However, Model 2 reveals that the impact of the war on the energy score diminishes over time. This pattern aligns with the LDA results discussed above, where the prominence of energy-related topics declines in 2023.

Finally, I conducted a Chow Test to assess whether there was a structural break at the onset of the war. The Chow statistic is 6.005, and the p-value is 0.002, allowing us to reject the null hypothesis of no structural break at the war date. This indicates a significant shift in how energy-related content was presented in news articles following the start of the Russia-Ukraine war. The relationship between time and the energy proportion metric changed fundamentally at this point, suggesting a structural break in the data.

3.4. Structural Topic Modeling

Finally, I apply a Structural Topic Model (STM) [2] to better understand the main themes in energy-related news coverage and how these themes vary with time and in response to major geopolitical events. STM is a topic modeling method that, unlike traditional models like LDA, allows for the inclusion of covariates—such as time or binary indicators like whether an article was published before or after the start of the Russia-Ukraine war. This makes STM especially powerful in my context, as it not only identifies latent topics within the text but also estimates how their prevalence changes over time or between conditions.

I fit the STM with five topics, consistent with my earlier LDA

Table 3. LDA Topics with Highest Probability and FREX Words

Topic #	Label	Highest Probability Words	FREX Words
1	Greenhouse Emissions	energi, countri, global, mani, russian, work, war	carbon, energi, inflat, russian, conflict, greenhous, emiss
2	US Politics	climat, presid, unit, washington, nuclear, nation, biden	biden', republican, presid, democrat, tax, washington, polici
3	Everyday Life	time, home, can, wednesday, hurrican, tuesday, govern	room, grill, recip, meat, ice, ingredi, music
4	Energy Crisis	year, one, chang, oil, day, ukrain, will	oil, fuel, will, day, world', crisi, one
5	Gas Price	state, said, gas, peopl, price, thursday, month	price, peopl, electr, rain, high, area, thursday

Note: Topics extracted using LDA; FREX words help interpret topic uniqueness and frequency.

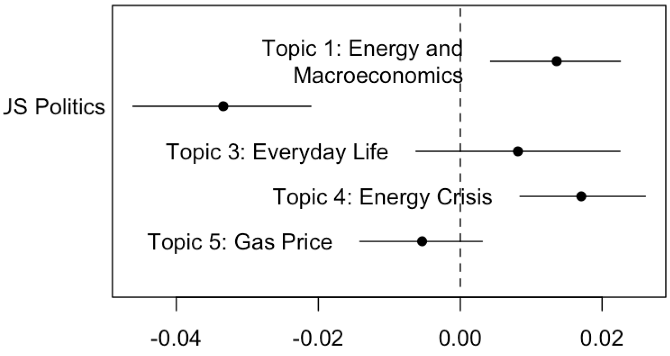


Figure 4. War Effect on Topics

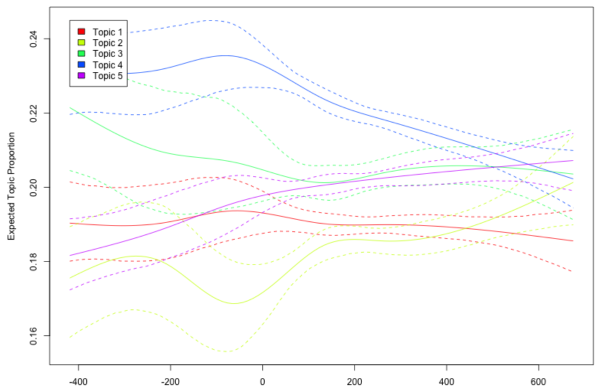


Figure 5. Topics Prevalence Over Time

model. STM itself generates these topics based on word co-occurrence patterns in the text.

To analyze changes in topic prevalence over time, I estimate the effects of time, war, and their interaction. Given the likelihood of non-linear trends in news coverage, I model time using regression splines (specifically natural splines with 5 degrees of freedom).

This choice is motivated by the fact that I cannot make strong assumptions about the underlying functional form of time. News coverage might follow a complex or even polynomial trend, but I do not know a priori what the appropriate degree would be. Including many polynomial terms could overfit or make the model unnecessarily complex. In this sense, regression splines provide a major advantage: they are a nonparametric regression method, which allows the model to flexibly fit non-linear patterns without requiring us to specify the exact nature of the curve. Moreover, splines can determine the optimal number and placement of knots through cross-validation, making them especially suitable for capturing smooth, data-driven variation in topic prevalence over time.

Altogether, this STM framework enables me to track how the discourse on energy—from gas prices to political framing—evolved between 2021 and 2023. This offers a meaningful window into media narratives and potentially broader public attention throughout the period.

Then, I analyze how the prevalence of each topic changed after the war (see figure 4). Positive values indicate that the topic is more prevalent after the war, Negative values inidcate the topic is less prevalent after the war.

Topics 1 and 4—related to energy and macroeconomics, and the energy crisis—become more prevalent after the onset of the war, which aligns with expectations given the central role of energy in geopolitical tensions. In contrast, Topic 5, focused on gas prices, sees a decline in prevalence after the war. This might initially seem counterintuitive, but it could reflect a shift in media framing—from specific price fluctuations to broader narratives about energy security and global markets. Additionally, this trend may be partially driven by temporal dynamics, as overall attention to gas prices was already declining over time.

Topic 2, which captures themes around US politics, also shows a slight decrease in prevalence post-war, while Topic 3, centered on everyday life, becomes marginally more prominent. However, it is important to note that the magnitude of these changes is relatively small. This suggests that although geopolitical events influence media coverage, much of the variation in topic prevalence may be explained by broader, longer-term trends rather than immediate shocks alone.

As a final step, I use spline regressions to model the effect of time on the topic prevalence to evaluate how these topic have been more or less talked temporarily. The following graph (figure 5) shows how the prevalence of these topics has changed as days elapsed from the war. negative numbers mean before the war and positive numbers after the war.

It is notable that several topics exhibit a clear shift around the onset of the war. Topic 1 (Energy and Macroeconomics) shows a subtle increase in attention leading up to the war, followed by a gradual decline as the conflict unfolds. Topic 4 (Energy Crisis) appears to spike just before the war, likely reflecting anticipatory coverage, but its prevalence diminishes over time. Interestingly, Topic 5 (Gas Prices) begins rising prior to the war and continues to gain attention afterward, possibly due to sustained economic impacts. In contrast, Topic 2 (U.S. Politics) experiences a sharp decline during the war period, suggesting a shift in media focus away from domestic political narratives.

4. Discussion

In discussing the shifts in energy policy discourse before and after the Russia-Ukraine war, several key takeaways emerge. First, the increase in negative sentiment following the conflict highlights how external events, like geopolitical crises, can quickly influence public perception and media narratives surrounding energy policy. The shift from a broader focus on climate and economic concerns to an emphasis on energy security and geopolitics suggests a recalibration of priorities in the face of a global crisis.

The gradual decline in the salience of energy issues, as indicated by the interrupted time series analysis, suggests that while the war initially intensified the conversation, the urgency of the issue lessened over time as other global concerns took precedence. This highlights the transient nature of media attention to issues like energy policy, especially when the immediate crisis subsides.

Finally, the results from structural topic modeling underscore how security considerations became more prominent, reflecting a broader shift in how energy policy is framed in times of crisis. The findings suggest that energy policy is not only a matter of economics and environment but also a vital component of national security, which may have long-term implications for how energy strategies are formulated in the future.

In conclusion, the analysis demonstrates how a major geopolitical event like the Russia-Ukraine war can significantly alter the discourse surrounding energy policy, shifting it toward security and geopolitical concerns. It also illustrates the dynamic nature of media coverage, which evolves in response to global crises and shifting priorities.

A. Appendix

A.1. Metrics to choose optimal K

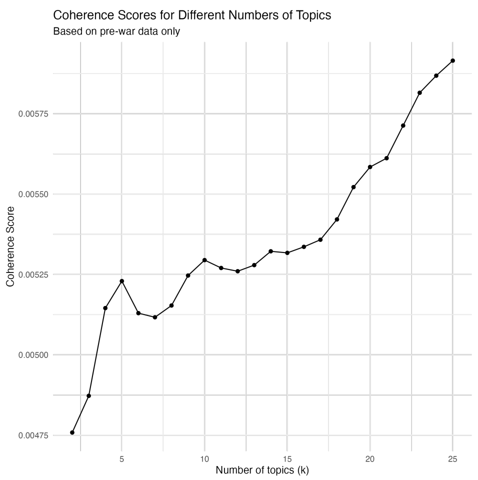


Figure 6. Coherence Score to choose optimal K

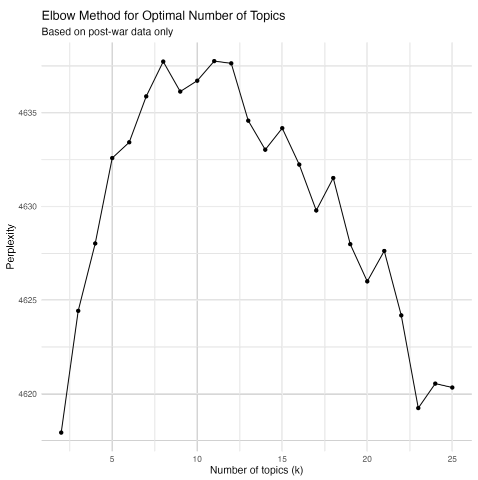


Figure 7. Perplexity Score to choose optimal K

A.2. LDA Topic Modeling by year

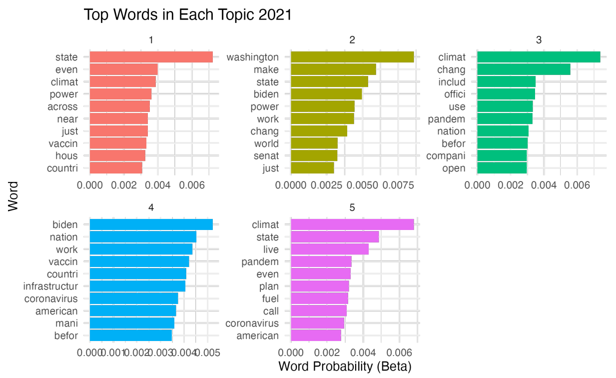


Figure 8. Topic distribution, 2021

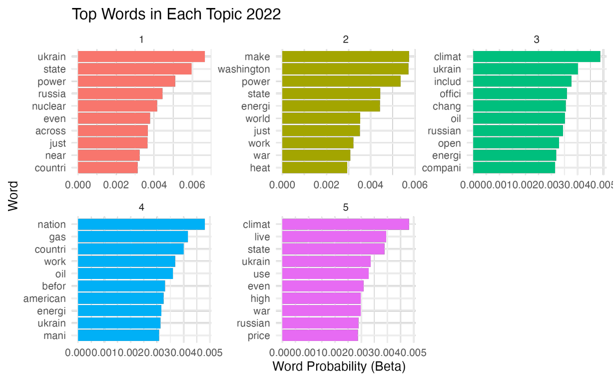


Figure 9. Topic distribution, 2022

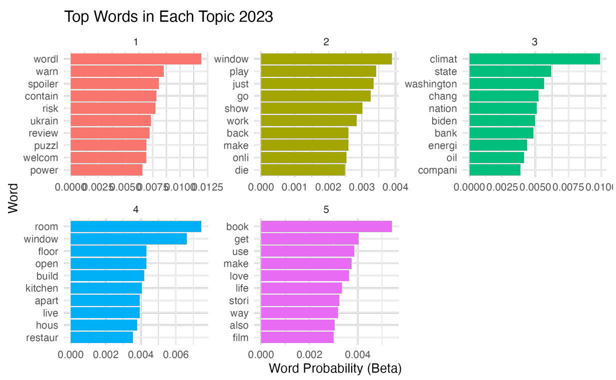


Figure 10. Topic distribution, 2023

References

[1] *Latent Dirichlet Allocation*. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[2] *R Package for Structural Topic Models*. [Online]. Available: https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf.

[3] *Topic Modelling Three Decades of Climate Change News in Denmark*. [Online]. Available: https://ssrn.com/abstract=4513921.