Universidade de Brasília - UnB

Faculdade de Economia, Administração,

Contabilidade e Gestão de Políticas Públicas - FACE

Departamento de Economia - ECO

Resonant Journalism in the Russo-Ukrainian War: A Topic Modeling Approach to Key Point Detection

Arthur Gomes Nery

Brasília, Brasil

2023

Arthur Gomes Nery

Resonant Journalism in the Russo-Ukrainian War: A Topic Modeling Approach to Key Point Detection

Monografia apresentada como requisito parcial à obtenção do título de Bacharel em Ciências Econômicas pela Universidade de Brasília.

Universidade de Brasília - UnB

Faculdade de Economia, Administração,

Contabilidade e Gestão de Políticas Públicas - FACE

Departamento de Economia - ECO

Orientador: Prof. Daniel Oliveira Cajueiro, Dr.

Brasília, Brasil

2023

Arthur Gomes Nery
Resonant Journalism in the Russo-Ukrainian War: A Topic Modeling Approach to Key Point Detection/ Arthur Gomes Nery. – Brasília, Brasil, 2023-
35p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Daniel Oliveira Cajueiro, Dr.

Monografia (Graduação) – Universidade de Brasília - UnB
Faculdade de Economia, Administração,
Contabilidade e Gestão de Políticas Públicas - FACE
Departamento de Economia - ECO, 2023.

1. Alocação Latente de Dirichlet. 2. Detecção de pontos-chave. 3. Fontes de Notícias 4. Guerra Russo-Ucraniana. 5. Análise de Conflitos. II. Universidade de Brasília. III. Faculdade de Administração, Contabilidade e Economia - FACE. IV. Departamento de Economia IV. Resonant Journalism in the Russo-Ukrainian War: A Topic Modeling Approach to Key Point Detection

Arthur Gomes Nery

Resonant Journalism in the Russo-Ukrainian War: A Topic Modeling Approach to Key Point Detection

Monografia apresentada como requisito parcial à obtenção do título de Bacharel em Ciências Econômicas pela Universidade de Brasília.

Trabalho aprovado. Brasília, Brasil, 13 de fevereiro de 2023:

Prof. Daniel Oliveira Cajueiro, Dr.
Orientador

Prof. Victor Rafael Rezende Celestino, Dr.
Convidado 1

Brasília, Brasil
2023

Resumo

A colocação de momentos especiais em uma sequência temporal de eventos pode melhorar a compreensão de sistemas sociais complexos. No contexto de uma guerra, a identificação de eventos críticos, tendências e padrões pode revelar conhecimentos aprofundados sobre as ações e motivações dos atores envolvidos, bem como informar a tomada de decisões. Este estudo visa investigar se é possível detectar pontos-chave de um evento global através da análise de dados textuais não estruturados de fontes de notícias. Para isso, coletamos 61.165 histórias de 11 fontes de notícias e aplicamos o modelo de tópicos alocação latente de Dirichlet (LDA) para quantificar e descrever vários eventos-chave da guerra Russo-Ucraniana após seu agravamento em fevereiro de 2022. Além disso, investigamos as flutuações na cobertura de tópicos para cada fonte de notícias em comparação com as tendências do sistema usando a Divergência de Kullback-Leibler (KLD), evidenciando as diferenças no potencial de criação de tendências de cada fonte. Os resultados mostram que o LDA é capaz de identificar eventos-chave cobertos por todas as fontes e que o KLD pode ser aplicado às distribuições de tópicos para evidenciar que algumas fontes de notícia exibem desvios significantes do potencial médio de estabelecimento de tendências do sistema completo. Este estudo evidencia o potencial de dados textuais não estruturados na descoberta de eventos importantes em situações de conflito.

Palavras-chave: Guerra Russo-Ucraniana, Detecção de pontos-chave, Alocação Latente de Dirichlet, Análise de conflitos, Fontes de notícias

Abstract

The placement of special moments in a temporal sequence of events may improve the understanding of complex social systems. In the context of war, the identification of critical events, trends and patterns can reveal deep insights regarding the actions and motivations of the involved actors, as well as support informed decision-making. This study aims to investigate if key points of a major global event can be detected through the analysis of news outlets' unstructured textual data. To do so, we collect 61,165 stories from 11 major news sources and apply latent Dirichlet allocation (LDA) topic modeling to quantify and describe several key events of the Russo-Ukrainian war after its escalation in February 2022. Moreover, we investigate the fluctuations in topic coverage for each news source when compared to the tendencies of the entire system using Kullback-Leibler Divergence (KLD), bringing light to differences in the trendsetting potential of each outlet. The results show that LDA is capable of identifying key events covered across all sources, and KLD may be applied to topic distributions to show that some news outlets exhibit significant deviation from the average expected trendsetting potential of the entire system. This research sheds light on the potential of unstructured text data in uncovering important events in conflict situations.

Keywords: Russo-Ukrainian war, Key point detection, Latent Dirichlet Allocation, Conflict analysis, News sources

List of Figures

List of Tables

List of abbreviations and acronyms

IAEA	International Atomic Energy Agency
KLD	Kullback-Leibler Divergence
LDA	Latent Dirichlet Allocation
NATO	North Atlantic Treaty Organization
SWIFT	Society for Worldwide Interbank Financial Telecommunication

Contents

1 Introduction

The placement of special moments in a temporal sequence of events may improve the understanding of social complex systems. Time is a particularly important dimension that allows for the identification of processes that are fundamental to modeling the evolution of social systems such as path dependence (Pierson, 2004) and positive feedback or self-reinforcement. Furthermore, the temporal ordering of the events is of ultimate importance to understanding the outcomes of a given incident.

Motivated by the relevance of the temporal sequence of facts in a given conflict incident, in this work, we propose a method to identify and characterize the critical moments in the ongoing Russia-Ukraine conflict using public news and natural language processing. In our work, a key moment in this incident is an event that is highly covered by multiple news sources within a limited time frame, presenting a peak in reporting. We apply latent Dirichlet allocation (Blei et al., 2003) to classify documents in a dataset comprised of news stories regarding the Russo-Ukrainian conflict from multiple news sources. We then quantify the topic modeling results and identify peaks in reporting connected to key events in the conflict, from the domains of political happenings, battle advances, artillery discussions, and supply chain shocks.

Our data comes from 11 international news outlets: ABC, Associated Press, CBS, CNN, Daily Mail, Express, Fox, The Guardian, Mirror, The New York Times, and Reuters. We choose these sources due to their ubiquity and scale, allowing for a good grasp of the general direction of news within English-speaking nations. The resulting corpus is comprised of 61,165 articles mentioning Ukraine and Russia from July 2021 to December 2022.

Moreover, we apply the measures of resonance, novelty, and transience introduced by Barron et al. (2018) to the topic distributions found for the documents in the corpus to investigate differences in the average trendsetting potential of each source. These measures rely on the Kullback-Leibler Divergence (KLD) as a notion of "surprise" that measures the statistical distance between the topic probability distributions of two subsequent documents. We can use the average surprise to measure the distance between one news story and an interval of neighboring stories. "Novelty" represents the comparison between a given story and previous ones, while "transience" relates to future stories. With these concepts in hand, we can define "resonance" as the difference between novelty and transience. Therefore, a story considered resonant presented a topic mixture that was novel when compared to the past time frame and appeared again in the future, showing a high trendsetting potential.

The limitations of this study should be considered when interpreting the results.

Firstly, our findings are limited to the specific 11 news sources herein analyzed, and thus, the peaks found may not reflect the full picture of the events. Secondly, the key points we found using our methodology require confirmation from another source of key event data. This is a major limitation as the peaks found are solely based on the news sources analyzed. Additionally, our methodology requires supervision in multiple steps, such as hyperparameter specification, topic selection, and interpretation of news stories. This need for human intervention may introduce bias and limit the robustness of the results.

The remainder of this study is structured as follows: In Chapter 2, we will conduct a comprehensive review of previous studies related to our application, which will provide a background for our analysis and set the foundation for the methodology and results sections. In Chapter 3, we will outline the data and methodology used for this study, including latent Dirichlet allocation (LDA) topic modeling and Kullback-Leibler Divergence (KLD) analysis using the measures of Novelty, Transience and Resonance. Chapter 4 will present the results of our analysis, including the identification and description of the key events found across the 11 different news sources, as well as the analysis of differences in topic coverage for each news source, investigating their trendsetting potential. Finally, in Chapter 5, we will provide a conclusion that summarizes the findings of this study and offers suggestions for future research.

2 Related Work

This section establishes an overview of the previous literature regarding applications of topic modeling using Latent Dirichlet Allocation and of Kullback-Leibler Divergence in the field of information retrieval for different domains of textual analysis, as well as different means of turning point identification in time series.

Our application naturally relates to other works that propose statistical approaches to identifying turning points in time series of data (Western; Kleykamp, 2004; Spirling, 2007b; Spirling, 2007a; Li; Lund, 2015; Ruggieri; Antonellis, 2016; Park; Yamauchi, 2022). The work by Spirling (2007b), concerned with detecting turning points in the Iraq conflict, is particularly interesting. It examines causality data with reversible-jump Markov chain Monte Carlo techniques and it finds evidence of four change points: The capture of Saddam Hussein; (2) The installation of the Iraqi Interim Government; (3) The legislative elections; (4) the assumptions of military responsibilities by the Iraqi government. After all of these breaks, the number of causalities increases. As is ours, the objective of these papers is to determine points in time that allow the data-backed claim that extraordinary changes took place. However, the main difference presented by our approach is that we use unstructured data that comes from news stories. Thus, the advantage of our natural language processing-based approach is that we can capture the perception of the war from different viewpoints (Cajueiro et al., 2021). On the other hand, an obvious disadvantage is that the news outlets may show political bias or conflicts of interest (Archer; Clinton, 2018; Garz; Rickardsson, 2022).

Applications of the well-established LDA model (Blei et al., 2003; Blei, 2012) are plentiful in social science. As compiled by Mohr and Bogdanov (2013), the suite of articles published in a special issue of *Poetics* (Vol. 41, no. 6) (DiMaggio et al., 2013; McFarland et al., 2013; Miller, 2013; Bonilla; Grimmer, 2013; Mohr et al., 2013; Marshall, 2013; Tangherlini; Leonard, 2013; Jockers; Mimno, 2013) display a variety of applications of this method. Particularly relevant to the domain of the present work, Bonilla and Grimmer (2013) leveraged topic modeling to track the attention directed by major news outlets towards terror threats after elevations of the US government's color-coded alert system. The authors find that the alerts exert brief yet substantial influence on the media's agenda, and the effects of these attention shifts on the public are muted. Miller (2013) modeled typologies of violence held by government administrators during the Qing dynasty of China by modeling topics on the Qing Veritable Records to attain patterns of "crime rates" and provide insight into how different epochs understood crime, rebellion, and unrest. For another application in political science, Tsur et al. (2015) paired LDA topic modeling with autoregressive-distributed-lag models to process four years of public statements issued by

members of the U.S. Congress, investigating the differences between the framing strategies of Democrats and Republicans. In the grand scheme of information retrieval, these are only a few of the applications of latent Dirichlet allocation available. To the best of our knowledge, our work is a novel application of LDA in its use for tracking key events in conflicts.

Concerning the analysis of the trendsetting potential of sources in a time series of documents, this work directly relies on the application of LDA-based topic modeling introduced by Barron et al. (2018). The authors analyze a corpus of over 40,000 speeches from debates held during the first National Constituent Assembly of the French Revolution, applying the KLD-based (Kullback; Leibler, 1951) measures of novelty (how unexpected a speech's patterns are, given past speeches), transience (the extent to which those patterns fade or persist in future speeches), and resonance (the imbalance between novelty and transience: a speech that has both high novelty and low transience is considered resonant) to investigate trendsetting characteristics of particular groups or individuals. The authors found that highly novel speeches tended to be highly transient. As for groups in the political spectrum, left-wing representatives collectively produce more innovative speech patterns when compared to right-wing representatives, who sustain low transience and maintained previous topics and patterns. On an individual level, some of the speakers stand out as highly innovative trendsetters (e.g. Robespierre, a notorious Jacobin with patterns of high resonance), some show high novelty but failed to establish resonance for future debates (speakers such as Armand-Gaston Camus and Théodore Vernier) and others maintain and discuss the *status-quo* (prominent political conservatives such as Jean-Siffrein Maury and Jacques de Cazalès), showing low novelty and high resonance.

The applicability of the method introduced by Barron et al. (2018) is evident by the works that succeed it. Correia and Mueller (2022) apply the method to a corpus of over 1,600 papers from the Brazilian Association of Graduate Programs in Economics (ANPEC) meeting annals to track patterns of innovation and influence within Brazilian economic research. As seen with debates from the French Revolution, the authors verify that novelty is highly correlated with transience for this application. However, of the few resonant research ideas, the particularly novel ones show a higher impact in the context of ANPEC. Nielbo et al. (2020) pair the methodology with nonlinear adaptive filtering and adaptive fractal analysis to present a novel approach to trend estimation on social media platforms. Particularly, the authors introduce "trend reservoirs", signals that display trend potential according to their relationship between novelty and resonance as well as their degree of persistence, when compared to a random baseline.

Degaetano-Ortlieb and Teich (2018) leverage KLD for textual analysis of the corpus of scientific publications of the Royal Society of London to detect features involved in diachronic linguistic change in scientific writing. The authors conclude via sample analysis

that waves of lexical expansion tend to be followed by periods of consolidation, in which grammatical variation is reduced. This represents the effort to balance expressivity and communicative efficiency in papers, guaranteeing that conveyance remains successful amidst language use changes. Lastly, other relevant uses of KLD in the domain of textual analysis are the applications of Jing et al. (2019) and Murdock et al. (2017). The former estimates novelty in a corpus of fanfiction and investigates the relationship between innovation and popularity within the fanfiction community. The latter is an analysis of the corpus of Charles Darwin's records of reading choices, using topic models to quantify his local (text-to-text) and global (text-to-past) reading preferences and investigate his decisions between the exploitation of prior knowledge in a given area and further exploration of different knowledge.

3 Method and Data

The methodology section of this study presents the models and peak detection algorithm used for the analysis of the data, as well as the steps taken to collect and pre-process the corpus. Our Python implementation of the methods described in the following subsections is available on <https://github.com/r2nery/ukraine-media>.

3.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) (Blei et al., 2003; Blei, 2012) is a generative probabilistic model of a given corpus. It takes a vocabulary size w and a number of topics z as hyperparameters, and each document from the corpus is assumed to be generated from Dirichlet distributions of documents on the $(z-1)$-simplex of topics and of topic distributions on the $(w-1)$-simplex of words. Each document is represented by a multinomial distribution of topics, and each topic by a multinomial distribution of words.

The generative process is comprised of three steps: first, a Dirichlet distribution α of documents over the $(n-1)$-simplex of topics is randomly chosen. Then, for each word (selected from the multinomial distribution of words within the topic θ) in a given document, a topic is randomly chosen from the Dirichlet distribution η of topics over the $(w-1)$-simplex of words. Finally, a word is chosen from its corresponding multinomial distribution of words β over the fixed vocabulary. This process is reiterated, and the algorithm can then estimate latent topics within the corpus by comparing the generated documents to the original corpus. By the end of the generative process, each document in the corpus should exhibit all topics in different proportions based on the optimal Dirichlet distributions.

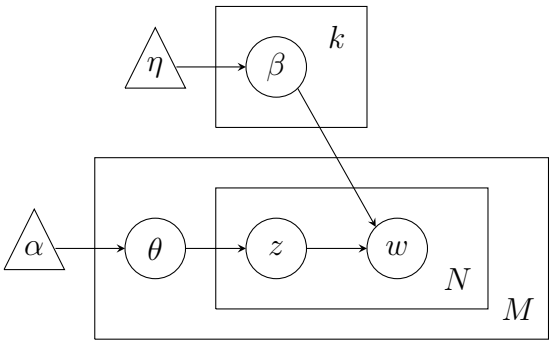

Figure 1 – LDA model. η: Dirichlet distribution of topics over the vocabulary, k: Topics, β: Multinomial distributions of words, α: Dirichlet distribution of documents over topics, M: Corpus, N: Document, θ: Multinomial distribution of topics, z: List of topics drawn from θ, w: List of words comprising a generated document.

We apply latent Dirichlet allocation to our corpus using a high number of topics ($z = 200$) as a hyperparameter. This allows the model to identify topics with high precision and granularity, a setting that is ideal in order to pinpoint specific happenings within the time series. A lower number of topics would result in broader classifications.

3.2 Kullback-Leibler Divergence

The Kullback-Leibler Divergence (Kullback; Leibler, 1951), also known as relative entropy, is a measure of information loss when an observed probability distribution p is estimated using a theoretical distribution q:

$$\text{KLD } (p|q) = \sum_{i=1}^{k} p_i \log_2 \frac{p_i}{q_i} \tag{3.1}$$

Therefore, it's a divergence between probability distributions: $\text{KLD}(p|q = p)$ would yield a result of zero and the KLD of two vastly different distributions would be high, signifying a great loss of information due to misspecification. It's important to note that the Kullback-Leibler Divergence is asymmetric, and therefore cannot be considered a distance. It is also unbounded.

The methods described by Barron et al. (2018) pair LDA topic modeling with relative entropy by investigating the distance between the measured topic probability distributions of different documents in a temporal sequence. From an information retrieval perspective, relative entropy may be interpreted as a measure of "surprise" when one document is expected and another is observed. Given an LDA-generated set of probability distributions $p^{(j)} = (p_1^{(j)}, p_2^{(j)}, \ldots, p_k^{(j)})$, where j indexes chronological order and k indexes topics, the measure of novelty $\mathcal{N}_w(j)$ of the j-th document is defined by the average KLD between itself and an interval of w past documents:

$$\text{KLD } (p^{(j)}|p^{(j-1)}) = \sum_{i=1}^{k} p_i^{(j)} \log_2 \frac{p_i^{(j)}}{p_i^{(j-1)}} \tag{3.2}$$

$$\mathcal{N}_w(j) = \frac{1}{w} \sum_{d=1}^{w} \text{KLD } (p^{(j)}|p^{(j-1)}) \tag{3.3}$$

When applied to an interval of future documents, the measure is defined as transience. The measure of resonance is defined as the difference between novelty and transience:

$$\mathcal{T}_w(j) = \frac{1}{w} \sum_{d=1}^{w} \text{KLD } (p^{(j)}|p^{(j+1)}) \tag{3.4}$$

$$\mathcal{R}_w(j) = \mathcal{N}_w(j) - \mathcal{T}_w(j) \tag{3.5}$$

We may interpret the resonance of a document in a corpus of news stories as an indicator of a novel subject that is capable of influencing the general direction of outlets,

being written about again in the future. A high average resonance for a given source when compared to the system average is an indicator of its trendsetting potential in reporting the Russo-Ukrainian conflict. Given that w is a document count, the exploration of multiple intervals reveals insights about the persistence of tendencies within the time series. We apply $\mathcal{N}$,$\mathcal{T}$ and $\mathcal{R}$ to several timescales w between 1 and 1000 to investigate these tendencies and use $w = 10$ as the main timeframe throughout this work.

3.3 Peak Detection Algorithm

One of the challenges we face in this work is the selection of distinct peaks of interest in reporting amidst the noise of all the news stories collected. For this purpose, we use the z-Scores-based robust peak detection algorithm by Brakel (2014). This algorithm is particularly robust due to its use of a separate moving mean and deviation, allowing the detection threshold to remain uncorrupted throughout the time series, especially for real-time data. The algorithm takes the lag of the moving window, the z-score (the number of standard deviations by which a data point is above or below the mean) at which the algorithm signals, and the influence of new signals on the mean and standard deviation as inputs.

Algorithm 1 z-Scores-Based Robust Peak Detection Algorithm (Brakel, 2014)

Require: $len(\mathtt{y}) \geq \mathtt{lag} + 2$
  $\mathtt{signals} \leftarrow [0, \ldots, 0]$ of length $= len(\mathtt{y})$ ▷ Initialize signal results
  $\mathtt{filteredY} \leftarrow [\mathtt{y}[1], \ldots, \mathtt{y}[lag]]$ ▷ Initialize filtered series
  $\mathtt{avgFilter} \leftarrow$ Null ▷ Initialize average filter
  $\mathtt{stdFilter} \leftarrow$ Null ▷ Initialize std. filter
  $\mathtt{avgFilter}[\mathtt{lag}] \leftarrow mean(\mathtt{y}[1, \ldots, \mathtt{y}[lag])$ ▷ Initialize first value average
  $\mathtt{stdFilter}[\mathtt{lag}] \leftarrow std(\mathtt{y}[1, \ldots, \mathtt{y}[lag])$ ▷ Initialize first value std.
  for $\mathtt{i} \leftarrow \mathtt{lag} + 1$ to $len(\mathtt{y}) - 1$ **do**
    if $abs(\mathtt{y}[\mathtt{i}] - \mathtt{avgFilter}[\mathtt{i} - 1]) > \mathtt{threshold} * \mathtt{stdFilter}[\mathtt{i} - 1]$ **then**
      if $\mathtt{y}[\mathtt{i}] > \mathtt{avgFilter}[\mathtt{i} - 1]$ **then**
        $\mathtt{signals}[\mathtt{i}] \leftarrow 1$ ▷ Positive signal
      else
        $\mathtt{signals}[\mathtt{i}] \leftarrow -1$ ▷ Negative signal
      end if
      $\mathtt{filteredY}[\mathtt{i}] \leftarrow \mathtt{influence} * \mathtt{y}[\mathtt{i}] + (1 - \mathtt{influence}) * \mathtt{filteredY}[\mathtt{i} - 1]$
    else
      $\mathtt{signals}[\mathtt{i}] \leftarrow 0$ ▷ No signal
      $\mathtt{filteredY}[\mathtt{i}] \leftarrow \mathtt{y}[\mathtt{i}]$
    end if
    $\mathtt{avgFilter}[\mathtt{i}] \leftarrow mean(\mathtt{filteredY}[\mathtt{i} - \mathtt{lag} + 1], \ldots, \mathtt{filteredY}[\mathtt{i}])$
    $\mathtt{stdFilter}[\mathtt{i}] \leftarrow std(\mathtt{filteredY}[\mathtt{i} - \mathtt{lag} + 1], \ldots, \mathtt{filteredY}[\mathtt{i}])$
  end for

3.4 News Stories Corpus

We construct our corpus by collecting the body text of articles published on the websites of 11 news outlets (ABC, Associated Press, CBS, CNN, Daily Mail, Express, Fox, The Guardian, Mirror, The New York Times, and Reuters) from July 1 2021 to November 31 2022. This date window allows for the recognition of topics that existed since before the aggravation of the Russo-Ukrainian conflict in March 2022, as well as topics that surfaced after this event. However, the analyses done in this work will emphasize the year 2022. In order to only collect stories that somehow relate to the conflict, we conduct queries within the websites with the terms "Russia" and "Ukraine". We conduct queries for each term separately and discarded duplicate results. This is the case for all websites except for Mirror, from which we collect all stories in the dedicated "Russia-Ukraine War" section. Due to this, the Mirror time series begins in February 2022. This results in a corpus comprised of 61,165 articles as described in Table 1:

The choice of the search terms "Russia" and "Ukraine" also results in the inclusion of stories with little relation to the conflict, such as stories covering the Olympics. We found that filtering out these subjects during preprocessing was unnecessary due to the sorting nature of LDA with a high number of topics. However, we conduct stopword removal as well as the removal of URLs prior to count vectorization utilizing the 10.000 most common words in the corpus before inputting the data into the LDA model. As each document within the corpus is labeled with its source, we may observe the measures of $\mathcal{N}$, $\mathcal{T}$ and $\mathcal{R}$ introduced in Chapter 3.2 throughout the entire corpus, as well as filter results for specific sources to identify trendsetting characteristics and compare different outlets. We choose the inter-source approach for the objective of mapping key events within the conflict (the common subject between all sources) utilizing LDA alone.

Source	Story Count	Avg. Text Length	Avg. Word Count
Reuters	16015	2137	345
Express	10172	3135	514
DailyMail	7953	9007	1496
AP	6019	6621	1061
Guardian	4928	5216	856
Fox	4526	3530	575
CBS	3408	3664	607
NYT	2548	6847	1121
CNN	2469	5702	932
Mirror	2136	3397	564
ABC	991	5060	825
Sum	61165	54316	8896
Mean	5560	4938	809

Table 1 – Descriptive statistics - News stories corpus

4 Results

The results section of this study presents the findings of our unstructured data analysis for key event retrieval and trendsetting potential identification. The section begins by presenting the quantitative results of the topics found using latent Dirichlet allocation (LDA) topic modeling as presented in Chapter 3.1. These results provide a detailed understanding of the main topics covered in the news outlets after the escalation of the Russo-Ukrainian war in February 2022. We then provide a brief description of the main peaks identified using the robust z-score peak detection algorithm presented in Chapter 4.2. Finally, the trendsetting potential results found using Kullback-Leibler Divergence (KLD) as presented in section 3.2 are presented. These results help to identify the sources that are most influential in shaping the narrative around key events in the conflict.

4.1 Topics

After LDA iteration, the model divides the news stories that comprise the full corpus into 200 topics. Figure 2 displays the daily counts for each topic throughout 2022. This unfiltered representation is quite noisy, and this fact warrants the application of a method of peak detection so we can better identify the key events in the time series.

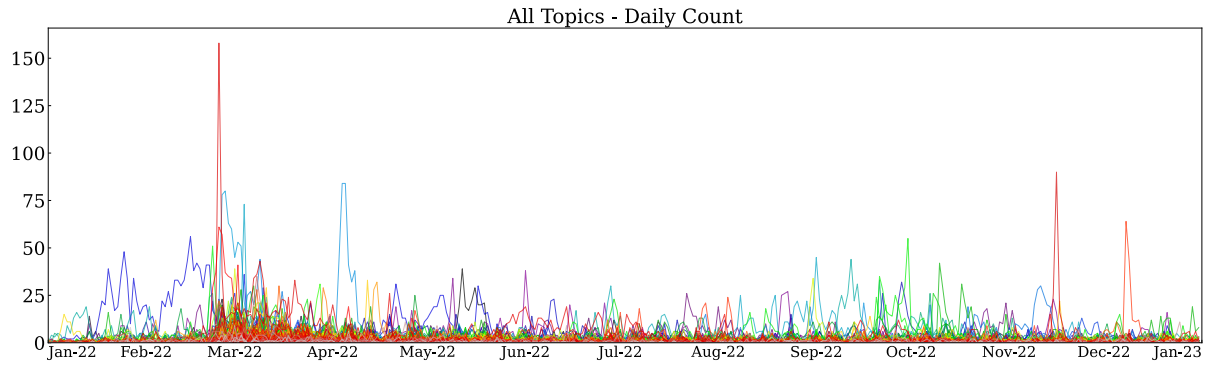

Figure 2 – 200 topics within the news story corpus, unfiltered

Table 2 shows that LDA topic modeling was effective in capturing specific events in the corpus through the identification of the likeliest words for each topic. Some events are particularly identifiable by their distinct set of likeliest words. For example, Topic 185 relates to the event of a missile strike in Polish territory on November 15th, 2022, and Topic 166 refers to the widely-covered prisoner swap that freed Women's National Basketball Association player Brittney Griner, detained in Russia on February 17, 2022. Some topics, such as Topic 39, show highly likely words that are less distinctive on their

own. However, by observing the placement of such topics in the time series and analyzing individual stories from their peaks, these events can be identified precisely.

Topic 180 (Peak = 158)	Topic 185 (Peak = 90)	Topic 56 (Peak = 84)	Topic 61 (Peak = 80)	Topic 65 (Peak = 73)
ukraine	poland	bucha	kyiv	nuclear
putin	polish	russian	russian	plant
attack	warsaw	bodies	city	power
ukrainian	said	civilians	forces	zaporizhzhia
military	ukraine	kyiv	ukrainian	ukraine
russian	missile	war	capital	russian
country	duda	crimes	troops	said
russia	ukrainian	said	ukraine	ukrainian
said	nato	ukrainian	said	shelling
thursday	russia	mass	fighting	chernobyl

Topic 166 (Peak = 64)	Topic 169 (Peak = 61)	Topic 39 (Peak = 56)	Topic 115 (Peak = 55)	Topic 72 (Peak = 44)
griner	sanctions	ukraine	russia	ukrainian
russia	russia	russia	ukraine	russian
russian	russian	russian	regions	kharkiv
whelan	banks	troops	russian	forces
brittney	financial	border	donetsk	region
said	putin	putin	luhansk	troops
release	economic	invasion	territory	ukraine
detained	economy	military	people	city
reed	swift	eastern	ukrainian	territory
moscow	measures	invade	putin	said

Table 2 – 10 likeliest words of topics that achieved the highest peaks in daily counts.

4.2 Peak Detection

Before conducting peak detection, we remove all topics with a maximum daily count lower than 25 stories in order to isolate very relevant peaks. This filter represents the removal of 39,029 stories from the corpus (63.8%) and also eliminates topics with little relation to the conflict and/or very specific topics with little coverage throughout the news outlets. This threshold is subjective: a more lenient cutoff would result in a larger selection of key events, and this choice is entirely dependent on the particular characteristics of the major event being studied. Thereafter, we apply the z-scores-based robust peak detection algorithm by Brakel (2014) (Chapter 4.2). For our purpose, the "influence" parameter of the algorithm was disabled and both the lag and the z-score threshold were modified manually for each topic, given the particular characteristics of each topic between highly isolated peaks and somewhat homogenous peaking. Figure 3 displays the algorithm's performance for topics from the corpus with different peaking characteristics.

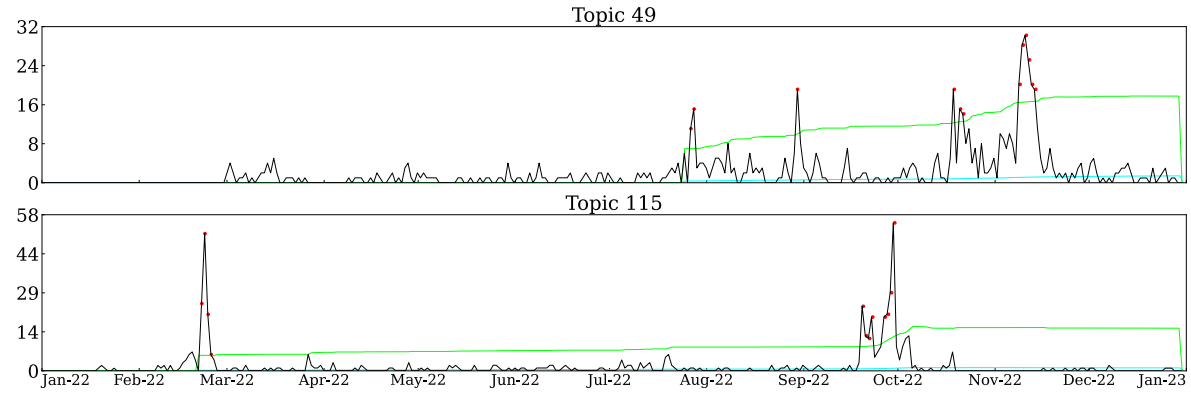

Figure 3 – z-Scores-based robust peak detection algorithm. In cyan, is the moving average, in green, is the detection threshold, and in red, are the detected peaks.

After filtering for documents classified by the algorithm as part of a peak day, we manually divide the resulting topics into 4 categories concerning their subjects for clarity. "Advances, Battles, Attacks and Effects of War" refers to the reporting of skirmishes, shellings, killings, and the direct effects of those factors, such as architectural destruction, casualty counts, victims of forced displacement and civil conflict survivors. "Politics" refers to stories reporting political addresses by national leaders as well as political measures. "Prices, Supply Chain and Sanctions" refers to topics covering inflation, supply of major commodities such as natural gas, grain, and oil, corporate retaliation, stocks, and economic sanctions of any nature. Lastly, "Artillery and Troops" refers to stories reporting troop counts, developments regarding weapons, vehicles, and any other forms of artillery involved in the conflict. Figure 4 shows the result of our peak detection methodology.

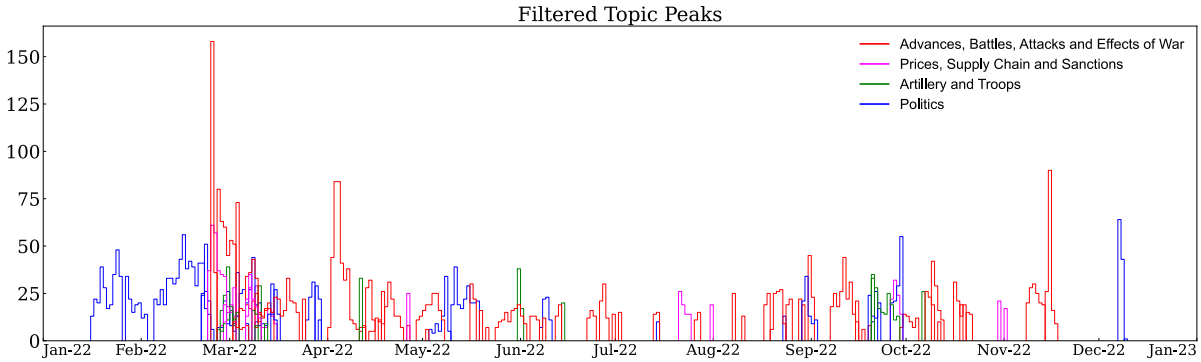

Figure 4 – Topics after peak detection and reclassification

4.2.1 Advances, Battles, Attacks and Effects of War

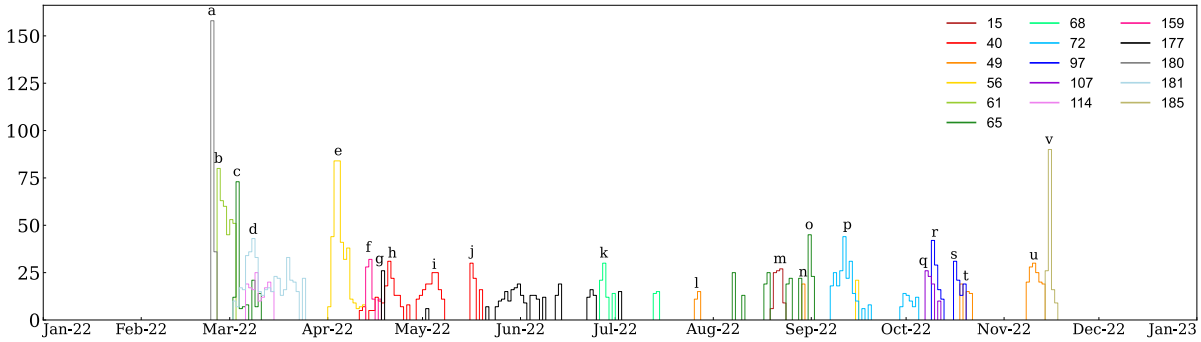

Figure 5 – Advances, Battles, Attacks and Effects of War - Peaks

Peak	Date	Topic	Count	Event
a	Feb 24	180	158	Invasion day, martial law
b	Feb 26	61	80	Russia targets Kyiv
c	Mar 4	65	73	Russia attacks Zaporizhizhia nuclear plant
d	Mar 9	181	43	Russia attacks Mariupol hospital
e	Apr 5	56	85	Russia attacks Bucha, major civilian casualties
f	Apr 15	159	32	Ukraine sunks russian flagship
g	Apr 19	177	24	Russia launches offensive in Donbas
h	Apr 21	40	31	Putin calls off plan to storm Azovstal steel plant in Mariupol
i	May 5	40	25	Russia attacks Azovstal steel plant, Civilian evacuation
j	May 17	40	30	Surrendered Ukrainian fighters leave Azovstal steel plant
k	Jun 28	68	30	Russia attacks Ukrainian mall
l	Jul 27	49	15	Ukraine's begins counterattack in Kherson
m	Aug 23	15	27	Daughter of Alexander Dugin, Putin ally, killed in car attack
n	Aug 30	49	19	Uraine's counteroffensive increases in Kherson
o	Sep 1	65	45	IAEA convoy inspects Zaporizhzhia plant
p	Sep 12	72	44	Major Ukrainian retakes in Kherson and Kharkiv
q	Oct 8	107	26	Crimean bridge bombing
r	Oct 10	97	42	Russia retaliates Crimea bridge bombing with multiple missile strikes
s	Oct 17	97	31	Russian attacks with Iranian drones
t	Oct 19	49	19	Putin declares martial law in Donetsk, Luhansk, Kherson and Zaporizhzhia
u	Nov 11	49	30	Russia retreats from Kherson
v	Nov 16	185	90	Missile kills two in Poland, near border with Ukraine

Table 3 – Advances, Battles, Attacks and Effects of War - Description of the events found

"Advances, Battles, Attacks and Effects of War" is the largest category among all peaks found. The topics within it contain 8,715 news stories, 14% of the entire corpus. This category also contains the single largest peak in daily story counts, which happened on the first day of the Russian invasion of Ukraine (February 24th) (Schwirtz et al., 2022). However, it is important to note that the reclassification of topics found is a subjective task. For example, Peak **s** (Topic 97) contains stories discussing the Russian attacks using Iranian drones (Barnes, 2022). While some stories focus on the aftermath of the attacks, many reports also focus on the characteristics of the drones and their provenance (which could be considered an Artillery discussion).

While some peaks are very distinct and singular, such as the missile strike to Polish territory seen in peak **v** (Jakes, 2022), others represent recurring subjects capable of creating peaks on multiple dates. Topic 40 is an example of this, containing multiple surges in reporting the major battle for the Azovstal steel plant that took place between

April and June 2022 (**h**, **i**, **j**) (Zinets, 2022). The so-called "Bucha Massacre" (Rankin; Boffey, 2022) (**e**) is another major event in the series, characterized by the mass murder of Ukrainian civilians and prisoners of war by the Russian Armed Forces during the fight for and occupation of the Ukrainian city of Bucha. Even though this attack happened in March, the evidence became available to the public after the withdrawal of Russian forces from the site on April 1, when a peak in stories began to form. Topic 65 showed two major peaks (**c**, **o**): The first reported the Russian attacks on the Zaporizhzhia power plant (Borger; Henley, 2022), which raised worries regarding nuclear contamination across the affected area, and the second reported the posterior evaluation of the site by an International Atomic Energy Agency (IAEA) convoy.

Even though the four peaks detected in Topic 49 (**l**, **n**, **t**, **u**) were small when compared to the entire series, this topic reports the lengthy Ukrainian counteroffensive on the Kherson oblast, that culminated in the retreat of Russian Armed Forces from the region (Santora et al., 2022). Topic 72 (**p**) also displayed a spread-out peak regarding major Ukrainian advances in this region as well as in Kharkiv. Topics 181 and 68 (**d**, **k**) describe Russian attacks with significant civilian casualties on a maternity hospital in Mariupol (Guy et al., 2022) and a shopping mall in Kremenchuk (Faulconbridge et al., 2022). Both topics show significant peaking posterior to the attacks, mostly comprised of stories reporting the devastation and civilian casualties endured, as well as the reaction of political leaders to the attacks.

4.2.2 Politics

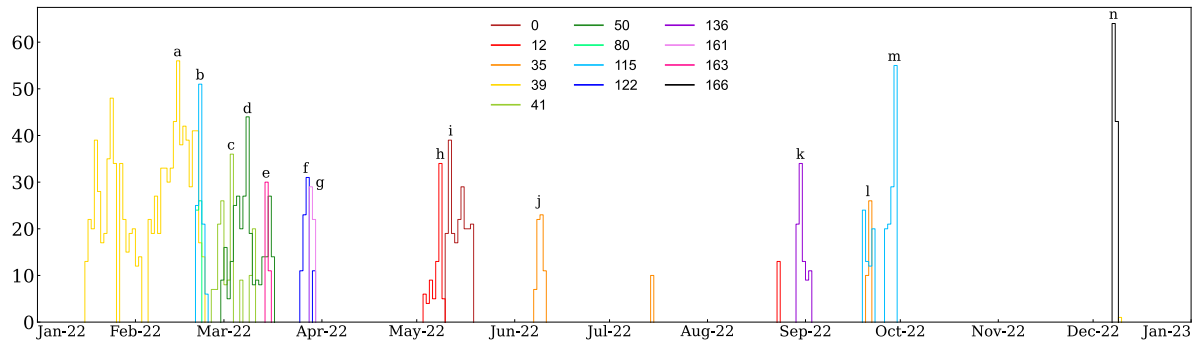

Figure 6 – Politics - Peaks

Peak	Date	Topic	Count	Event
a	Feb 14	39	43	Speculation before the first invasion, nations react to threats
b	Feb 22	115	51	Putin orders troops to Ukraine after recognizing breakaway regions
c	Mar 4	41	36	Russia blocks access to social media and media outlets
d	Mar 9	50	44	NATO declines no-fly zone over Ukraine, Poland offers jets for Ukraine
e	Mar 15	163	30	Russian court fines woman for anti-war protest on state TV
f	Mar 28	122	31	Biden speech: "Putin cannot remain in power"
g	Mar 29	161	29	Russia and Ukraine engage in peace talks
h	May 9	12	34	Putin Victory Day speech
i	May 12	0	39	Finland and Sweden join NATO
j	Jun 10	35	23	Two British soldiers sentenced to death by Russia
k	Aug 31	136	34	Death of Mikhail Gorbachev
l	Sep 22	35	25	Major Russia-Ukraine prisoner swap
m	Sep 30	115	55	Putin signs decree officially annexing four Ukrainan regions
n	Dec 8	166	64	Brittany Griner freed in a prisoner swap

Table 4 – Politics - Description of the events found

"Politics" is the second largest category, with 5,856 stories classified under its topics, or 9.6% of the corpus. It contains the reaction to multiple national addresses by political leaders and major events such as Finland and Sweden joining the North Atlantic Treaty Organization (NATO) (**i**) (O'Neil, 2022). This particular event had major coverage during the discussion stage in May, with both nations applying to join the organization on May 18, even though the accession protocol was only signed by NATO on July 5. While many peaks are distinct and isolated, such as the widely-covered prisoner deal that freed Women's National Basketball Association player Brittany Griner on December 8 (**n**) (Shear; Baker, 2022) and the announcement of the unilateral annexation of Donetsk, Kherson, Luhansk, and Zaporizhzhia oblasts by Russian President Vladimir Putin on September 30 (**m**), this category also represents spread-out events like the lengthy reporting of the aggravation of the political animosity between Ukraine and Russia (**a**) and the social media and news outlets restrictions put into practice in Russia during February and March (**c**) (Troianovski; Safronova, 2022).

4.2.3 Prices, Supply Chain and Sanctions

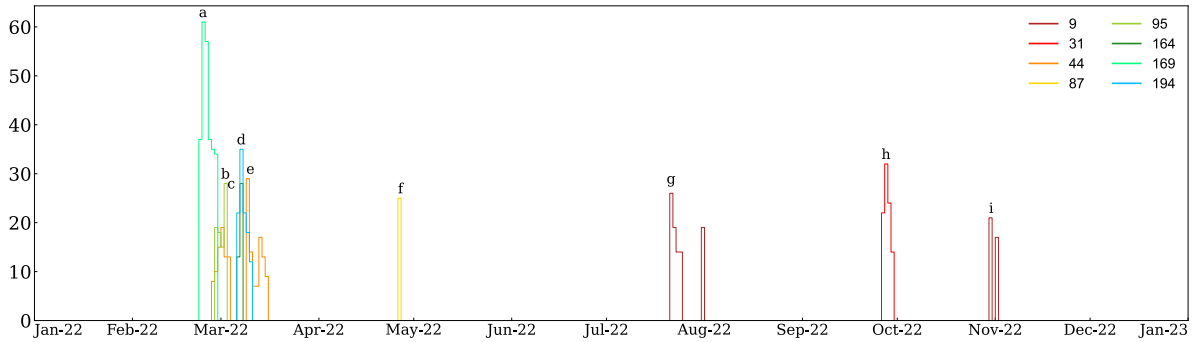

Figure 7 – Prices, Supply Chain and Sanctions - Peaks

Peak	Date	Topic	Count	Event
a	Feb 24	169	61	First European sanctions, SWIFT bans
b	Mar 3	95	28	Wall Street pushes measures to remove Russian assets
c	Mar 8	164	28	Biden bans all imports of Russian oil into the U.S.
d	Mar 8	194	35	Sources report record fuel prices in the U.S.
e	Mar 10	44	29	Roman Abramovich, the owner of Chelsea F.C., hit by sanctions
f	Apr 27	87	25	Russia cuts off natural gas to Poland and Bulgaria
g	Jul 22	9	26	Ukraine and Russia sign UN-backed deal to restart grain exports
h	Sep 28	31	31	Nord Stream gas pipeline leaks
i	Oct 31	9	21	Russia suspends grain deal with Ukraine

Table 5 – Prices, Supply Chain and Sanctions - Description of the events found

This category represents 7.7% of the corpus, with 4,718 stories classified under its topics. Peak **a**, contemporaneous to peak **a** in "Battles", reported the immediate reaction of nations to the invasion, as well as the discussion of SWIFT bans on major Russian banks (Blenkinsop, 2022). These bans were applied on March 1, and this dynamic is represented by a high peak on the most inflammatory moment of the invasion, which tapers off towards the dates surrounding the execution of the bans. The only recurring topic (Topic 9 - **g**, **i**) references the Initiative on the Safe Transportation of Grain and Foodstuffs from Ukrainian ports, an UN-backed deal conducted to address the global food crisis directly caused by the conflict (Boffey et al., 2022). The deal was signed on July 22, and Russia suspended its participation on October 29, returning to the deal on November 2. On the first of August, the first ship loaded with Ukrainian grain left the port of Odesza. These four dates were well represented in our findings. Topics 164 and 194 peaked on the same date (**c**, **d**), and both report effects of the oil crisis on the American economy, with record-high fuel prices and the signing of an Executive Order by President Joe Biden banning the imports of Russian oil, liquefied natural gas, and coal into the U.S. (Isidore, 2022).

4.2.4 Artillery and Troops

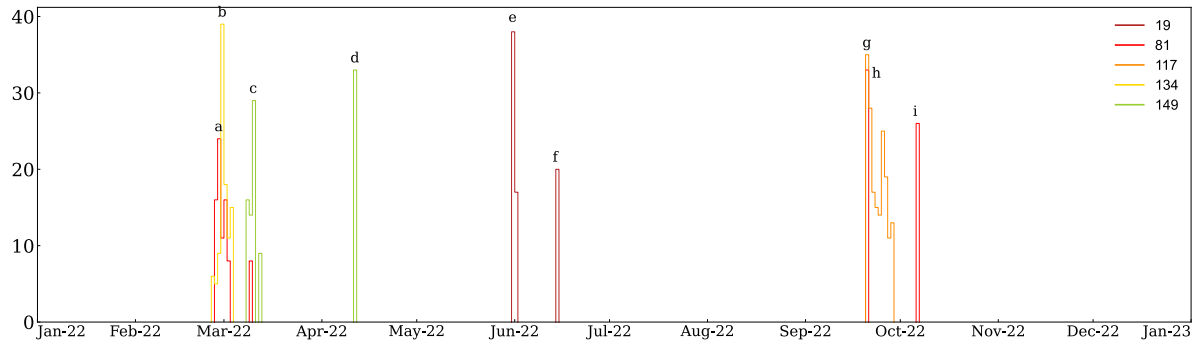

Figure 8 – Artillery and Troops - Peaks

Peak	Date	Topic	Count	Event
a	Feb 28	81	24	Putin declares nuclear alert
b	Mar 1	134	39	Russia uses Cluster and Thermobaric bombs in Kharkiv
c	Mar 11	149	29	Speculation on Russian biological weapons and U.S.-funded Labs in Ukraine
d	Apr 12	149	33	Speculation on Russian chemical attack in Mariupol
e	Jul 1	19	38	U.S. sends long-range rocket system to Ukraine
f	Jul 15	19	20	Biden Administration announces additional $1B in military aid for Ukraine
g	Sep 21	117	31	Putin announces partial mobilisation of reservists
h	Sep 21	81	33	Putin issues nuclear threat to the West
i	Oct 7	81	26	Biden mentions Cuban Missile Crisis in speech, cites risk of "Armageddon"

Table 6 – Artillery and Troops - Description of the events found

This category represents 4.7% of the corpus, with 2,847 stories classified under its topics. It mainly covers political speeches regarding the use of specific artillery such as nuclear and biochemical weapons, and the peaks found were mostly well-defined and concentrated. Topics 149 (**c**, **d**), 81 (**a**, **h**, **i**) and 19 (**e**, **f**) were recurring, with weapons of the same nature (biochemical and nuclear) being reported on the former two (Qiu, 2022; Sanger; Broad, 2022), and major events of U.S. aid to Ukraine on the latter (Erlanger, 2022). Topic 117 (**g**) highlights the partial mobilization of military reservists announced by President Vladimir Putin on September 21, with Defense Minister Sergei Shoigu announcing a plan to mobilize 300,000 recruits (Sauer, 2022).

4.3 Novelty, Resonance and Transience

In this subsection, we apply the KLD-based measures of novelty, resonance and transience to the corpus and investigate trendsetting characteristics pertaining to the entire system and to each source.

Figure 9 shows a density plot of the transiences and novelties of stories in the entire corpus at a time scale of $w = 10$ stories. This scale is used to obtain the average KLD between a number of w past and future topic distributions and each document as described in Chapter 3.2. The identity line ($x = y$) shows stories that display the same surprise when compared to future and past documents. We consider stories below the identity line resonant ($\mathcal{N} > \mathcal{T}$), and therefore better aligned with future stories rather than previous ones. Stories above the identity line received little reward in resonance for the amount of novelty achieved, that is, did not follow the past trend but failed to establish a new reporting pattern. In Figure 10, we present the results of resonance and novelty for time scales w equal to 10, 100 and 1000 stories. Given that the measure of transience is dependent on the reception of the system to a given story, this characteristic is out of each source's control, as opposed to novelty. We fit a linear model to our data to measure the expected resonance of any story given some level of novelty:

$$E[\mathcal{R}|\mathcal{N}] = \beta_{int} + \beta_{\mathcal{N}}\mathcal{N} + \sigma \tag{4.1}$$

As shown in Figure 10, the system shows a general bias towards innovation ($\beta_{\mathcal{N}} > 0$) from short to very long time scales. This is expected since new topics are generally covered across multiple sources after their introduction. However, the increase in w results in a noticeable decrease in the slope of the fit line, indicating that the tendency for novelty bias is less present for long timeframes. The comparison between the average expected resonance of each source and the average system-wide expected resonance allows us to measure how each source was capable of overcoming system-wide trends:

$$\Delta\overline{E[\mathcal{R}|\mathcal{N}]}_{source} = \overline{E[\mathcal{R}|\mathcal{N}]}_{source} - \overline{E[\mathcal{R}|\mathcal{N}]}_{system} \tag{4.2}$$

Figure 11 compares the system averages of $\mathcal{N}$, $\mathcal{R}$ and $\Delta E[\mathcal{R}|\mathcal{N}]$ for a range of scales w by their Pearson Correlation to $w = 10$. The results show that novelty (and transience, accordingly) is the most consistent result across timescales, with a high correlation (Pearson > 0.8) from $w = 4$ to $w = 50$. This is not the case for $\mathcal{R}$ and $\Delta E[\mathcal{R}|\mathcal{N}]$, which fall in correlation much sooner. $\Delta E[\mathcal{R}|\mathcal{N}]$ shows slightly higher correlation than $\mathcal{R}$ for all timeframes below $w = 300$.

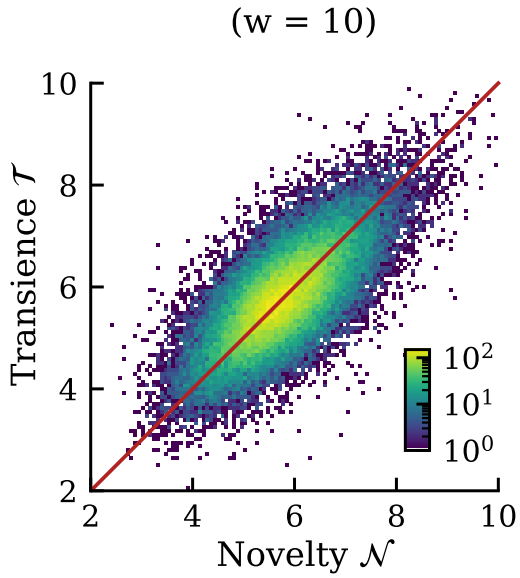

Figure 9 – Density of novelty and transience across the entire system.

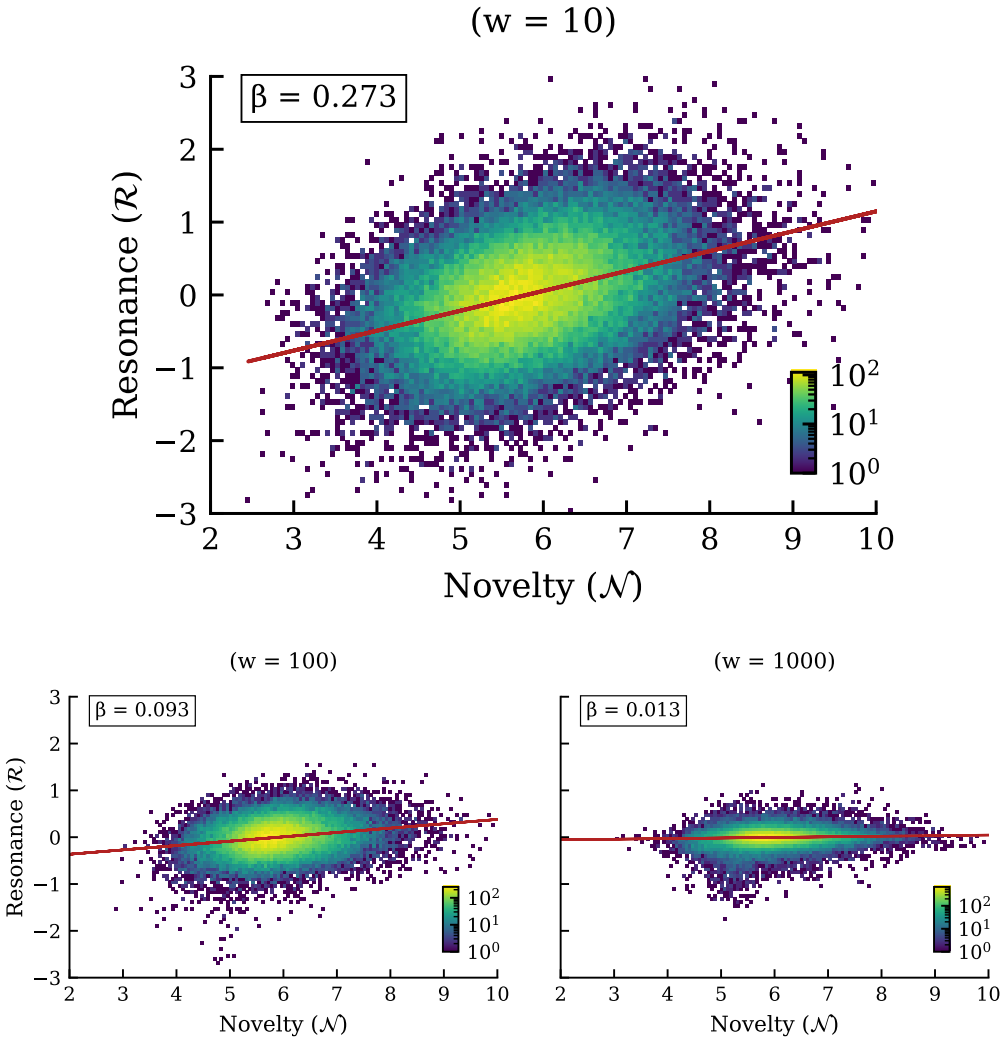

Figure 10 – Novelty and resonance for different timescales (entire system).

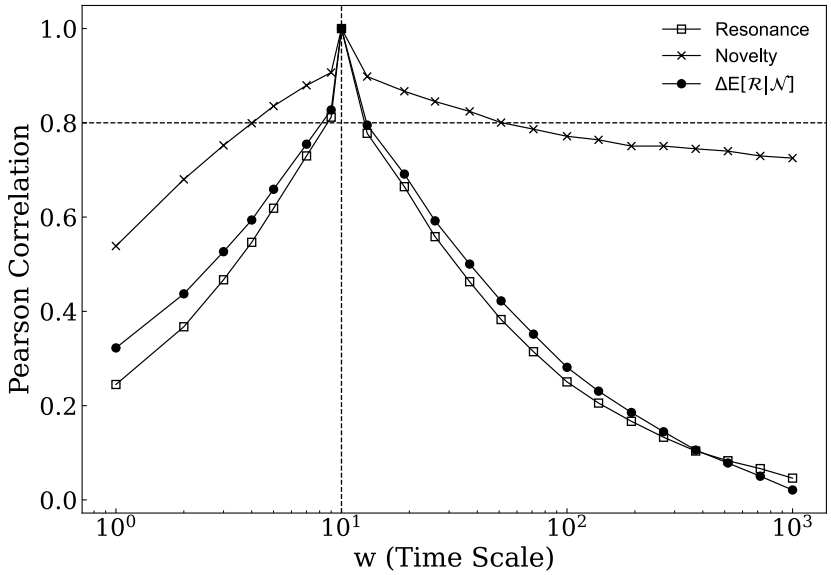

Figure 11 – Pearson Correlation between different timescales w and $w = 10$ for average $\mathcal{R}$, $\mathcal{N}$ and $\Delta E[\mathcal{R}|\mathcal{N}]$.

We isolate the results from each source from the comprehensive dataset in order to explicit different trendsetting characteristics and then compare the filtered results with the original dataset to identify variations between the sources and the system. Comparisons of the measures described previously between each source and the entire system are shown in Table 7, and the results shown for $\Delta\mathcal{R}$ and $\Delta\mathcal{N}$ are plotted in Figure 12, where the highest trendsetters are placed in the first two quadrants, with a higher resonance than the system average. We conduct t-tests between the $\mathcal{N}$, $\mathcal{T}$ and $\Delta E[\mathcal{R}|\mathcal{N}]$ series of each source and of the entire system.

The comparison of document resonances from each source to the entire system showed low significance, meaning that there is not enough evidence to conclude that the resonances of documents from one source are significantly different from the entire series. This indicates that each source does not have a distinct impact on prior tendencies and creating new tendencies, compared to the entire group of documents. However, 9 of the 11 sources were found to be significant at a 5% level, and 8 at a 1% level, for both novelty and deviation from expected resonance. This suggests that the deviation from the expected resonance and novelty of documents from these sources is distinct from what would be expected based on the overall pattern, and may represent unique characteristics or behaviors of these sources. The values in Table 7 show the difference between the means of each source and the entire system.

Source	$\Delta\overline{\mathcal{R}}$	$\Delta\overline{\mathcal{N}}$	$\Delta\overline{E[\mathcal{R}\|\mathcal{N}]}$
AP	-0.0037	0.6715***	-0.1787***
ABC	0.0012	0.3408***	-0.0876***
CBS	0.0036	0.0206	-0.0018
CNN	0.0095	0.1845***	-0.0386***
DailyMail	0.0032	0.3072***	-0.0769***
Express	0.0093	-0.1994***	0.0613***
Fox	0.0017	-0.1596***	0.0433***
Guardian	-0.0140	0.0546***	-0.0282***
Mirror	0.0053	-0.0003	0.0054
NYT	-0.0154	-0.0388**	-0.0053**
Reuters	-0.0028	-0.2977***	0.0747***

Table 7 – $\Delta\overline{\mathcal{R}}$, $\Delta\overline{\mathcal{N}}$ (Difference between average observed resonance/novelty of each source and of the system) and $\Delta\overline{E[\mathcal{R}|\mathcal{N}]}$ (Difference between the average expected resonance of each source and of the system) for all sources. The significance shown represents t-tests between each source and the entire corpus at the following levels: * - 10%, ** - 5%, *** - 1%.

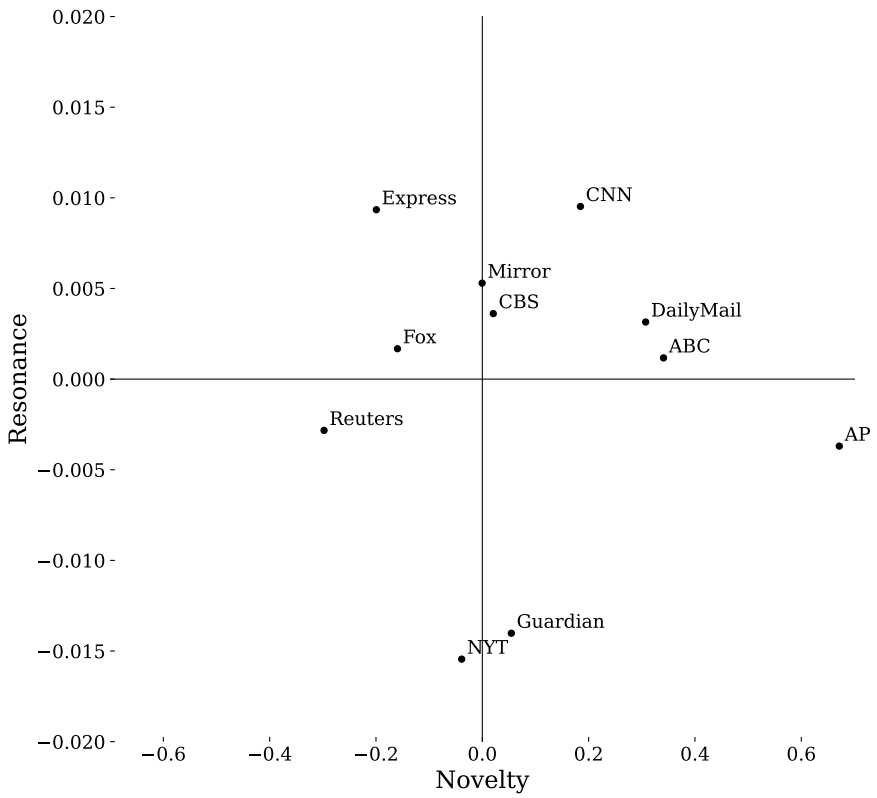

Figure 12 – Novelty and Resonance deviations from system average for each source.

5 Final Considerations

In conclusion, we investigated if we can detect key events of a global event by analyzing unstructured textual data from news outlets and whether different outlets show significantly unique trendsetting characteristics in reporting when compared to the entire system. Through the application of latent Dirichlet allocation (LDA) topic modeling, we have identified and described several key events in the Russo-Ukrainian war reported in 11 different news sources after the escalation of the conflict in February 2022. Additionally, we used Kullback-Leibler Divergence (KLD) to analyze the fluctuations in topic coverage for each news source, which highlighted differences in the trendsetting potential of each outlet.

Our results demonstrate that LDA is capable of detecting key conflict events covered across multiple news outlets and that relative entropy can be applied to topic distributions to highlight differences in trendsetting potential for this particular application, which is aligned with the findings of works in other domains followed by the first use of this methodology in Barron et al. (2018). The likeliest words of each topic found give insight into the events being reported, confirming the effectiveness of LDA in capturing specific events. Furthermore, our results show that the largest category of peaking topics detected within our corpus, with the most news stories represented, was related to battles, which supports the idea that despite the presence of other major repercussions of a conflict, major media outlets tend to favor the reporting of direct conflict advancements.

However, it's important to note that confirming the peaks we found using our methodology requires another source of key event data. This is a major limitation of this study, as the peaks found are solely based on the news sources analyzed. Despite this limitation, the results of this study shed light on the role of unstructured text data in detecting key points in conflict situations, and further research could focus on validating the findings with other sources of information. One of our key findings is that identifying key points of any major event widely covered by media is possible using topic modeling alone, and this fact could be explored in future research regarding different events.

References

Archer, A. M.; Clinton, J. Changing owners, changing content: Does who owns the news matter for the news? *Political Communication*, Taylor & Francis, v. 35, n. 3, p. 353–370, 2018. Cited on page 12.

Barnes, J. Iran sends drone trainers to crimea to aid russian military. *The New York Times*, 2022. Available on: <https://www.nytimes.com/2022/10/18/us/politics/iran-drones-russia-ukraine.html>. Cited on page 22.

Barron, A. T. et al. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 115, n. 18, p. 4607–4612, 2018. Cited 4 times on pages 10, 13, 16, and 31.

Blei, D. M. Probabilistic topic models. *Communications of the ACM*, ACM New York, NY, USA, v. 55, n. 4, p. 77–84, 2012. Cited 2 times on pages 12 and 15.

Blei, D. M.; Ng, A. Y.; Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Cited 3 times on pages 10, 12, and 15.

Blenkinsop, P. Eu bars 7 russian banks from swift, but spares those in energy. *Reuters*, 2022. Available on: <https://www.reuters.com/business/finance/eu-excludes-seven-russian-banks-swift-official-journal-2022-03-02/>. Cited on page 25.

Boffey, D. et al. Ukraine and russia sign un-backed deal to restart grain exports. *The Guardian*, 2022. Available on: <https://www.theguardian.com/world/2022/jul/22/ukraine-russia-sign-un-backed-deal-restart-grain-exports>. Cited on page 25.

Bonilla, T.; Grimmer, J. Elevated threat levels and decreased expectations: How democracy handles terrorist threats. *Poetics*, Elsevier, v. 41, n. 6, p. 650–669, 2013. Cited on page 12.

Borger, J.; Henley, J. Zelenskiy says 'europe must wake up' after assault sparks nuclear plant fire. *The Guardian*, 2022. Available on: <https://www.theguardian.com/world/2022/mar/04/ukraine-nuclear-power-plant-fire-zaporizhzhia-russian-shelling>. Cited on page 23.

Brakel, J. v. *Robust peak detection algorithm using z-scores*. 2014. Available on: <https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data/22640362#22640362>. Cited 2 times on pages 17 and 20.

Cajueiro, D. O. et al. A model of indirect contagion based on a news similarity network. *Journal of Complex Networks*, Oxford University Press, v. 9, n. 5, p. cnab035, 2021. Cited on page 12.

Correia, M. P. R.; Mueller, B. Pattern making and pattern breaking: measuring novelty in brazilian economic research. *Revista Brasileira de Inovação*, SciELO Brasil, v. 21, 2022. Cited on page 13.

Degaetano-Ortlieb, S.; Teich, E. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In: *Proceedings of the second joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*. 2018. p. 22–33. Cited on page 13.

DiMaggio, P.; Nag, M.; Blei, D. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, Elsevier, v. 41, n. 6, p. 570–606, 2013. Cited on page 12.

Erlanger, S. U.s. and allies pledge additional arms for ukraine, but kyiv wants more. *The New York Times*, 2022. Available on: <https://www.nytimes.com/2022/06/15/world/europe/biden-ukraine-weapons.html>. Cited on page 26.

Faulconbridge, G.; Graff, P.; Macfie, N. Russia denies hitting ukrainian shopping centre with missiles. *Reuters*, 2022. Available on: <https://www.reuters.com/world/europe/russia-says-hit-weapons-depot-kremenchuk-caused-fire-shopping-center-2022-06-28/>. Cited on page 23.

Garz, M.; Rickardsson, J. Ownership and media slant: Evidence from swedish newspapers. *Kyklos*, Wiley Online Library, 2022. Cited on page 12.

Guy, J. et al. Russia's bombing of maternity and children's hospital an 'atrocity,' zelensky says. *CNN*, 2022. Available on: <https://edition.cnn.com/2022/03/09/europe/russia-invasion-ukraine-evacuations-03-09-intl/index.html>. Cited on page 23.

Isidore, C. Why us gas prices are at a record and why they'll stay high for a long time. *CNN*, 2022. Available on: <https://edition.cnn.com/2022/03/09/energy/record-gas-price-causes/index.html>. Cited on page 25.

Jakes, L. Here's what we know about the s-300 missile, which was involved in the poland blast. *The New York Times*, 2022. Available on: <https://www.nytimes.com/2022/11/16/world/europe/poland-missile-s-300.html>. Cited on page 22.

Jing, E.; DeDeo, S.; Ahn, Y.-Y. Sameness attracts, novelty disturbs, but outliers flourish in fanfiction online. *arXiv preprint arXiv:1904.07741*, 2019. Cited on page 14.

Jockers, M. L.; Mimno, D. Significant themes in 19th-century literature. *Poetics*, Elsevier, v. 41, n. 6, p. 750–769, 2013. Cited on page 12.

Kullback, S.; Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, JSTOR, v. 22, n. 1, p. 79–86, 1951. Cited 2 times on pages 13 and 16.

Li, Y.; Lund, R. Multiple changepoint detection using metadata. *Journal of Climate*, v. 28, n. 10, p. 4199–4216, 2015. Cited on page 12.

Marshall, E. A. Defining population problems: Using topic models for cross-national comparison of disciplinary development. *Poetics*, Elsevier, v. 41, n. 6, p. 701–724, 2013. Cited on page 12.

McFarland, D. A. et al. Differentiating language usage through topic models. *Poetics*, Elsevier, v. 41, n. 6, p. 607–625, 2013. Cited on page 12.

Miller, I. M. Rebellion, crime and violence in qing china, 1722–1911: A topic modeling approach. *Poetics*, Elsevier, v. 41, n. 6, p. 626–649, 2013. Cited on page 12.

Mohr, J. W.; Bogdanov, P. Introduction—topic models: What they are and why they matter. *Poetics*, Elsevier, v. 41, n. 6, p. 545–569, 2013. Cited on page 12.

Mohr, J. W. et al. Graphing the grammar of motives in national security strategies: Cultural interpretation, automated text analysis and the drama of global politics. *Poetics*, Elsevier, v. 41, n. 6, p. 670–700, 2013. Cited on page 12.

Murdock, J.; Allen, C.; DeDeo, S. Exploration and exploitation of victorian science in darwin's reading notebooks. *Cognition*, Elsevier, v. 159, p. 117–126, 2017. Cited on page 14.

Nielbo, K. L. et al. Trend reservoir detection: Minimal persistence and resonant behavior of trends in social media. *Proceedings http://ceur-ws. org ISSN*, v. 1613, p. 0073, 2020. Cited on page 13.

O'Neil, T. Finland, sweden file official applications to join nato amid russia-ukraine war. *Fox News*, 2022. Available on: <https://www.foxnews.com/world/finland-sweden-official-applications-nato-russia-ukraine>. Cited on page 24.

Park, J. H.; Yamauchi, S. Change-point detection and regularization in time series cross-sectional data analysis. *Political Analysis*, Cambridge University Press, p. 1–21, 2022. Cited on page 12.

Pierson, P. *Politics in Time: History, Institutions, and Social Analysis*. 2004. Cited on page 10.

Qiu, L. Theory about u.s.-funded bioweapons labs in ukraine is unfounded. *The New York Times*, 2022. Available on: <https://www.nytimes.com/2022/03/11/us/politics/us-bioweapons-ukraine-misinformation.html>. Cited on page 26.

Rankin, J.; Boffey, D. Killing of civilians in bucha and kyiv condemned as 'terrible war crime'. *The Guardian*, 2022. Available on: <https://www.theguardian.com/world/2022/apr/03/eu-leaders-condemn-killing-of-unarmed-civilians-in-bucha-and-kyiv>. Cited on page 23.

Ruggieri, E.; Antonellis, M. An exact approach to bayesian sequential change point detection. *Computational Statistics & Data Analysis*, Elsevier, v. 97, p. 71–86, 2016. Cited on page 12.

Sanger, D.; Broad, W. Putin declares a nuclear alert, and biden seeks de-escalation. *The New York Times*, 2022. Available on: <https://www.nytimes.com/2022/02/27/us/politics/putin-nuclear-alert-biden-deescalation.html>. Cited on page 26.

Santora, M. et al. Russia orders retreat from kherson, a serious reversal in the ukraine war. *The New York Times*, 2022. Available on: <https://www.nytimes.com/2022/11/09/world/europe/ukraine-russia-kherson-retreat.html>. Cited on page 23.

Sauer, P. Putin announces partial mobilisation and threatens nuclear retaliation in escalation of ukraine war. *The Guardian*, 2022. Available on: <https://www.theguardian.com/world/2022/sep/21/putin-announces-partial-mobilisation-in-russia-in-escalation-of-ukraine-war>. Cited on page 26.

Schwirtz, M. et al. Here's how the russian attack is unfolding. *The New York Times*, 2022. Available on: <https://www.nytimes.com/2022/02/24/world/europe/how-russia-attacked-ukraine.html>. Cited on page 22.

Shear, M.; Baker, P. Brittney griner is freed as part of a prisoner swap with russia. *The New York Times*, 2022. Available on: <https://www.nytimes.com/2022/12/08/us/politics/brittney-griner-russia.html>. Cited on page 24.

Spirling, A. Bayesian approaches for limited dependent variable change point problems. *Political Analysis*, Cambridge University Press, v. 15, n. 4, p. 387–405, 2007. Cited on page 12.

Spirling, A. "turning points" in the iraq conflict: Reversible jump markov chain monte carlo in political science. *The American Statistician*, Taylor & Francis, v. 61, n. 4, p. 315–320, 2007. Cited on page 12.

Tangherlini, T. R.; Leonard, P. Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics*, Elsevier, v. 41, n. 6, p. 725–749, 2013. Cited on page 12.

Troianovski, A.; Safronova, V. Russia takes censorship to new extremes, stifling war coverage. *The New York Times*, 2022. Available on: <https://www.nytimes.com/2022/03/04/world/europe/russia-censorship-media-crackdown.html>. Cited on page 24.

Tsur, O.; Calacci, D.; Lazer, D. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015. p. 1629–1638. Cited on page 12.

Western, B.; Kleykamp, M. A bayesian change point model for historical time series analysis. *Political Analysis*, Cambridge University Press, v. 12, n. 4, p. 354–374, 2004. Cited on page 12.

Zinets, N. Mariupol defenders surrender to russia but their fate is uncertain. *Reuters*, 2022. Available on: <https://www.reuters.com/world/europe/ukrainian-troops-evacuate-mariupol-ceding-control-russia-2022-05-17/>. Cited on page 23.