Thematic analysis of economic inequality coverage over time in The New York Times and US Congress (1980-2024)

Master Thesis

Submitted by:

Agustina Pesce

Student ID: 1247019

at

Universität Konstanz

Period of completion: October 2024 - March 2025

Thesis submitted in partial fulfillment of the requirements for the degree of the Master of Social and Economic Data Science at the University of Konstanz

1st Supervisor: Prof. Dr. David Garcia

2nd Supervisor: Prof. Dr. Claudia Diehl

Konstanz, 25th March 2025

Table of contents

Abstract

This thesis examines how economic inequality has been addressed between 1980 and 2024 by two influential actors in US public debate: the legacy media outlet *The New York Times* and the US Congress. Specifically, it analyzes the issue's salience and the themes most frequently associated with it. We retrieved relevant documents through full-text keyword queries targeting economic and inequality-related terms, resulting in a corpus of 11,403 articles and 3,777 speeches. We implemented a probabilistic topic model (Structural Topic Model, STM) to identify themes within this corpus. Based on the STM results, we developed a codebook and used it to perform zero-shot deductive annotations using Llama 3, a large language model capable of context-aware classification.

The findings support previous research that identified the Occupy Wall Street movement as a key moment in raising attention towards economic inequality. They further suggest a recent decline in the issue's centrality, an area not yet widely explored. The most prominent associated themes across the studied period were horizontal inequalities, welfare policies, and macroeconomic policies. In recent years, the incidence of these themes has shifted, with greater relative prevalence of horizontal inequalities —particularly for racial disparities— and welfare policies, alongside a decline in macroeconomic discussions.

1. Introduction

1.1 Economic inequality: an issue in the rise

Since 1980, economic inequality in the United States has steadily risen. This trend is evident in the upward trajectory of the Gini index for income distribution (Figure 1.1) and the growing wealth share concentrated among the top 1% percentile (Board of Governors of the Federal Reserve System, 2024; Saez & Zucman, 2016; World Bank, 2024). Greater economic inequality has been linked to detrimental social implications, including declining trust and cooperation, increased violence, reduced political participation, and weakened support for democratic institutions (Jetten et al., 2021). Additionally, periods of economic crisis, such as the 2007-2009 Great Recession or the 2020-2021 COVID-19 pandemic, have recently heightened financial instability.

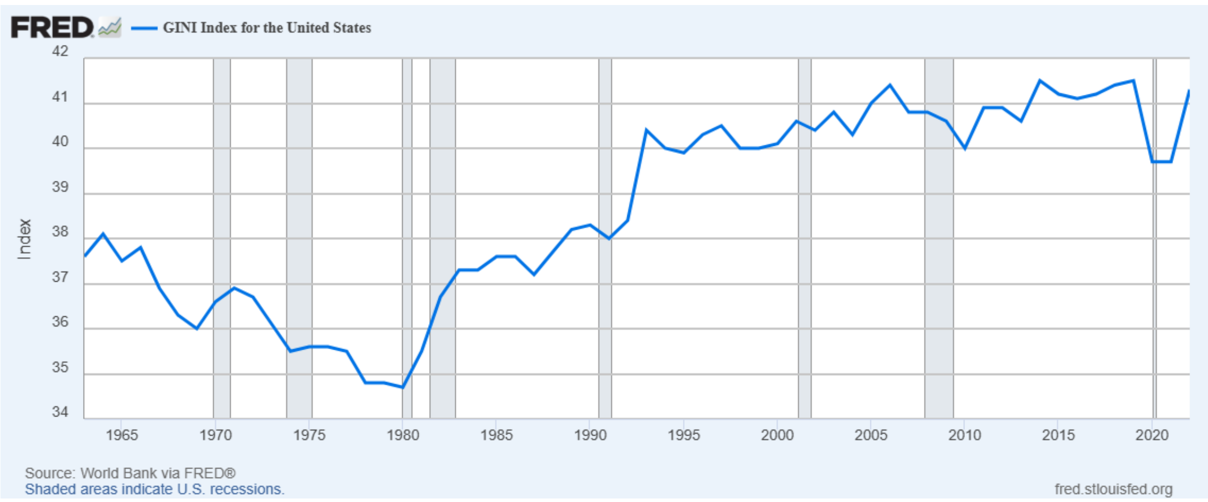

Figure 1.1: Gini index for the United States (1963-2022) (World Bank, 2024)

Despite the deepening of economic inequality and its societal repercussions, US public opinion data suggests that there has not been a corresponding growth in the perception of rising economic inequality (Franko, 2017; McCall, 2013). This discrepancy prompts questions about how the public conversation about economic inequality has developed over time, particularly the level of attention it has received across different spheres of debate and the topics with which it has been most commonly associated.

1.2 Media and Congress: sources of influence for the public debate

Since economic inequality within the total population distribution is not directly observable in everyday life, it is essential to study how it is debated in public discourse. Actors like media outlets and politicians play a central role in shaping the public conversation on current issues. Additionally, the emphasis these sources place provides insights into how they aim to influence public perceptions and reflect their interpretations of society's primary concerns (Grisold & Preston, 2020; McCall, 2013).

In mass media, communicators decide which aspects of a topic to highlight and which to disregard (Stecula & Merkley, 2019). This becomes particularly relevant when research on media coverage and economic inequality has reported that sustained media coverage over several days affects public perceptions of social justice (Diermeier et al., 2017). The mechanisms underlying this influence include the activation of cognitive schemas in audiences and the indirect effects of media narratives through interpersonal discussions (DiMaggio et al., 2013).

Recent shifts in media landscapes have further complicated the dynamics of citizens' information diets. Traditional legacy media relevance declined, while alternative news outlets, particularly those on social media, have gained prominence since the 2000s (McGovern et al., 2020; Vaughan, 2024). However, this thesis focuses on print media for two reasons. First, traditional newspapers provide a continuous source that enables a longitudinal analysis through the above-mentioned period of rising economic inequality. Second, despite the loss of centrality, legacy print media continues influencing public opinion (McGovern et al., 2020). While we acknowledge the growing influence of social media, in this initial approach to the issue, we prioritized exploring a media source with consistent coverage since 1980.

Democratic political institutions are another major entity influencing the public debate. Previous research has highlighted that once an issue gains traction within Congress, it may later be picked up by mass media (Edwards & Wood, 1999). Furthermore, parliamentary debates are central to representative democracy, as they provide a platform for legislators to express their ideas publicly and communicate with voters (Osnabrügge et al., 2021). Beyond facilitating communication and position-taking for representatives, these debates also influence policymaking by persuading other parliamentary members, thereby shaping legislative outcomes (Proksch & Slapin, 2012).

Legacy media and the Congress play a critical role in shaping public debate. This thesis examines how these two institutions have addressed economic inequality over time, analyzing the level of attention given to the issue and the themes most commonly associated with it.

1.3 Themes present in the economic inequality debate

The prevailing themes in the public conversation provide valuable insights to assessing the dimensions considered most central within the broad issue of economic inequality. For instance, discussions may focus on the relevance of fiscal balance or the effects of economic inequality on health. Analysing this development within the public conversation hints at how different actors seek to influence public perceptions and what their interpretations of society's primary concerns are (Grisold & Preston, 2020; McCall, 2013).

Among the various themes associated with economic inequality, we were particularly interested in the evolution of debates concerning horizontal inequalities, which refer to disparities "among culturally determined groups, groups that have salience for their members and/or others in society; for example, among races, ethnic groups, religions, religious sects, regions, and so on"

(Stewart, 2009, p. 316). This concept contrasts with the study of vertical inequality, which examines disparities among individuals across the entire population. Horizontal inequalities extend beyond economic gaps, including broader domains such as inequalities in educational or employment opportunities. Such persistent group-based disparities are relevant in restricting access to economic mobility, resulting in structural disadvantages that raise concerns about fairness, poverty reduction, and the fulfilment of human rights.

1.4 Research questions and thesis structure

This thesis examines how economic inequality has been made salient since the beginning of its pronounced increase after 1980. Specifically, it explores the extent to which economic inequality has been addressed and the themes most commonly associated with it. To do so, we focussed on two influential sources for public debate and democracy, analysing articles from a major legacy newspaper, *The New York Times* (NYT), and speeches from the US Congress.

The deriving research questions for our analyzed time frame, 1980-2024, are three:

RQ1 - How prevalent has the subject of economic inequality been in the NYT and US Congress?

RQ2 - What other topics have been predominantly discussed alongside economic inequality in the NYT and US Congress?

RQ3 - How prevalent has the coverage of horizontal inequalities been when economic inequality is addressed in the NYT and US Congress?

The thesis is structured as follows. The *Research Background* section reviews prior literature on the coverage of economic inequality and methods for thematic analysis. The *Data* section describes the collection, cleaning, and preprocessing of the NYT and US Congress corpora. The *Methods* section outlines the computational techniques we used, including topic discovery with a Structural Topic Model (STM) and the corpus annotation using a Llama 3 large language model. This section also details the validation procedures and the statistical tests we applied. The *Results* section presents the findings on economic inequality coverage and the prevalence of the associated themes over time. The *Discussion* further interprets our findings and contextualizes them within existing previous research. Finally, the *Conclusion* summarizes the main findings and reflects on the methodological limitations. [1]

2. Research background

2.1 Economic inequality in the media and political discourse

2.1.1 Coverage

Despite consistent increases in income and wealth concentration in the US since 1980 (Board of Governors of the Federal Reserve System, 2024; Saez & Zucman, 2016; World Bank, 2024) and periods of economic hardship in recent decades, media and political coverage of economic inequality does not appear to have followed a steady upward trajectory (Baumann & Majeed, 2020; Vaughan, 2024).

A study from McCall (2013), where she explored the content of three major American newsweeklies through articles' subject terms (1980– 2010), found that media coverage of economic inequality occurred in waves. However, there was no clear evidence of a sustained increase in media attention to this subject over time, as coverage levels at the end of the period did not exceed those observed at the beginning. Additionally, the author emphasized that the coverage of economic inequality in the 1990s was similar to that of the 2000s, when recessions and financial crises took place. It should be noted that using subject terms for article extraction instead of full-text queries can hinder the recall of all the subject occurrences. Furthermore, broader

[1] A GitHub repository with the source code and supplementary materials for this thesis is available at: https://github.com/agustinapesce/economic-inequality-coverage.

queries implemented included unrelated concepts such as "Poor/Statistics" or "Skilled labor", possibly introducing noise into the data collection by lowering precision.

An analysis of *The New York Times* and *The Wall Street Journal* (1990–2015) using article content retrieved through a query with 19 keywords (e.g., "pay inequality", "wage inequality"), produced different findings. This study identified two moderate coverage peaks —one in the late 1990s and another in the late 2000s— followed by a substantial and sustained upward trend from 2011 onward, right after the Occupy Wall Street (OWS) movement[2] (McGovern et al., 2020). Similarly, a study of US and Canadian newspapers searching for the keyword phrase "economic inequality" (2000–2014), found a small surge for 2007 and a steady trend from 2011. Researchers contrasted this trend behavior with the one of "poverty" coverage, which has been on the rise with a continuous rhythm over the whole period (Baumann & Majeed, 2020). For similar results see also Eshbaugh-Soha & McGauvran (2018) and Gaby & Caren (2016).

In the field of political speeches, a previous study analyzed the prominence of income inequality attention in US presidential speeches (1999-2013). This was carried out by counting the number of sentences containing phrases from a keyword list (Eshbaugh-Soha & McGauvran, 2018). The findings indicate that this topic received the most attention during the final years of the Clinton administration (1999-2000) but declined significantly during the Bush administration (2001–2008) and under the early years of the Obama presidency (2009–2013). In the case of Obama, attention to income inequality fluctuated over time, yet the OWS protests did not predict an increase in presidential attention to the issue.

Findings in both fields of interest suggest that, while economic inequality has been a recurring topic in traditional media and political speeches, its coverage has not closely relied on objective economic indicators like the surge in economic inequality or recent economic shocks.

[2]The OWS movement emerged in the US in 2011, following the Great Recession. One of its central goals was to highlight economic inequality, popularizing the slogan “We are the 99%” (Baumann & Majeed, 2020).

Building on existing research, three areas for improvement emerge. First, previously implemented queries and data selection methods have often been either too broad —capturing related but noisy terms— or too narrow. Second, existing studies conclude their analysis a decade ago, leaving recent developments unexamined. To our knowledge, the period of the Trump and Biden presidencies, as well as the COVID-19 pandemic and post-pandemic years, has not yet been explored in the context of economic inequality discussions. Finally, there seems to be a notable gap in the analysis of political speeches, a major field of influence in the public discussion. A longitudinal analysis spanning from the beginning of the economic inequality rise to the present would give insights on changes about the evolution of the discussion about this economic tendency.

Drawing from prior studies, we formulated two hypothesis for our coverage research question:

RQ1: How prevalent has the subject of economic inequality been in the NYT and US Congress?

H1.1 - Economic inequality coverage did not experience a sustained increase prior to the onset of the OWS movement in 2011.

H1.2 - Economic inequality coverage is not related to economic inequality objective measures.

2.1.2 Content analysis

Studies applying text analysis techniques to economic inequality discussions have predominantly employed content and framing approaches guided by researchers' interests in specific aspects of how the issue is portrayed. These include studies on whether economic inequality is framed as a social concern or wether this concern is relativized, for instance, with meritocratic arguments (Vaughan, 2024). Other studies have focused on the type of causes and solutions portrayed in media (Baumann & Majeed, 2020) on more general sentiment valorations in the media and political speeches (Eshbaugh-Soha & McGauvran, 2018; McGovern et al., 2020).

In a non-systematic manner, McCall (2013) presents insights from her close reading of 57 articles on the topic published between 1980 and 2010. She highlights the prominence of discussions on taxes during the Reagan presidency (1981–1988), as she explains: "(...) tax cuts topped Reagan's policy agenda, and these often went hand in hand with deficit-reducing slashes in social program expenditures (...)" (McCall, 2013, p. 83).

The 2000s are the first period in which the author stresses the presence of articles referring to the economic hardships and their uneven impact on various groups, such as young people and minorities, and therefore the first appearance of horizontal inequalities in her analysis.

Although this type of qualitative analysis is enriched by the researchers' discussions on the implicit content within the articles, it is a limitation that the inferential process from the documents' text to the conclusions remains obscure, making it difficult to determine the extent to which researchers' perspectives shaped the resulting categorizations. A similar approach is found in McGovern et al. (2020) (N = 240), that developed a codebook with eight categories, such as "conceptualization", "type of change", and "attribution level", to annotate their retrieved corpus.

This can be partly overcome by computational methods that enable themes discovery in a data-driven fashion through word co-occurrences. To our knowledge, the only study applying these methods family to unveil articles content on economic inequality is that of Gaby & Caren (2016). They implemented a non-negative matrix factorization to articles with the word "inequality" (2002-2013) over eight news sources (N = 7,024). The discovered fourteen major themes include contents related to economics, politics, civil rights, and art, among others.

Their analysis of the topics prevalence before and after the OWS movement showed an increase in economic topics (e.g. minimum wage, welfare, income) and a decrease in group based inequalities topics (e.g. gender, LGBT, affirmative action). These results support the hypothesis of a sink in the salience of horizontal inequalities towards concepts related to non grouped-defined economical issues.

In contrast, research on media and social identities reports a rising trend in social identity mentions in media posts on Twitter and Facebook after 2011 (Hopkins et al., 2024). A similar pattern emerges in the frequency of "identity politics" references, a term describing political movements centered on specific identity groups, such as gender or ethnicity. Between 1990 and 2020, the usage of this concept peaked locally in 2008 and increased significantly after 2014 (Amira & Abraham, 2022).

Finally, we did not thematic analysis for political speeches on economic inequality in the US.

Building on previous research, this thesis aims to contribute to the thematic analysis of economic inequality in three areas. First, by implementing automated computational methods, we seek to analyze all the occurrences of economic inequality within the sources of interest over a forty-year period. Second, systematic frameworks reduce researcher degrees of freedom, thereby increasing reproducibility. Third, while existing computational research shares these advantages, it has often been limited to specific periods. In contrast, this study takes a broader explorative apporach, examining the evolution of different themes through an extended timeframe, namely 1980-2024.

Based on the previously mentioned research, we formulated two hypotheses related to the two research questions on themes associated with economic inequality. Since political speeches have been less studied in this context, our hypotheses apply to the NYT and the US Congress.

RQ2 - What other topics have been predominantly discussed alongside economic inequality in the NYT and US Congress?

H2 - After the OWS movement, economic topics such as minimum wage, income, middle class, taxes and welfare increased its prevalence in the discussions about economic inequality.

RQ3 - How prevalent has coverage of horizontal inequalities been when economic inequality is addressed in the NYT and US Congress?

H3 - After the OWS, the prevalence of topics related to horizontal inequalities decreased.

2.2 Thematic analysis and computational methods

Thematic analysis involves identifying "recognizable recurring topics, ideas, or patterns (themes) occurring within the data that provide insight into communication" (Allen, 2017, pp. 1756–1757). The discovered themes serve to provide a comprehensive description of how the studied event is communicated.

Due to the limited research on the themes associated with economic inequality, we adopted an inductive approach, deriving themes from the data under investigation (see next section). Since we were particularly interested in the treatment of horizontal inequalities, this was the only topic category defined ahead of the text analysis.

2.2.1 Probabilistic topic modeling with a Structural Topic Model (STM)

Probabilistic topic models are a group of algorithms built to simultaneously discover and annotate large amounts of documents based on the themes they address (Blei, 2012). These models operate in a “mixed-membership” fashion, meaning that texts are assigned a vector of proportions representing the fraction of words associated with each topic (Roberts et al., 2014). Within the family of machine learning algorithms, topic models are unsupervised methods, as they infer topic structures directly from the data without requiring labeled training sets.

Since its simplest approach, the Latent Dirichlet allocation (LDA), these methods are based in four nested concepts: the corpus (or document collection), documents, topics, and terms (or words). Each document is conceptualized as a distribution over latent topics and each topic as a distribution over terms probabilities (Maier et al., 2018). Grounded in the assumption that semantically similar words tend to appear together, word co-occurrence information is leveraged to uncover the underlying structure of these distributions.

These models are particularly well-suited for an initial exploratory analysis of textual data due to their transparent algorithmic design and low computational cost. Even without any direct understanding of word semantics, using a "bag-of-words" assumption, topic models effectively identify linguistic contexts easy to interpret by humans (DiMaggio et al., 2013). Additionally, they have been recognized as a strong inductive method for uncovering previously overlooked categories (Chen et al., 2023). Regarding its coherence with human thematic coding, previous research on economic inequality has addressed an impressive overlap between topic modeling algorithms and hand-coded articles (Nelson et al., 2021).

Given that this thesis analyzes two corpora from distinct discursive contexts, it is particularly relevant to examine how the model handles widely used words that are specific to each dataset. Not all the words used in the documents are thematically relevant, some of them instead reflect the stylistic and procedural conventions of their respective discursive contexts. For instance, terms such as "letter", "summary", or "reporting" are common in journalistic texts, while words like "speaker", "thank", or "tonight" frequently appear in congressional speeches. Topic models tend to group these terms into what are known as “boilerplate topics”. Identifying such topics enhances the interpretability of the remaining thematic clusters and improves comparability across different datasets (Maier et al., 2018).

When researchers believe that covariates within the dataset affect the mentioned document-topic and topic-word distributions, a Structural Topic Model (STM) can be implemented to incorporate them in the model estimation. This approach allows document-level variables to influence the prevalence of specific topics and the most representative words within each topic (Roberts et al., 2014). In the case of this research, this feature is particularly useful given its heterogeneous dataset, which includes two very distinct sources and spans over four decades.

Another relevant characteristic of topic models is that their results are sensitive to initialization values. Since the mixed-membership optimization problem involves a non-convex posterior, parameter estimation can converge to different local optima depending on the initialization values. However, previous research has identified an alternative deterministic option, the Spectral initialization. This procedure relies on applying a method of moments estimation to the decomposed document-term matrix, and it has been shown to yield consistent results compared to stochastic approaches, "while retaining guarantees of globally optimal convergence" (Roberts et al., 2016, p. 31).

One of the most critical parameters in topic modeling is the number of topics (K), which must be specified by the researcher. This parameter directly influences the granularity of the model: a smaller K results in broader, more generalized topic categories, while a larger K may lead to highly specific, overlapping topics that are difficult to distinguish from one another. The standard approach for determining an optimal K involves building models with different values of K and subsequently evaluating their interpretability and coherence (Maier et al., 2018; Roberts et al., 2014).

Different procedures allow for topic model selection and validation. Researchers often rely on face validity to identify the most informative models, assessing the number of interpretable topics. This evaluation involves examining the internal cohesion of a topic's most representative words (semantic validity) and analyzing the documents with the highest topic proportions to determine whether they share a consistent underlying theme (content validity). This procedure allows for selecting the most useful model, one that humans perceive as having a cohesive underlying meaning and that holds theoretical relevance in the studied area (Bernhard et al., 2023).

Once the model is selected, interpretations of the different topic latent variables must be done to evaluate the topics that should be further investigated. Topics that are difficult to interpret through their top-words, topics with high prevalence in documents that refer to different themes and boilerplate topics should be removed. In addition, similar topics can be grouped (Maier et al., 2018).

After model selection and in-depth interpretation, internal and external validation methods must be applied to assess the topic model validity. Internal coherence can be assessed through systematic methods that include human assessment, such as the topic intrusion task. Developed by Chang et al. (2009), this task involves presenting external evaluators with a set of words from the same topic, along with a randomly inserted intruder term. If participants consistently identify the intruder, the topic is considered to align well with human conceptualization. This measure is more effective than purely statistical metrics, such as Normalized Pointwise Mutual Information (NPMI), which prioritize word co-occurrence, an aspect already optimized by the topic model's algorithm. In fact, previous research has found a negative correlation between NPMI scores and human interpretability (Bernhard et al., 2023; Chang et al., 2009).

As external validity measures, historical events can serve as criteria to evaluate the model's accuracy to reflect real-world patterns. In this case, we chose to examine the emergence of a COVID-19 topic in 2020 and the prevalence of the taxes topic during the Reagan presidency as emphasized by McCall (2013).

2.2.2 LLM assisted codebook annotations

The adoption of neural network architectures for text-related research has increased rapidly driven by advances in the text-comprehension capabilities of models. This area has had unprecedented progress thanks to two factors: the introduction of the attention mechanism, which enhances word dependency modeling, and the ability to train models on unsupervised tasks with large-scale corpora (Devlin et al., 2019; Vaswani et al., 2023).

Additionally, large language models (LLMs), such as Meta's Llama 3, have incorporated post-training pipelines that further refine their capabilities. These include supervised fine-tuning (SFT) techniques, such as instruction tuning and Direct Preference Optimization (DPO), which improve response accuracy and enhance human-AI interaction (Meta, 2024). Thus, these models represent a methodologically distinct approach from probabilistic topic models for text analysis. The mentioned strengths of probabilistic topic models in unsupervised topic discovery can therefore be complemented by LLM-based annotations, which incorporate context understanding.

Previous research has evaluated the accuracy of various LLM approaches and established guidelines for improving their performance in the field of deductive coding, which involves identifying themes based on a predefined typology (Allen, 2017). Although the careful fine-tuning of BERT-based models sets a very high baseline (Ziems et al., 2024), it is particularly encouraging that methods that do not require labeled datasets for training have demonstrated human-equivalent accuracy for some tasks (Dunivin, 2024). Furthermore, studies on news article topic labeling with zero-shot prompts, prompts without examples, have reported higher human-model agreement than human-human agreement for most coding categories (Chew et al., 2023). This advantage may be attributed to LLMs' stronger performance in annotation tasks where category meanings closely align with their pre-trained conceptual representations, like in topic annotations. Contrarily, their effectiveness is limited when dealing with categories requiring the interpretation of complex latent variables or concepts that require systematic domain-specific explanations (Halterman & Keith, 2025).

Guidelines for codebook development emphasize the use of codebook-centered prompts rather than example-centered ones (Xiao et al., 2023). Additionally, incorporating Chain-of-Thought (CoT) reasoning instructions, which guide the model through a structured sequence of steps for the given task, enhanced annotation accuracy (Dunivin, 2024).

Finally, validations for the models annotations mostly rely in precision mesures. Here, we also assessed historical events for external validation as done for the STM model.

3. Data

3.1 Text data retrieval, cleaning and preprocessing

3.1.1 *The New York Times* articles

We retrieved articles from the NYT using Nexis Uni, a database service for universities provided by LexisNexis. This service grants access to NYT articles dating back up to June 1980. Consequently, our dataset includes documents published from this date until December 2024.

To construct the dataset, we designed a query to identified articles containing an economic- and inequality-related term appearing within a five-token window[3]. We refined the selection of keywords through an iterative process, incrementally testing economic and inequality-related terms while assessing false positives using the platform's preview tool (King et al., 2017). For instance, we discarded the keyword *difference* due to its frequent occurrence in unrelated contexts. We downloaded the articles matching the query in batches of 500, as permitted by the platform, resulting in a final dataset of 14.538 articles.

We used the NYT Archive API available on their developer's site to retrieve general metadata for all NYT articles published along the same period. After downloading the monthly data, we aggregated the values to compute yearly totals that allow prevalence calculations. Figure 3.1 shows the annual frequency distribution of all NYT articles.

[3]Query terms: (economic, income, salary, wage, compensation, pay, or wealth) [5-token window] (inequ or gap)

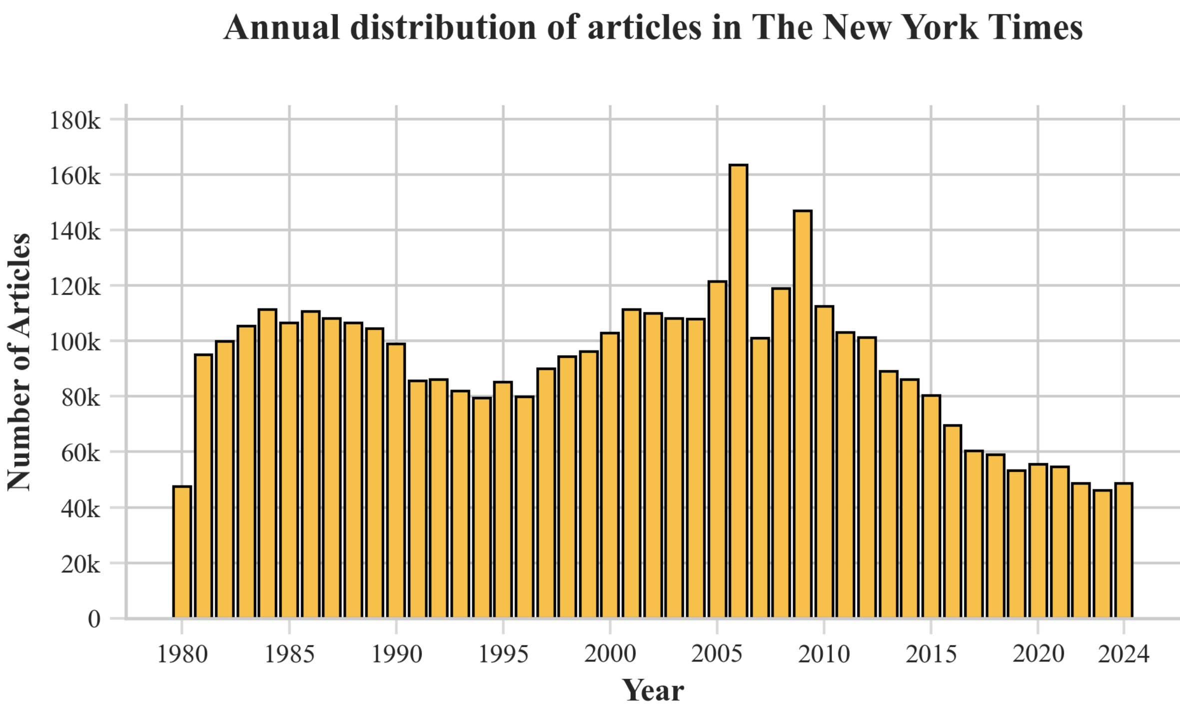

Figure 3.1: Number of articles in the NYT between June of 1980 and 2024 according to NYT Archive API.

To process the economic inequality articles, we first parsed the retrieved `.doc` files to extract the title, body, and publication date, which we later used for analysis.

Duplicates are a usual pitfall in Nexis Uni document extractions. While analyzing the NYT API results, we observed that most duplicates originated from discrepancies between the articles' "main" and "print title". Variations in punctuation, such as differences in apostrophe usage, frequently appeared in the body text of these different versions. To address this, we tokenized the body text and removed punctuation using the spaCy package before assessing text similarity (Honnibal & Montani, 2017). Next, we computed the Jaccard index on the word sets of articles published within a seven-day window (Maier et al., 2018). After testing multiple thresholds, we set the similarity cutoff at 0.80.

Beyond duplicates, we identified erroneous retrievals of book "Best Sellers" lists, which we excluded before the coverage analysis. In total, we detected 3,173 duplicates and 98 incorrectly retrieved articles, with some overlap, leading to a final count of 11,403 articles on economic inequality for coverage analysis.

Further examination of the articles revealed that some consisted of daily or weekly news summaries containing multiple short news, of which only one was related to economic inequality. To detect these cases, we searched for summary cues within the title and body (e.g., "news summary", "the speed read") during document parsing. We excluded these articles from the thematic analysis to prevent overrepresenting topics not originally meant to be related to economic inequality.

Additionally, we removed for the thematic analysis some unusual publications, including long transcripts of political speeches and TV interviews, as well as instances where the article contained only a photo without any text. Filtering out 599 summary articles, 60 transcripts, and 2 articles without a text body resulted in a final set of 10,681 NYT documents for thematic analysis.

3.1.2 United States Congress speeches

We extracted the congressional speeches from 1980 to 2022 using the publicly available repository from Aroyehun et al. (2024) [4]. This dataset compiles speeches collected by Gentzkow et al. (2018) up to 2017 and includes additional data scraped from the Congressional Records website for the years 2018 to 2022 using an automated script. To extend the dataset to 2024, we applied the same automated tool to retrieve the most recent speeches (Judd et al., 2017).

Additionally, we enriched the dataset with complementary speaker attributes, including party affiliation and gender, using publicly available information from the Congressional Records website.

A major challenge in estimating the prevalence of speeches on specific topics within the con-

[4] combined_congress1879_till_2022_filtered_nonprocedural.csv.gzip

gressional data is the presence of procedural speeches. These include pro forma interventions such as "nay", "the objection is heard", or "I reserve the balance of my time", which do not directly contribute to substantive discussions. To ensure a more precise estimation of the coverage of economic inequality, we excluded procedural speeches from the analysis (Card et al., 2022). Since Aroyehun et al. (2024) had already performed this data cleaning for earlier years, we applied the same procedure to the newly downloaded data for 2023 and 2024.

To filter procedural speeches we trained a BERT transformer model (`google-bert/bert-base-uncased`), following the methodology described by Card et al. (2022) to create training and test set labels. We applied the classifier to speeches with lengths between 20 and 420 characters achieving an F1 score of 0.97 on the test set. Procedural speeches accounted for 55% of the total number of speeches. The cleaned congressional dataset (1980–2024) included 2,180,206 speeches, distributed as 51% from Democrats, 48% from Republicans, and less than 1% from other parties (Figure 3.2).

Finally, we extracted economic inequality speeches by emulating the described query retrieval method for the NYT, applying text tokenization and economic and inequality term matching in a 5 word window. We identified 3,777 inequality speeches, 69% from Democrats, 31% from Republicans, and less than 1% from other parties.

3.1.3 Text preprocessing for probabilistic topic modelling

Since topic modeling methods such as STM rely on token probabilities, several text preprocessing steps are needed before model fitting. We followed the guidelines from the prescriptive articles from Maier et al. (2018) and Roberts et al. (2014). We used the "en_core_web_sm" pipeline from spaCy to tokenize and lemmatize both datasets text content (Honnibal & Montani, 2017). We removed stopwords and kept only alphabetic contents in their lowercase form. Additionally, language-distribution qualities make very infrequent tokens usual. These tokens are

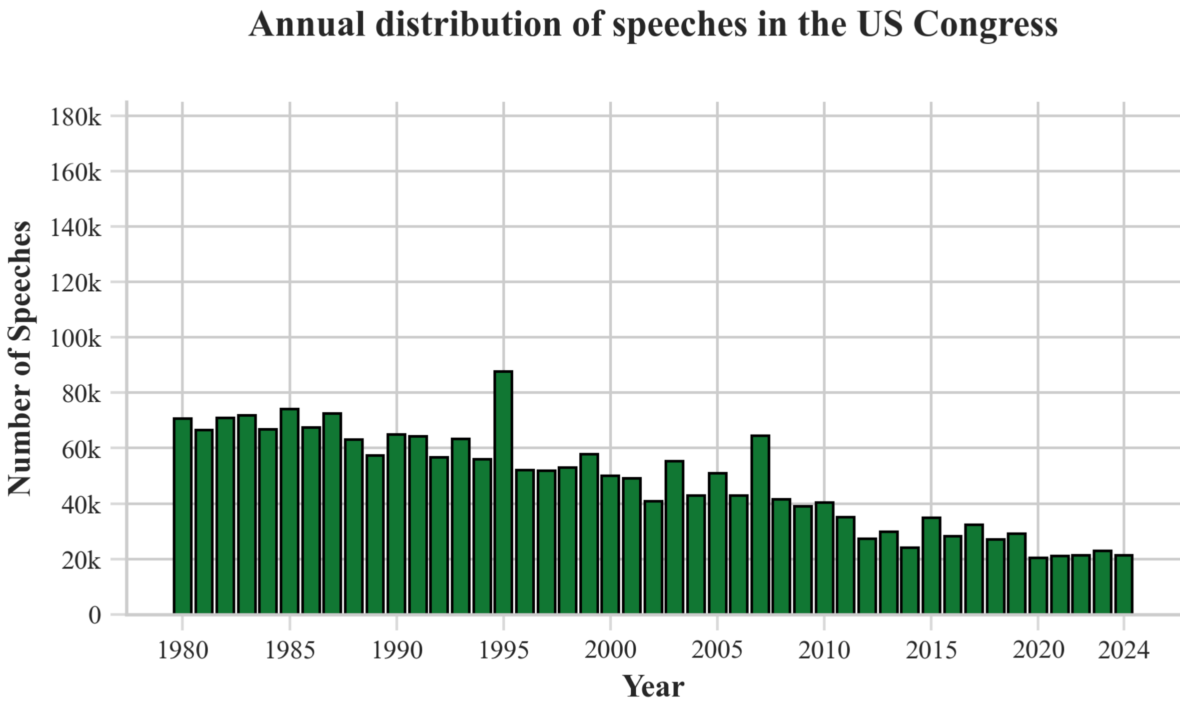

Figure 3.2: Number of speeches for Congressional sessions across both House and Senate between 1980 and 2024. Procedural speeches are excluded.

not informative for word co-occurence methods and take to much higher computational costs due to document-term matrix sparcity. To overcome this issue, we applied relative pruning using the STM package in R (Roberts et al., 2019), removing tokens that appeared in fewer than 1% of the documents. The matrix retained all documents after pruning.

3.2 Objective economic inequality indicators

The Gini index is a widely used economic inequality measure. It assesses "the extent to which the distribution of income or consumption expenditure among individuals or households within an economy deviates from a perfectly equal distribution" (World Bank, 2024). We obtained the yearly values of this indicator from 1980 to 2022 from the World Bank development indicators.

The top 1% wealth concentration is another widely used indicator of economic inequality (Saez & Zucman, 2016), which is conceptually related to the OWS claim "We are the 99%". As an additional robustness check for our results, we extracted the quarterly data for the distribution of household wealth from 1989 Q1 to 2024 Q3. We grouped the quarterly data by year and averaged it to construct a yearly measure (Board of Governors of the Federal Reserve System, 2024).

4. Methods

4.1 Structural Topic Model (STM)

4.1.1 Model parameters and model selection

Given the heterogeneity of this study's datasets, we implemented a STM model as probabilistic topic modeling approach. This model enables the inclusion of covariates in the estimation of the document-topic and topic-word distributions. This flexibility was necessary to allow document-topic prevalences variations along the NYT articles and US Congress speeches (dataset variable) and across time (year variable). We modeled the influence of the year variable using a natural cubic spline with three breakpoints (knots) evenly spaced at 1991, 2002, and 2013. Additionally, since the words with which the topics are addressed may also change depending on the dataset context, we enabled topic-word distributions to vary according to the document source. Since the STM outcomes are sensitive to initialization conditions we implemented the "Spectral" initialization, a reproducible method with a good consistent performance (see Section 2.2.1). We fitted the STMs with the "stm" package for R (Roberts et al., 2019).

To choose the number of topics (K parameter) that asserts the most coherent results for both datasets, we created models with K values between 40 and 75 with steps of 5 topics. For these models we examined the 7 most prominent words and the 15 documents with the highest prevalence to evaluate internal coherence. For instance, when a topic featured words like "employees, overtime, employee, morale, compensation, retention, salary" (Topic 9) we expected to see mostly documents that dealt with labor relations for both datasets. When a topic was accurate for only one of the datasets, we annotated this as inconsistent. We selected the K = 70 model since it provided the most cohesive topics for both datasets and avoided overlapping topics for the same underlying subject.

As final step before assessing model validity, we distinguished theoretically meaningful topics from boilerplate and mixed-topics. We excluded 10 topics for containing boilerplate terms and 11 topics for including mixed-topics (see examples in Figure 4.1). Meaningful topics represented around 60% of the words or more along all the timeframe (Figure A1). The full list of meaningful topics and their clustering is shown in Table A1.

```
Boilerplate topic example

Topic 1
Top words: headline, dan, edit, column, read, sunday.
Explanation: Words related to usual terms in the context of
newspaper articles.

Mixed-topic example

Topic 55
Top words: impeachment, democracy, dictator, rig, electoral,
voting.
Explanation: Top documents on speeches over the constitution,
democracy and voting for the US Congress, and articles on the Trump
impeachment process for the NYT.
```

Figure 4.1: Examples for a boilerplate and a mixed-topic in the STM (K = 70) results.

4.1.2 Validation

To assess the selected topic model internal validity we performed the intrusion task from Chang et al. (2009) (see Section 2.2.1). We retrieved the 5 most prevalent words within each of the 49 STM meaningful topics and randomly selected a 6th intruder word from other topic's top words. For STM topics nested within the same theme, we excluded accompanying topics from the random intruder assignment to reflect the created topic structure.

We designed an online survey using the Potato annotation tool and deployed it on Prolific with a sample of 20 participants (Pei et al., 2022). We requested Prolific to pre-screen participants to ensure they were US residents and native English speakers. The survey included the 49 intrusion tasks presented in randomized order and 3 attention checks. The median completion time was 13 minutes. We excluded two participants from the analysis, one due to incomplete responses and the other for an unusually fast completion time (less than 8 minutes). The intruder detection accuracy was 0.68, significantly exceeding the random baseline of 0.17.

Finally, we assessed the evolution of the topics referring to the external events mentioned in Section 2.2.1. The selected model included a COVID-19 topic that had a surge in 2020 with proportions of 2.4% for NYT and 2.2% for the US Congress. Additionally, in the Congress dataset the taxes topic had a prevalence of over 4% for most of the Reagan's presidency, value that was not surpassed later in the time series. Finally, although we did not hypothesise it previously as a validation check, the selected topic model shielded an environment topic that has been rising in prevalence since 2010 for the NYT data (Appendix A2).

4.2 LLM annotations

After implementing the STM model as a systematic, data-driven method for topic discovery, we annotated the dataset using Llama 3, an open-source LLM. Unlike the STM, which relies on a "bag-of-words" assumption, Llama 3 captures sentence dependencies, allowing for context understanding of the analysed text. Additionally, while STM follows a mixed-membership approach, assigning documents to multiple topics, Llama 3 classified each document into a single category.

4.2.1 Prompt and model implementation

We interpreted the STM topics and nested them hierarchically to develop a codebook with topic and subtopic categories (see Table A1). In Figure 4.2 an example for the Horizontal Inequalities codebook entry is shown. Following previous research, the codebook category descriptions were codebook-centered, meaning that they did not focus on examples, but on instructions (Xiao et al., 2023). We derived the descriptions for the categories from the most representative words within each subtopic and the common themes in the documents with high topic prevalence, as identified by the STM.

`1. Horizontal Inequalities:` **—> TOPIC 1**
`Disparities based on social identity, affecting access to rights, resources, and opportunities.`
`A. Gender: Disparities in the workplace and family or private life based on sex or gender.` **—> SUBTOPIC 1A**
`B. Race: Historical and ongoing racial injustices, including slavery, segregation, incarceration, reparations, and racial liberation.` **—> SUBTOPIC 1B**
`C. Affirmative Action: Policies and measures that promote the access to education, employment and other opportunities to specific social groups.` **—> SUBTOPIC 1C**

Figure 4.2: Example codebook entry for the Horizontal Inequalities topic and its subtopics. Codebook descriptions in bold added for figure clarity.

Additionally, the prompt incorporated chain-of-thought (CoT) reasoning, as recommended by Dunivin (2024). It included a role assignment, a general task description, the codebook, a justification requirement for responses, and output formatting instructions, all provided before the text to be classified (for detailed prompt structure, see Appendix B1). We employed a zero-shot classification approach, where the LLM assigned the categories without prior task-specific training, relying solely on the provided codebook descriptions.

We implemented the Llama 3 open-source model to perform the deductive annotations with the mentioned prompt (`meta-llama/Meta-Llama-3-70B-Instruct`). We selected this model as previous research with a smaller version of it demonstrated higher classification accuracy compared to Mistral, other popular open-source LLM alternative (Halterman & Keith, 2025). We implemented the model in an A100 GPU and initialized it with a 4-bit quantization (FP4). The model followed the output instructions mostly without issues, it produced 310 invalid answers, and it did not assign correct subtopics to 288 cases. We dismissed these cases for further analysis (4.1% of the dataset).

4.2.2 Validation

We validated the Llama annotations by randomly selecting 100 cases for manual review (50 from each dataset). The precision for the NYT annotations was of 0.68 and for the US Congress of 0.64.

Additionally, to assess the precision of the topics used for external validation and those analyzed in the results section, we randomly selected and annotated 20 documents for each category. The precision scores for the validation-related subtopics were as follows: COVID-19 and Taxes both achieved perfect precision (1.00) and the Environment topic reached 0.95. Among the most prevalent categories, the Horizontal Inequality topic had a precision of 0.90, Social Welfare Policy 0.85, Macroeconomic Policy 0.80, Business 0.60, Health 1.00, and Inequality Increase 0.85. Regarding the Horizontal Inequality subtopics, Gender had a precision of 1, and Race of 0.9. These evaluations are documented in the GitHub repository.

As with the STM series, the Llama model results conformed to the external events validations. The COVID-19 topic had a surge in 2020 for both datasets, with a prevalence of 3.7% for the NYT and 2.1% for the US Congress. The discussion on taxes was at its highest values, over 9%, at the beginning of the Reagan administration for the Congress dataset and also showed a local maximum in the NYT dataset. Finally, the environmental topic has steadily increased its prevalence for the last 15 years in the NYT dataset (Appendix B2).

4.3 Statistical analyses and visualizations

We tested our economic inequality coverage hypothesis by computing Pearson correlations between coverage prevalence and year. Furthermore, we calculated the correlations between coverage prevalence and two objective economic inequality indicators: the Gini index and wealth concentration within the top 1% of the distribution.

For the analysis of thematic trends, we applied a six-year sliding window smoothing to the prevalence data from the LLM deductive coding. This technique improves the clarity of medium- and long-term trends, consistent with the focus of our research questions. To account for uncertainty in subtopic prevalence, we calculated 95% confidence intervals by bootstrapping 1,000 samples per year and calculating the corresponding upper and lower bounds. We then smoothed the confidence intervals using the abovementioned procedure.

To analytically assess recent thematic shifts in topics associated with economic inequality, we computed the correlations between the most prevalent topics and the years following the OWS and the Great Recession. We also performed a correlation between the whole period and horizontal inequality to assess the evolution of the prevalence of this theme throughout the whole timeframe. We conducted the latter with results from the STM and Llama annotations for greater robustness. Additionally, we assessed possible topics' trade-offs by calculating crosstopic correlations for the whole timeframe.

Finally, we conducted boxplot visualizations and applied one- and two-sample t-tests to determine whether the plotted trends descriptive features were statistically significant.

5. Results

5.1 Coverage

Figure 5.1 presents the coverage proportions for the economic inequality subject across the NYT and US Congress datasets. The most notable feature in both time series is the break in their predominantly stable trajectories in 2011, coinciding with the onset of the OWS movement. Although coverage between 1980 and 2010 was significantly greater than zero for both datasets (NYT: $t = 19.16$, $p < 0.01$, US Congress: $t = 5.13$, $p < 0.01$), the period from 2011 to 2024 shows a significant increase. Specifically, average coverage rose by 0.79% in the NYT (95% CI [0.56, 1.03], $p < 0.01$) and by 0.31% in the US Congress (95% CI [0.19, 0.43], $p < 0.01$). Notably, the NYT boxplot for 1980–2010 displays an outlier in 2007, the first year of the Great Recession (Figure C1).

Prior to 2011, neither dataset exhibits a pronounced increase or decrease in coverage. Nevertheless, both show a significant positive correlation between coverage and year (1980-2010) (NYT: $r = 0.61$, 95% CI [0.33, 0.80], $p < 0.01$; US Congress: $r = 0.58$, 95% CI [0.26, 0.78], $p < 0.01$), suggesting a gradual upward trend.

When comparing the two sources, there was no statistically significant difference in coverage prevalence between 1980 and 2014. A one-sample t-test on the yearly coverage differences yielded a mean difference not significantly different from zero ($t = 1.11$, $p = 0.28$). However, beginning in 2015, the gap between both sources coverage widened, reaching its largest divergence in 2020, when NYT coverage exceeded that of the US Congress by more than 1%. The 2015–2024 period had a statistically significant mean difference in coverage (mean = 0.64%; 95% CI [0.48, 0.81]; $t = 8.87$, $p < 0.01$). Despite these differences, the trends were strongly correlated for the whole period ($r = 0.84$, 95% CI [0.73, 0.91], $p < 0.01$). Boxplots illustrating

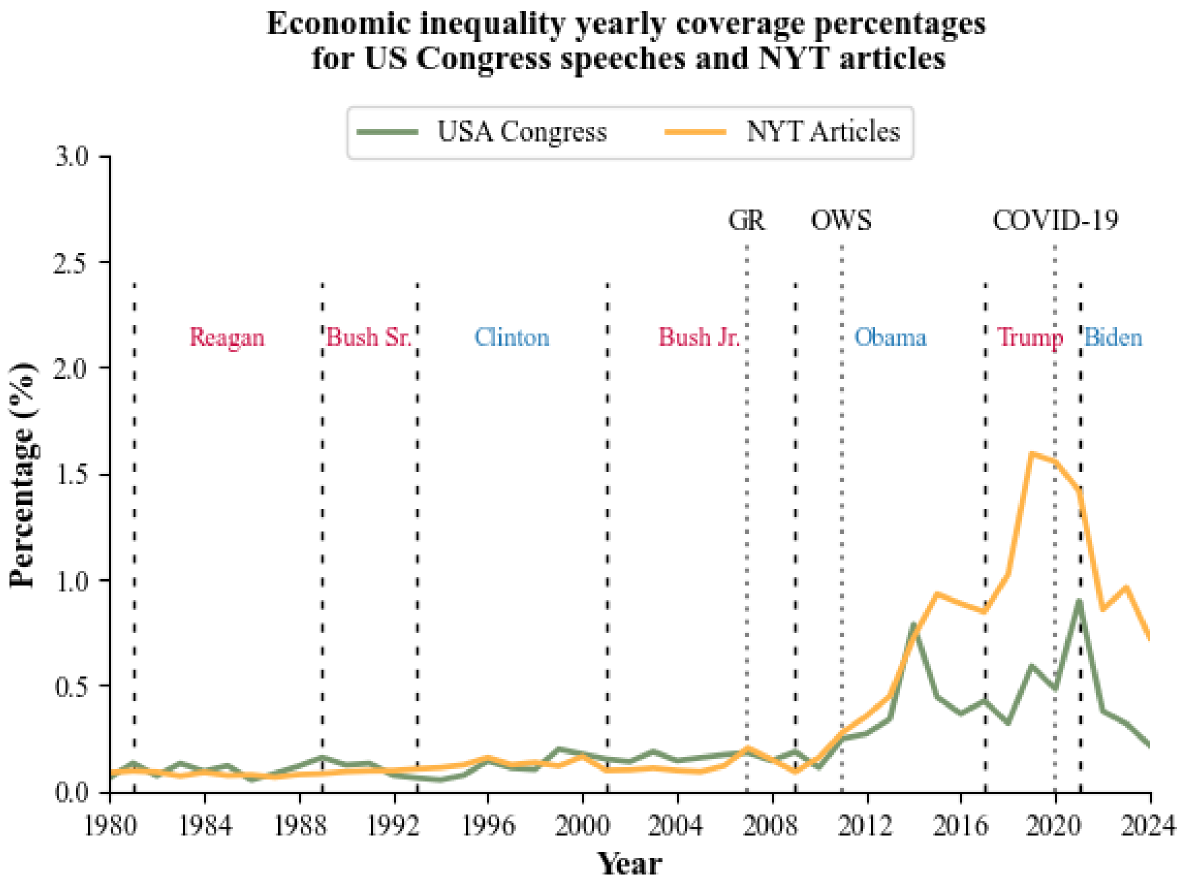

Figure 5.1: Yearly coverage percentages for the economic inequality subject for the US Congress speeches and the NYT articles. GR: Great Recession. OWS: Ocuppy Wall Street movement.

the mean differences for both periods are available in Appendix C3.

In recent years, both datasets have shown a sharp decline from 2021 to 2022. By 2024, while the NYT coverage point estimate remains more than twice as high as in 2011, the Congressional estimate has dropped below the levels observed during the OWS movement year.

Finally, as shown in Table 5.1, both measures of economic inequality produced different correlation strengths with coverage across both datasets. While in the case of the Gini index the relationship is moderate, for the top 1% of wealth accumulation is rather strong. However, correlations may be induced by the fact that they are both increasing trends along time, as can be seen in the "year" column. An analysis with a detrended series would be necessary for better inspection. Figure C4 displays all four trends simultaneously.

Table 5.1: Correlation table for economic inequality objective measures and coverage prevalence of economic inequality in the NYT articles and US Congress speeches

	Gini index	Top 1%	Year
NYT	0.40 [0.12, 0.63]	0.72 [0.51, 0.85]	0.75 [0.57, 0.85]
US Congress	0.44 [0.16, 0.65]	0.67 [0.44, 0.82]	0.73 [0.56, 0.85]

Note. Correlations are computed using Pearson's coefficients. Values in brackets indicate 95% CIs. All correlations are significant with $p < 0.01$.

5.2 Thematic analysis

As mentioned in the methods section, we applied a six-year sliding window smoothing to the thematic analysis graphs and the data used for the correlations. We applied this technique to improve clarity in identifying medium and long-term trends, aligning with the focus of our research question. However, it is important to note that years with prevalence peaks, such as those driven by discussions of specific congressional bills, may not be accurately captured.

For more insights on the original STM-topics later interpreted and clustered in the codebook categories, see Table A1.

5.2.1 General topics overview

The most prominent topics found through the codebook-based LLM annotations were Horizontal Inequality, Macroeconomic Policy, Social Welfare Policy, Business, Health and Inequality Statistics (Figure 5.2). However, Horizontal Inequality, Macroeconomic Policy and Social Welfare Policy were the most prominent for both datasets, surpassing a prevalence of 10% in almost all years and therefore entailing together more than a 40% of the covered topics along time. Regarding the results of the two sources, the NYT appeared to have a more stable topic prevalence behavior than the US Congress.

In the NYT dataset, since 2011, the abovementioned most prominent topics have undergone some variation. While Horizontal Inequality has a relative increase in prevalence within eco-

Table 5.2: Most prevalent topics descriptions and examples

Topic	Description	Top words	NYT example	Congress example
Horizontal Inequality	Documents that address disparities based on social identity and their impact on group rights and opportunities. The gender pay gap and advances in African American rights are the primary focus within this category, while discussions on other minority groups are much less prominent.	breadwinner, pregnant, luther king, apartheid	A Man's Place - Correction Appended When the subject is women's economic progress, it's easy to get lost in the controversies of the moment.	Mr. Speaker. I think it is absolutely appalling. irresponsible. and downright unethical. for a college or university president to say lowtest scores of African American students are linked to their genetic.
Macroeconomic Policy	Documents on fiscal balance, budget, taxation, and economic expansion strategies within the US. Additionally, it includes discussions on macroeconomic policies in other countries and regions, such as China and Europe.	discretionary, taxation, inflation, trade	Economists Sharply Split Over Trade Deal Effects WASHINGTON – Lawmakers and presidential candidates are having their say about the 12-nation Pacific Rim trade accord that is President Obama's top economic priority in his final year in office.	Mr. Speaker, I rise in opposition to this legislation, and perhaps for no better reason than it is a $270 billion cost that the Congressional Budget Office showed with no pay-fors, no offsets in the Federal budget.
Social Welfare Policy	Documents on the challenges faced by the unemployed, workers, the poor, and the middle class, as well as policy strategies aimed at improving their socioeconomic conditions (e.g., raising the minimum wage, entitlements). This category also includes union demands and ideological debates on right- and left-leaning policy approaches.	hunger, bargaining, earner, communism	A Pay Raise's Impact The Clinton Administration, which appears to be advocating a higher minimum wage as one of its major policy objectives.	Mr. President, I am here to speak out in favor of working families and how we can empower American workers to obtain good jobs, to secure a safe retirement after a lifetime of hard work.

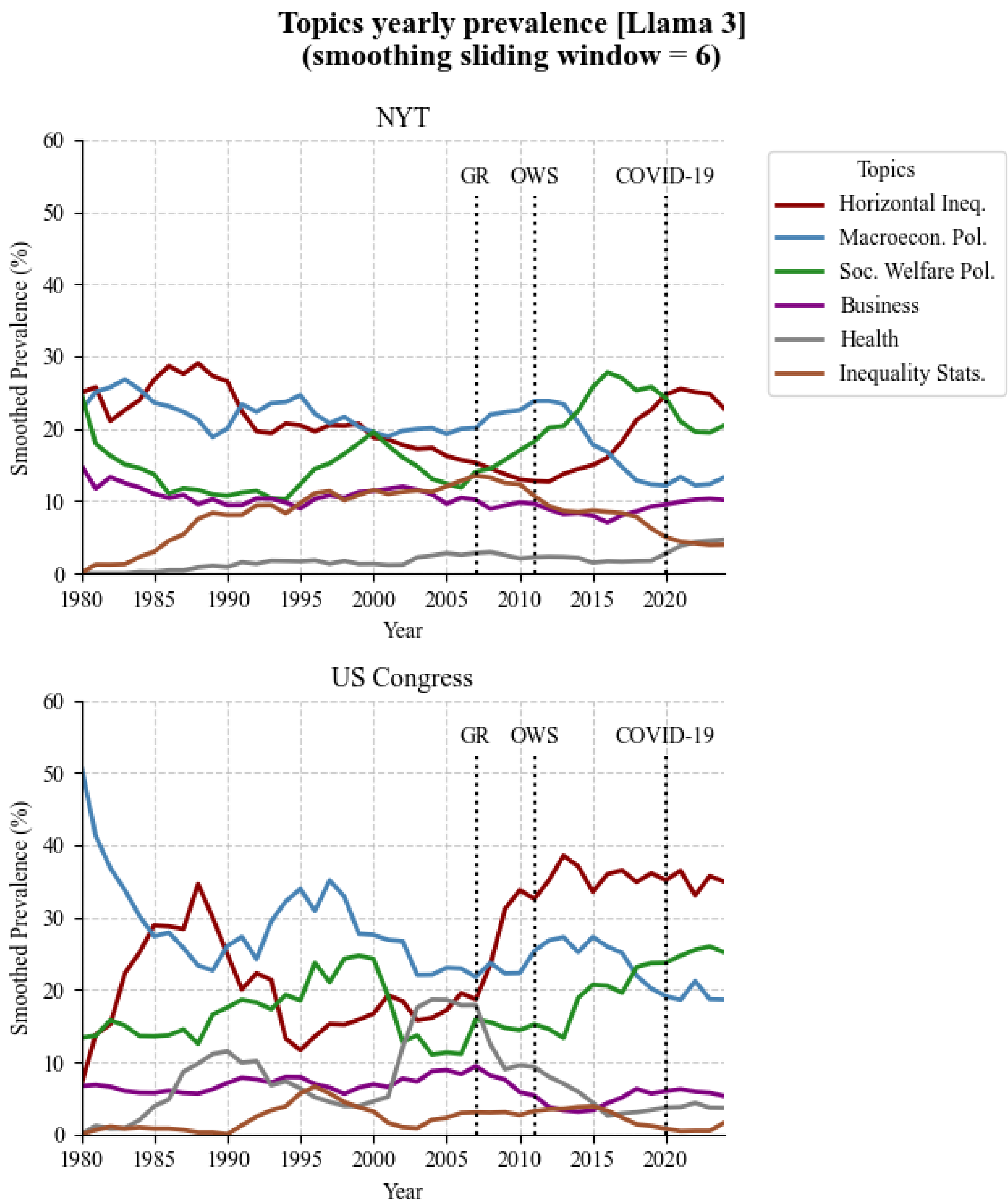

Figure 5.2: Yearly prevalence for most covered themes associated to economic inequality in the NYT articles and US Congress speeches. GR: Great Recession. OWS: Ocuppy Wall Street movement.

nomic inequality coverage (r = 0.94, 95% CI [0.82, 0.98], $p < 0.01$), Macroeconomic Policy has shown a strong decline (r = -0.90, 95% CI [-0.97, -0.71], $p < 0.01$). Social Welfare Policy has fluctuated, with an overall negative relation with time, but a broad confidence interval (r = -0.56, 95% CI [-0.84, -0.03], $p < 0.01$).

The crosscorrelation between the three trends along the whole period was only significant for Macroeconomic Policy and Social Welfare Policy, which were negatively related (r = -0.57, 95% CI [-0.74, -0.33], p = 0.04).

For the US Congress, the topics' time series exhibit greater volatility. The clearest example is Horizontal Inequality, which follows a period of steady increase, followed by a decline from the start of the series until 1995. Additionally, after 2007, it undergoes an upward level shift, rising from a mean of 20% for the 1980-2007 period to over 35% for 2008-2024 (mean difference = 16%, 95% CI [9, 23], $t = 4.51$, $p < 0.01$; see boxplots in Figure C5).

Great variation is also observed for the Health category, where two distinct cycles of growth and decline can be identified: one spanning the 1980s and 1990s, and another between 2000 and approximately 2010 (for insights in these congressional speeches see Appendix D).

Turning to the recent developments of the two other most prevalent themes in the US Congress, Social Welfare Policy has shown a clear upward trend since 2007 (r = 0.95, 95% CI [0.87, 0.98], $p < .01$). In contrast, Macroeconomic Policy has remained comparatively stable, exhibiting a moderate negative correlation with year (r = -0.58, 95% CI [-0.83, -0.14], p = .01).

Finally, the topics crosscorrelation was only significant for Horizontal Inequality and Macroeconomic Policy, which were negatively related (r = -0.64, 95% CI [-0.79, -0.42], $p < 0.01$).

Other topics, such as Education, International Issues, Debt & Housing, and Environmental Issues, are also present in the dataset,s but with lower prevalence. The full visualization of topic prevalence can be accessed in the thesis GitHub repository.

5.2.2 Horizontal Inequality: Gender and Race

Horizontal Inequality was one of the most prevalent topics in both datasets. However, correlations between topic prevalence and year over the 1980–2024 period revealed divergent behaviors. In the NYT, results indicated a moderate downward trend, with correlations of $r = -0.43$ for LLM coding (95% CI [–0.64, –0.15]) and $r = -0.41$ for STM annotation (95% CI [–0.62, –0.13]). In contrast, the Congress dataset showed a strong upward trend, with correlations of $r = 0.62$ (95% CI [0.41, 0.78]) and $r = 0.70$ (95% CI [0.51, 0.82]) for LLM and STM, respectively. Correlations between the STM and LLM results were also very high for both datasets (NYT: $r = 0.98$, 95% CI [0.96, 0.99]; US Congress: $r = 0.96$, 95% CI [0.92, 0.98]). All reported correlations were statistically significant ($p < .01$).

Going deeper into these trend components, Figure 5.3 illustrates the differences in subtopic prevalence in the two datasets. In the NYT, Gender and Race have followed overlapping trajectories, with Race showing higher point estimates for most of the studied period. In contrast, the US Congress discussions on economic inequality have had a greater prevalence of the Gender category.

Additionally, all three peaks in the Horizontal Inequality time series correspond to increases in the prevalence of Gender. The two major peaks in Congress occurred in 1985 and 2009, driven by the Federal Equitable Pay Practices Act and the Paycheck Fairness Act (also known as the Lilly Ledbetter Fair Pay Act), respectively. Interestingly, while the NYT shows a peak around the 1985 event, primarily featuring articles on equal pay, no similar increase is observed in 2009.

Regarding recent years, the NYT dataset shows a continuous increase in the prevalence of Race since 2011 ($r = 0.97$, 95% CI [0.91, 0.99], $p < .01$). A similar trend is evident in the US Congress dataset, where Race displays the steepest upward trajectory in the studied period ($r = 0.97$, 95% CI [0.90, 0.99], $p < .01$). In 2014, Race accounted for approximately 5% of Congressional coverage, reaching nearly 20% by 2024, matching the level of Gender coverage over the past three years.

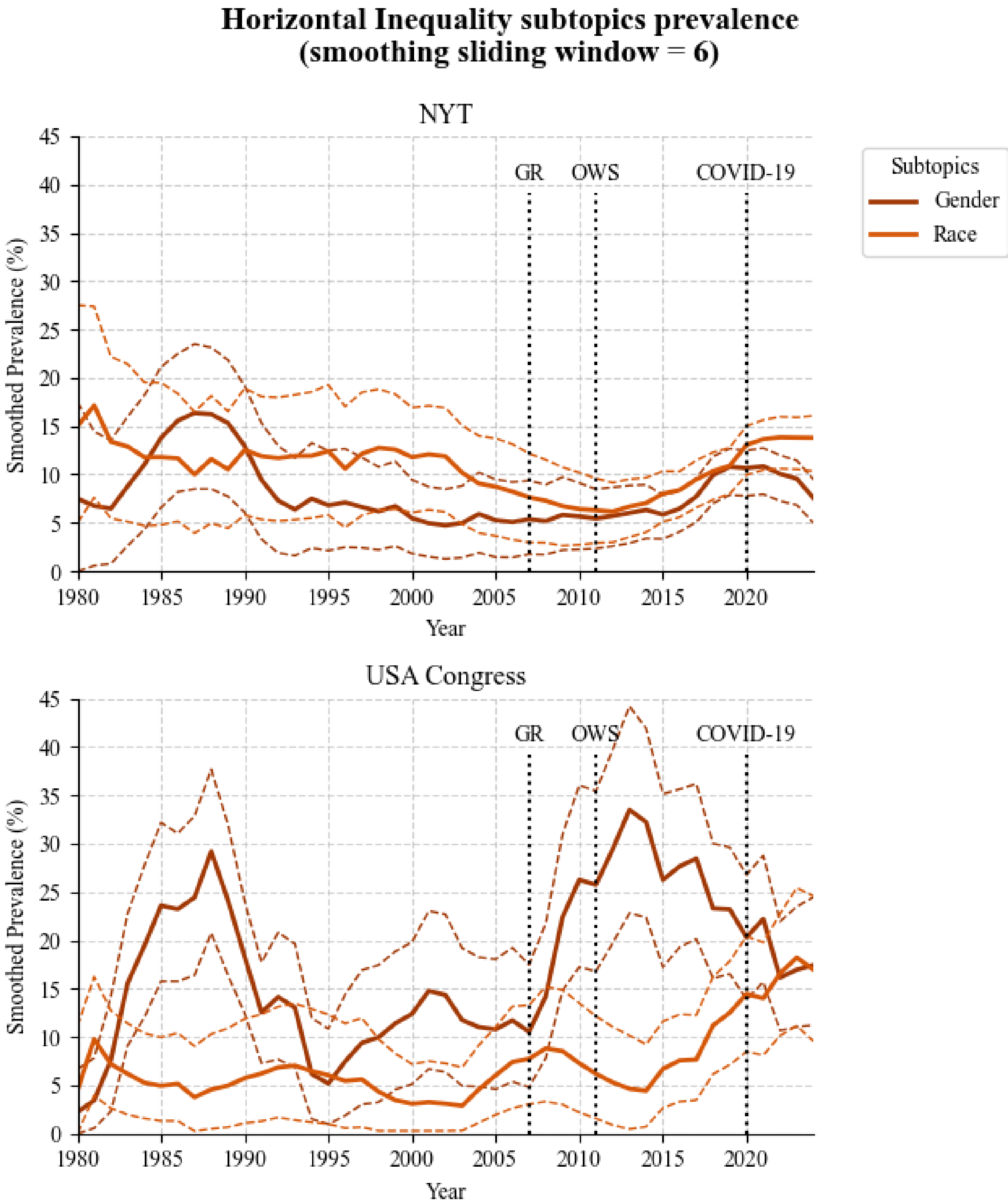

Figure 5.3: Yearly prevalence for the gender and race subtopics within the Horizontal Inequality topic. Annotations done by Llama 3 model with a codebook prompt. GR: Great Recession. OWS: Ocuppy Wall Street movement.

The US Congress subtopic trends provide further information into the overall stability of the Horizontal Inequality trend over the last decade. While its aggregate prevalence has remained steady, this stability results from the mentioned increase in the prevalence of Race and a simultaneous decline in Gender within the category, however, there is no significant negative correlation between Gender and year for this short time period (r = -0.23, 95% CI [-0.67, 0.34], p = .42).

6. Discussion

Because economic inequality within the broader population distribution is not directly observable in everyday life, it is essential to examine how the issue is discussed the public debate (Grisold & Preston, 2020). To contribute to this line of research, this thesis investigated how economic inequality has been addressed over time by two influential actors in US public discourse: the legacy media outlet the NYT and the US Congress. Specifically, we analyzed the degree of attention devoted to the issue and the themes most frequently associated with it over a 45-year period (1980–2024). For this analysis, we employed computational methods suited for exploratory topic discovery and corpus annotation.

Contrary to our H1.1 hypothesis, coverage of economic inequality exhibited a modest upward trend between 1980 and 2010 in both datasets. Still, between 2011-2014, the topic's prevalence more than doubled, marking an unparalleled break from previous patterns. These findings align with prior research identifying the OWS movement as a turning point in elevating the visibility of economic inequality in public discourse (Baumann & Majeed, 2020; McGovern et al., 2020). Additionally, it stresses the relevance of social movement claims since economic crises like the Great Recession (2007–2009) did not appear to significantly alter coverage patterns, aside from a small surge in NYT coverage in 2007. This observation also corresponds with earlier studies reporting only a brief spike in coverage during that year (Baumann & Majeed, 2020; McGovern et al., 2020).

During the most recent four-year period, coinciding with the Biden administration, coverage appears to have declined in both datasets. Due to the limited number of data points, however, no statistical tests were performed. Notably, by 2024, the estimated prevalence of economic inequality in Congressional speeches had dropped even below the level observed in 2011.

We also explored the relationship between coverage prevalence and objective indicators of economic inequality. Contrary to our H1.2 hypothesis, we found a relation between coverage and economic inequality indicators in both datasets. The strength of this association varied considerably depending on the indicator used. In particular, wealth concentration among the top 1% showed a stronger relation with coverage than the Gini index. This result is noteworthy as it relates to one of the central claims of the OWS movement. Awareness of this specific indicator may have grown among journalists and policymakers since 2011, leading to greater interest in the topic. Nevertheless, further analysis using detrended time series would be necessary to draw more robust inferences.

In mass media, communicators choose which dimensions of a subject to bring to the forefront and which to omit (Stecula & Merkley, 2019). With respect to our research questions on the thematic associations of economic inequality, three dominant topics consistently emerged across both datasets: horizontal inequalities, macroeconomic policy, and social welfare policy. Together, these accounted for over 40% of the thematic coverage over time.

Previous research has documented a rise in media attention to issues such as the minimum wage, income inequality, the middle class, and welfare —all of which fall within our Social Welfare Policy category— following the OWS movement (Gaby & Caren, 2016). In the NYT dataset, while coverage of social welfare topics increased between 2011 and 2015, this trend reversed in subsequent years, resulting in a slight overall decline in prevalence after 2011. In contrast, the US Congress dataset shows a steady and sustained increase in the coverage of social welfare issues over the past 15 years.

Our findings regarding hypothesis H2 are therefore mixed. While the NYT showed an increase in social welfare policy coverage after 2011, it was not sufficient to establish a sustained long-term trend. In contrast, the US Congress demonstrated a continuous rise in the prevalence of this theme over the same period. Additionally, in the NYT data, the prevalence of social welfare policies was negatively associated with that of macroeconomic policy topics, suggesting a potential trade-off: as attention to social welfare policies in the media increase, coverage of macroeconomic concerns tended to decrease.

Finally, with respect to our third research question concerning the prevalence of horizontal inequalities, this topic maintained consistently high coverage across the entire period in both datasets. While the overall trend across 1980–2024 was negative for the NYT, its coverage of horizontal inequalities has increased markedly since 2011. These findings contradict earlier research suggesting a decline in media attention to group-based inequalities following the OWS movement (Gaby & Caren, 2016). However, they align with recent studies that report increased mentions of social identities and identity politics in the media after 2011 (Amira & Abraham, 2022; Hopkins et al., 2024).

For the US Congress, on the other hand, the prevalence of horizontal inequality themes increased over the whole period. Over the past 15 years, approximately one-third of congressional speeches addressing economic inequality also referred to horizontal inequalities.

In sum, our findings contradict hypothesis H3. Rather than declining, NYT coverage of horizontal inequalities increased following the OWS movement. In the case of the US Congress, attention to this theme remained consistently prominent, with the greatest levels of prevalence observed across the entire study period.

This rise in coverage of horizontal inequalities appears to be driven by different subtopic dynamics in each dataset. For the NYT, the increase is primarily associated with the growing prominence of gender and race themes. In contrast, within the US Congress, the increase is more closely linked to discussions of racial disparities in the context of economic inequality.

Despite its contributions, this thesis has several limitations that should be acknowledged. First, the analysis was restricted to a single US legacy media outlet due to constraints in the available database. For a more comprehensive assessment, future research should incorporate additional newspapers representing a broader range of ideological perspectives, as well as social media platforms, which play an increasingly significant role in shaping public discourse. Second, the computational methods employed to retrieve relevant documents and identify associated themes are more effective at capturing explicit references to economic inequality, potentially overlooking more implicit or nuanced expressions. Third, prompt-based LLM approaches for deductive coding remain a developing area and require more extensive prompt refinement and parameter tuning to improve annotation accuracy. In this study, such iterative tuning was limited by computational constraints, which also precluded the implementation of a mixed-membership classification in the LLM-based annotation. Finally, the analysis of temporal trends primarily relied on correlations. To improve the robustness of such analyses, future research should employ detrended time series techniques to assess the relationship between coverage and objective economic indicators more accurately.

7. Conclusion

Our analysis of economic inequality coverage in both the NYT and the US Congress reveals a gradual increase since 1980. Nevertheless, we contribute to the research that highlights that 2011 —the year of the OWS movement— was a critical inflexion point, marking a substantial increase in attention to the issue across both media and political discourse. Furthermore, by leveraging the extended time frame of this study, we also observed a decline in coverage during the most recent years, coinciding with the Biden administration (2021–2024).

Although a range of themes are addressed in the context of economic inequality, group-based disparities, macroeconomic considerations, and policies aiming at reducing inequality have consistently emerged as the most prominent across both discursive contexts. Recent developments, however, indicate a shift in emphasis: discussions on group-based inequalities and social welfare policies have gained prevalence, while attention to macroeconomic constraints has declined.

This thesis also contributes to the understanding of how frequently different social groups are represented within the broader horizontal inequalities category. Our findings suggest that racial inequalities are receiving increasing attention, whereas gender disparities appear to be less prominently discussed in recent years. Whether this reflects a sustained trend or a temporary shift remains to be seen. Future research should assess whether attention to gender resurfaces in a new attention cycle, as observed during the 1980s and 2000s.

References

Allen, M. (Ed.). (2017). *The SAGE encyclopedia of communication research methods: 4* (Online-Ausg). SAGE Publications, Inc.

Amira, K., & Abraham, A. (2022). How the media uses the phrase “identity politics”. *PS: Political Science & Politics*, 55(4), 677–681.

Aroyehun, S. T., Simchon, A., Carrella, F., Lasser, J., Lewandowsky, S., & Garcia, D. (2024, May 12). Computational analysis of US congressional speeches reveals a shift from evidence to intuition. Retrieved March 16, 2025, from http://arxiv.org/abs/2405.07323

Baumann, S., & Majeed, H. (2020). Framing economic inequality in the news in canada and the united states. *Palgrave Communications*, 6(1), 42.

Bernhard, J., Teuffenbach, M., & Boomgaarden, H. G. (2023). Topic model validation methods and their impact on model selection and evaluation. *Computational Communication Research*, 5(1), 1.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.

Board of Governors of the Federal Reserve System. (2024, December 20). *Distributional financial accounts: Distribution of household wealth in the u.s. since 1989*. https://www.federalreserve.gov/releases/z1/dataviz/dfa/distribute/chart/#quarter:140;series:Net%20worth;demographic:networth;population:1,3,5,7,9;units:shares;range:1989.3,2024.3

Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., Abramitzky, R., & Jurafsky, D. (2022). Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31), e2120510119.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd international conference on neural information processing systems* (pp. 288–296). Curran Associates Inc. https://dl.acm.org/doi/10.5555/2984093.2984126

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. W. (2023). What we can do and cannot do with topic modeling: A systematic review. *Communication Methods and Measures*, 17(2), 111–130.

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023, June 23). LLM-assisted content analysis: Using large language models to support deductive coding. Retrieved March 4, 2025, from http://arxiv.org/abs/2306.14924

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). BERT: Pre-training of deep bidirectional transformers for language understanding. Retrieved March 16, 2025, from http://arxiv.org/abs/1810.04805

Diermeier, M., Goecke, H., Niehues, J., & Thomas, T. (2017). *Impact of inequality-related media coverage on the concerns of the citizens* (Discussion Pape No. 258, ISBN 978-3-86304-257-8,). Heinrich Heine University Düsseldorf, Düsseldorf Institute for Competition Economics (DICE). Düsseldorf. https://hdl.handle.net/10419/162781

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, 41(6), 570–606.

Dunivin, Z. O. (2024, February 12). Scalable qualitative coding with LLMs: Chain-of-thought reasoning matches human performance in some hermeneutic tasks. Retrieved March 4, 2025, from http://arxiv.org/abs/2401.15170

Edwards, G. C., & Wood, B. D. (1999). Who influences whom? the president, congress, and the media. *American Political Science Review*, 93(2), 327–344.

Eshbaugh-Soha, M., & McGauvran, R. J. (2018). Presidential leadership, the news media, and income inequality. *Political Research Quarterly*, 71(1), 157–171.

Franko, W. W. (2017). Understanding public perceptions of growing economic inequality. *State Politics & Policy Quarterly*, 17(3), 319–348.

Gaby, S., & Caren, N. (2016). The rise of inequality: How social movements shape discursive fields. *Mobilization: An International Quarterly*, 21(4), 413–429.

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2018, January 16). Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. Retrieved October 10, 2024, from https://data.stanford.edu/congress_text

Grisold, A., & Preston, P. (2020, September 14). News media and economic inequality: Reflections and requisite reforms. In A. Grisold & P. Paschal (Eds.), *Economic inequality and news media* (1st ed., pp. 189–212). Oxford University Press. Retrieved March 4, 2025, from https://academic.oup.com/book/33644/chapter/288173362

Halterman, A., & Keith, K. A. (2025, January 9). Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. Retrieved March 4, 2025, from http://arxiv.org/abs/2407.10747

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Hopkins, D. J., Lelkes, Y., & Wolken, S. (2024). The rise of and demand for identity-oriented media coverage. *American Journal of Political Science*, ajps.12875.

Jetten, J., Peters, K., Álvarez, B., Casara, B. G. S., Dare, M., Kirkland, K., Sánchez-Rodríguez, Á., Selvanathan, H. P., Sprong, S., Tanjitpiyanond, P., Wang, Z., & Mols, F. (2021). Consequences of economic inequality for the social and political vitality of society: A social identity analysis. *Political Psychology*, 42, 241–266.

Judd, N., Drinkard, D., Carbaugh, J., & Young, L. (2017). *Congressional-record: A parser for the congressional record*. @unitedstates. Retrieved January 16, 2025, from https://github.com/unitedstates/congressional-record

King, G., Lam, P., & Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, 61(4), 971–988.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2), 93–118.

McCall, L. (2013, March 29). *The undeserving rich: American beliefs about inequality, opportunity, and redistribution* (1st ed.). Cambridge University Press.

McGovern, P., Obradovic, S., & Bauer, M. W. (2020). *Income inequality and the absence of a tawney moment in the mass media,* (No. 53). International Inequalities Institute, London School of Economics and Political Science. London, UK. https://eprints.lse.ac.uk/107535/

Meta. (2024, July 23). *The llama 3 herd of models*. Retrieved March 16, 2025, from https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1), 202–237.

Osnabrügge, M., Hobolt, S. B., & Rodon, T. (2021). Playing to the gallery: Emotive rhetoric in parliaments. *American Political Science Review*, 115(3), 885–899.

Pei, J., Ananthasubramaniam, A., Wang, X., Zhou, N., Dedeloudis, A., Sargent, J., & Jurgens, D. (2022). Potato: The portable text annotation tool. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Proksch, S.-O., & Slapin, J. B. (2012). Institutional foundations of legislative speech. *American Journal of Political Science*, 56(3), 520–537.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016, January 31). Navigating the local modes of big data: The case of topic models. In R. M. Alvarez (Ed.), *Computational social science* (1st ed., pp. 51–97). Cambridge University Press. Retrieved March 16, 2025, from https://www.cambridge.org/core/product/identifier/CBO9781316257340A009/type/book_part

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). **stm**: An *R* package for structural topic models. *Journal of Statistical Software*, 91(2).

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082.

Saez, E., & Zucman, G. (2016). Wealth inequality in the united states since 1913: Evidence from capitalized income tax data. *The Quarterly Journal of Economics*, 131(2), 519–578.

Stecula, D. A., & Merkley, E. (2019). Framing climate change: Economics, ideology, and uncertainty in american news media content from 1988 to 2014. *Frontiers in Communication*, 4, 6.

Stewart, F. (2009). Horizontal inequality: Two types of trap. *Journal of Human Development and Capabilities*, 10(3), 315–340.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August 2). Attention is all you need. Retrieved March 16, 2025, from http://arxiv.org/abs/1706.03762

Vaughan, M. (2024, October 17). *Communication about economic inequality in hybrid media: A systematic review*.

World Bank. (2024, September 19). *GINI index for the united states [SIPOVGINIUSA], retrieved from FRED, federal reserve bank of st. louis* [Federal reserve economic data]. Retrieved March 15, 2025, from https://fred.stlouisfed.org/series/SIPOVGINIUSA

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. *28th International Conference on Intelligent User Interfaces*, 75–78.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1), 237–291.

Appendix

A. STM topic structure and validations

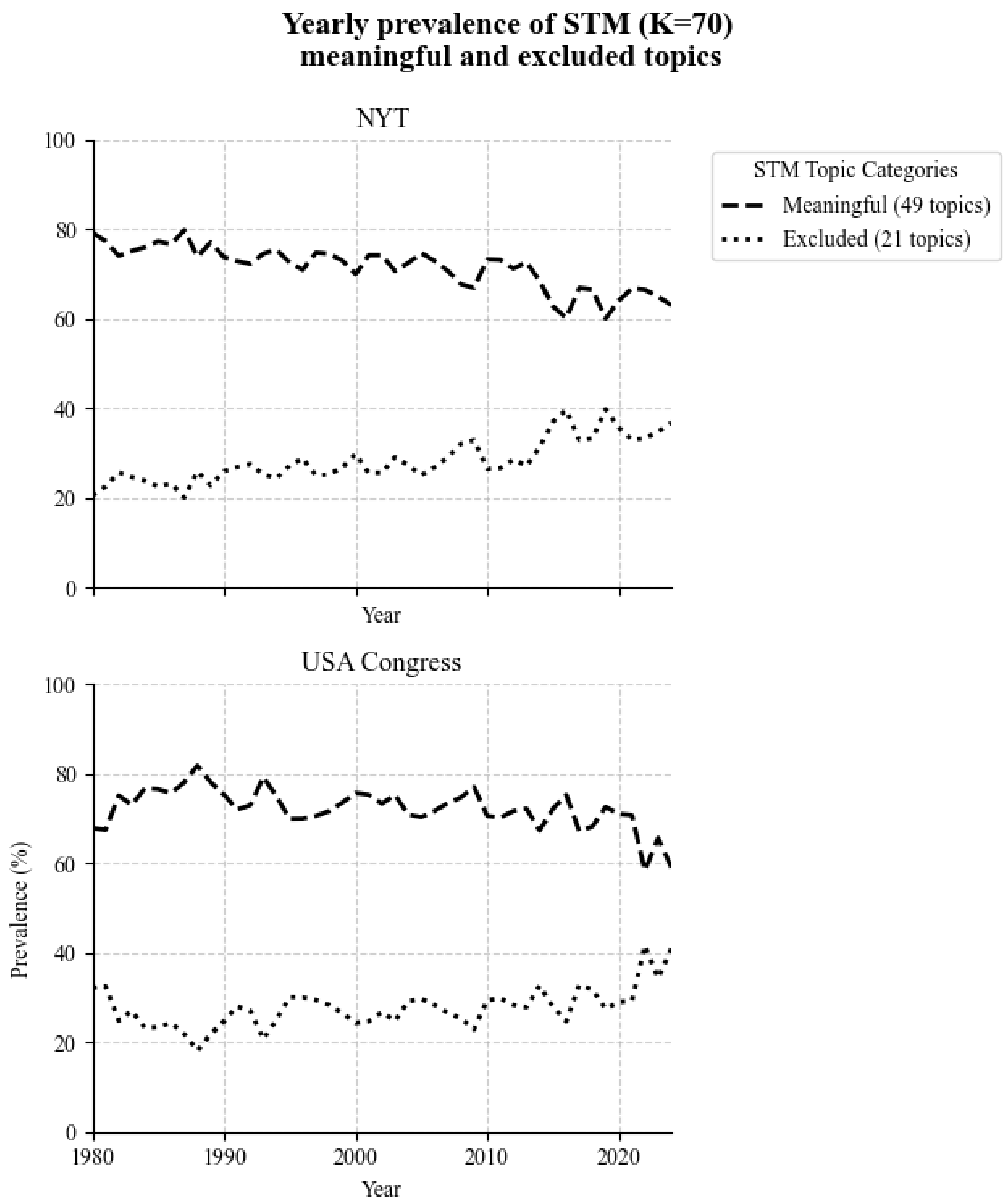

Figure A1: Yearly prevalence of STM (K=70) meaningful and excluded topics.

Table A1: STM (K = 70) results and interpretation for meaningful topics.

n°	Topic	Subtopic	Code	Top 3 words
3	horiz inequalities	gender	1A	ledbetter, breadwinner, discrimination
35	horiz inequalities	gender	1A	occupational, occupation, comparable
43	horiz inequalities	gender	1A	breast, pregnant, pregnancy
51	horiz inequalities	gender	1A	amendment, ratify, constitution
24	horiz inequalities	race	1B	luther, king, commemorate
31	horiz inequalities	race	1B	apartheid, africa, crow
52	horiz inequalities	affirmative action	1C	affirmative, discrimination, discriminatory
22	macroeconomics	budget fiscal balance	2A	discretionary, baby, reduce
12	macroeconomics	taxes	2B	deduction, taxation, bracket
25	macroeconomics	monetary recession	2C	advisers, inflation, summers
68	macroeconomics	trade tariffs	2D	tariff, trade, trading
29	stats trends	census statistics	3A	median, hispanic, hispanics
37	stats trends	inequality rise	3B	redistribution, earner, wealth
14	debt housing	debt housing	4A	voucher, assistance, funding
41	debt housing	debt housing	4A	borrower, lender, loan
61	debt housing	debt housing	4A	rental, homeowner, renter
45	businesses corporations	small business innovation	5A	entrepreneurship, entrepreneur, entrepreneurial
54	businesses corporations	merges antitrust big tech	5B	merger, antitrust, monopoly
38	businesses corporations	corporate compensations	5C	shareholder, stock, securities
62	businesses corporations	workplace conditions	5D	customer, lawsuit, ward
9	public employment nyc	public employment	6A	employees, overtime, employee
30	public employment nyc	public employment	6A	readiness, pentagon, personnel
10	public employment nyc	nyc politics	6B	brooklyn, city, mayor
42	welfare idiology	poverty relief	7A	hunger, nutrition, jobless
20	welfare idiology	unions	7B	workers, organizing, bargaining
6	welfare idiology	wages	7C	minimum, hourly, earner
4	welfare idiology	idiology discussions	7D	communism, faction, liberty
7	civic engagement	protest police	8A	policing, detain, police
34	civic engagement	charity community	8B	library, volunteer, nonprofit

Continued on next page

(Continued from previous page)

n°	Topic	Subtopic	Code	Top 3 words
18	civic engagement	religious leaders	8C	prayer, christian, religious
8	health	health insurance	9A	physician, reimbursement, hospital
16	health	health insurance	9A	deductible, drug, medication
49	health	health insurance	9A	uninsured, obamacare, affordable
56	health	covid19	9B	pandemic, covid, vaccine
28	education	early childhood educ	10A	kindergarten, literacy, childhood
13	education	school administration learning	10B	schools, superintendent, charter
17	education	colleges universities	10C	bachelor, graduation, undergraduate
47	education	teacher salary capacitation	10D	teacher, teaching, teachers
15	international	militar conflicts	11A	south, korean, north
44	international	militar conflicts	11A	missile, soviet, diplomatic
53	international	militar conflicts	11A	israel, assassination, jews
59	international	militar conflicts	11A	greece, ireland, greek
26	international	china politics	11B	china, taiwan, chinese
63	international	latam democracy aid	11C	repression, caribbean, latin
60	environment	environ pollution	12A	carbon, greenhouse, fossil
27	arts entertainment	free time events	13A	arts, artistic, musical
65	arts entertainment	broadcast social media	13B	broadcast, cable, funny
70	arts entertainment	theatre awards	13C	actor, star, adapt
19	sports	sports	14A	soccer, baseball, athlete

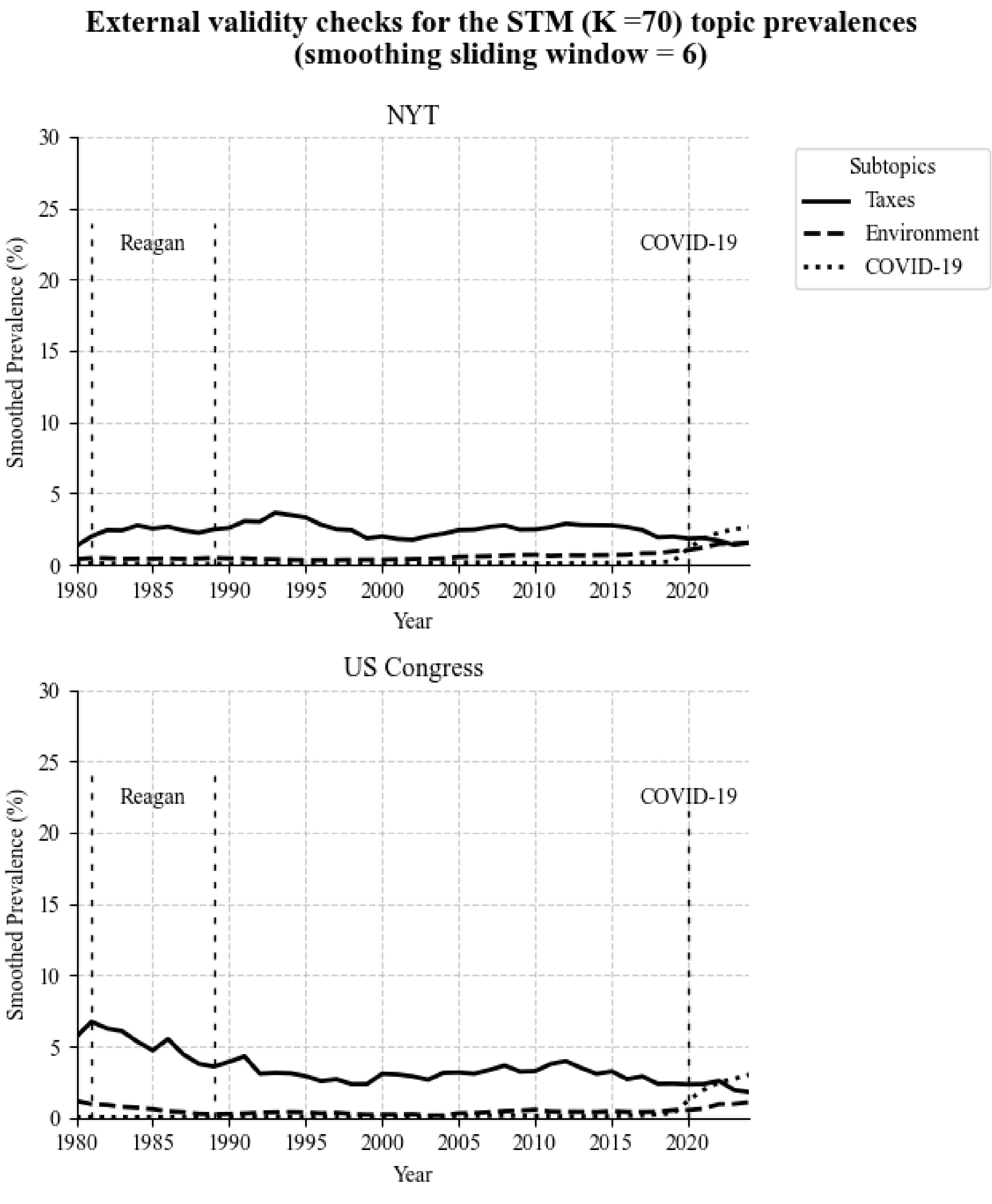

Figure A2: Prevalence of topics used as external validity measures along time for the STM (K = 70) model.

B. Llama prompt and validations

```
<begin_of_text_token>You are a classifier for economic
inequality-related texts based on a codebook.  You will receive a
text from the {text_type} (year:  year) that deals with economic
inequality, and your task is to classify it according to the most
prominent topic and subtopic.  Use the codebook provided below.

## Topic and Subtopic Codebook:  {ecoineq_topics_codebook}

## Instructions:  1.  Justify:
Provide a justification of why you applied the selected code.
2.  Determine the TOPIC:
Select the single primary topic (numbered 1 to 14) that best matches
the text’s main theme based on the descriptions provided below.
Every text has a topic, choose the one that most closely aligns with
the dominant message or has according keywords.
3.  Determine the SUBTOPIC:
Select the corresponding primary subtopic (letter A to D). Use
’none’ as subtopic if no subtopic aligns with the text within the
chosen topic.

## JSON Output description:
{json_examples}
- When multiple topics are present, classify the text based on the
most emphasized topic or the one most strongly represented by the
keywords and context.
- Do not answer anything additional to the given formatted JSON.

## Text to Classify:
{text}

## JSON Output:
```

Figure B1: Prompt structure used for the Llama 3 annotations.

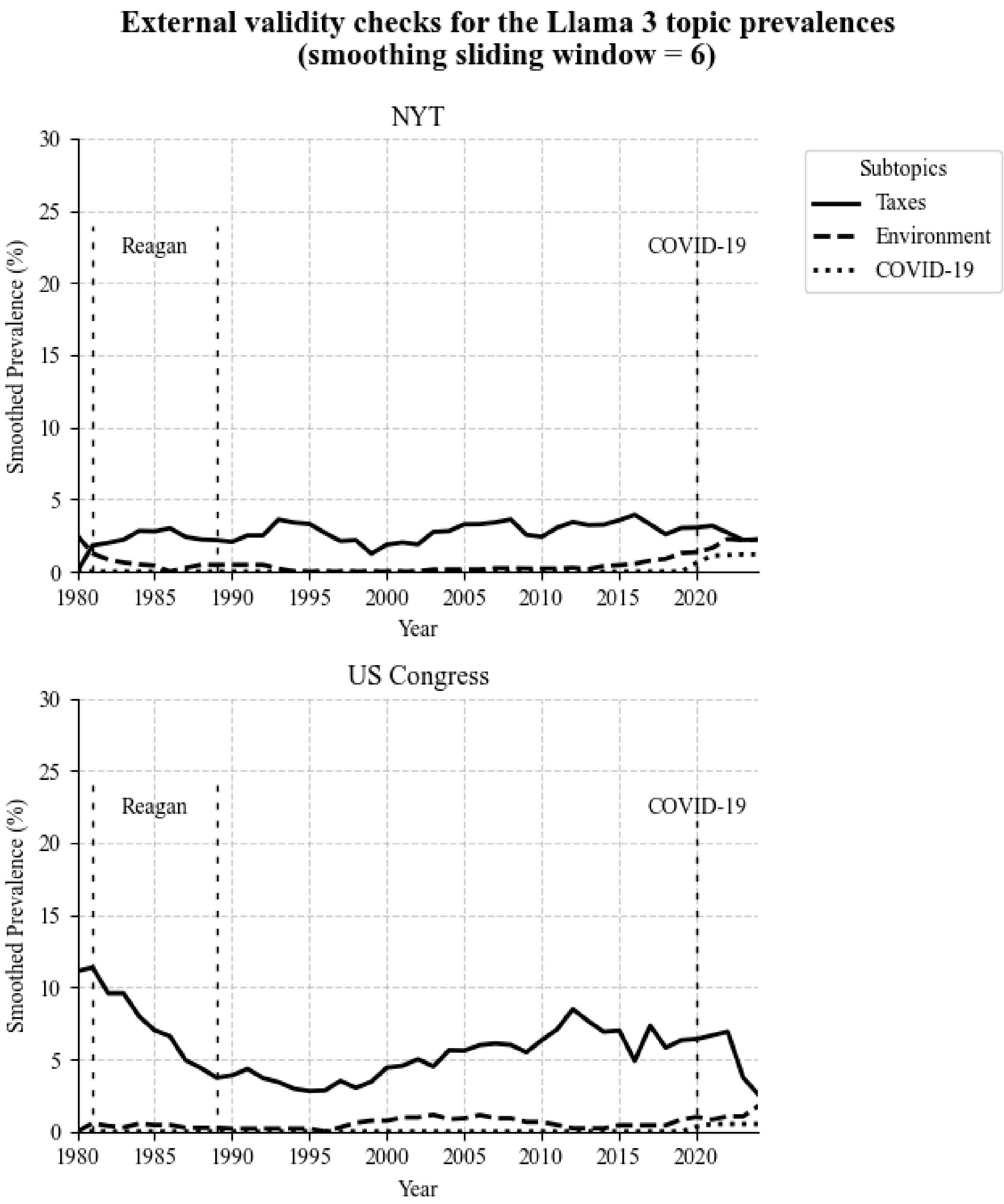

Figure B2: Prevalence of topics used as external validity measures along time for the Llama 3 model.

C. Additional results visualizations

936

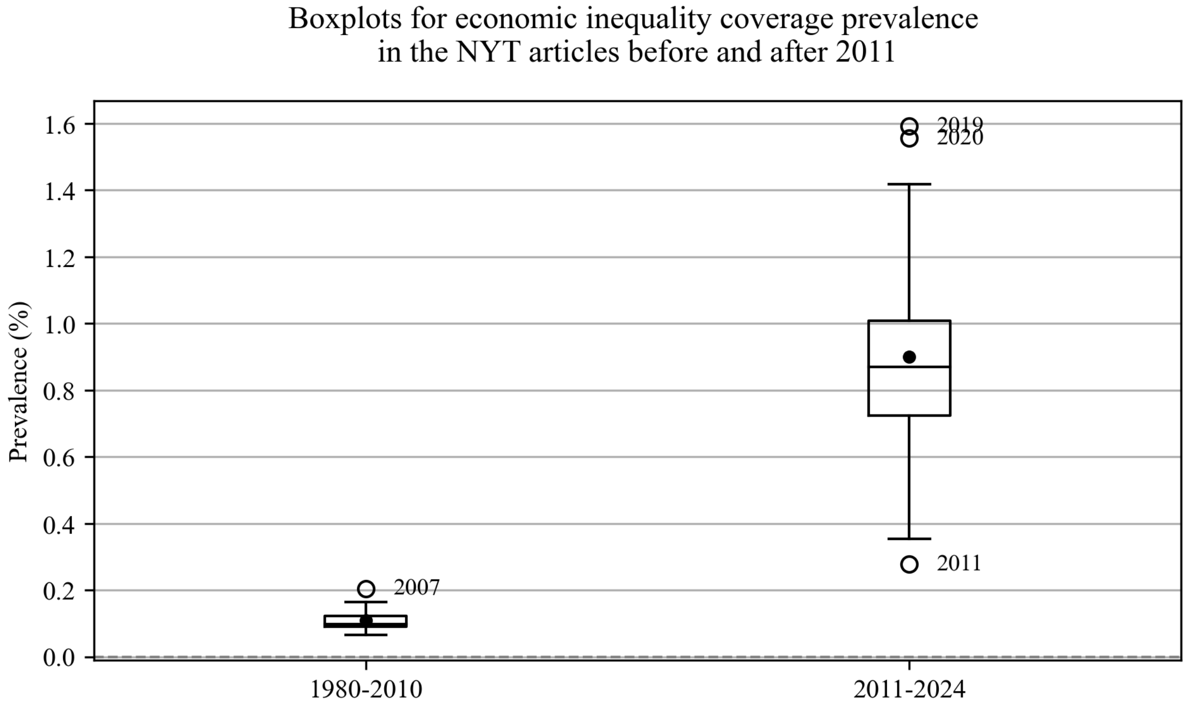

Figure C1: Boxplots for economic inequality coverage prevalence in the NYT articles before and after 2011 (mean difference: 0.79% (95% CI [0.56, 1.03]).

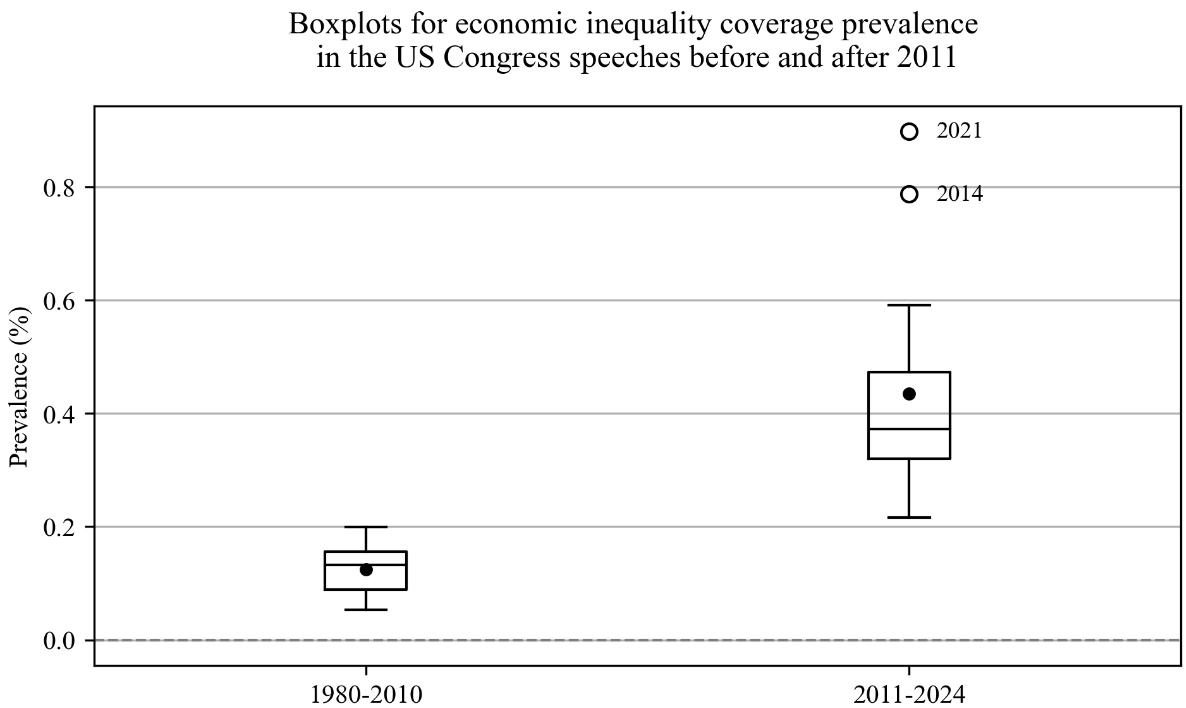

Figure C2: Boxplots for economic inequality coverage prevalence in the US Congress speeches before and after 2011 (mean difference: 0.31% (95% CI [0.19, 0.43]).

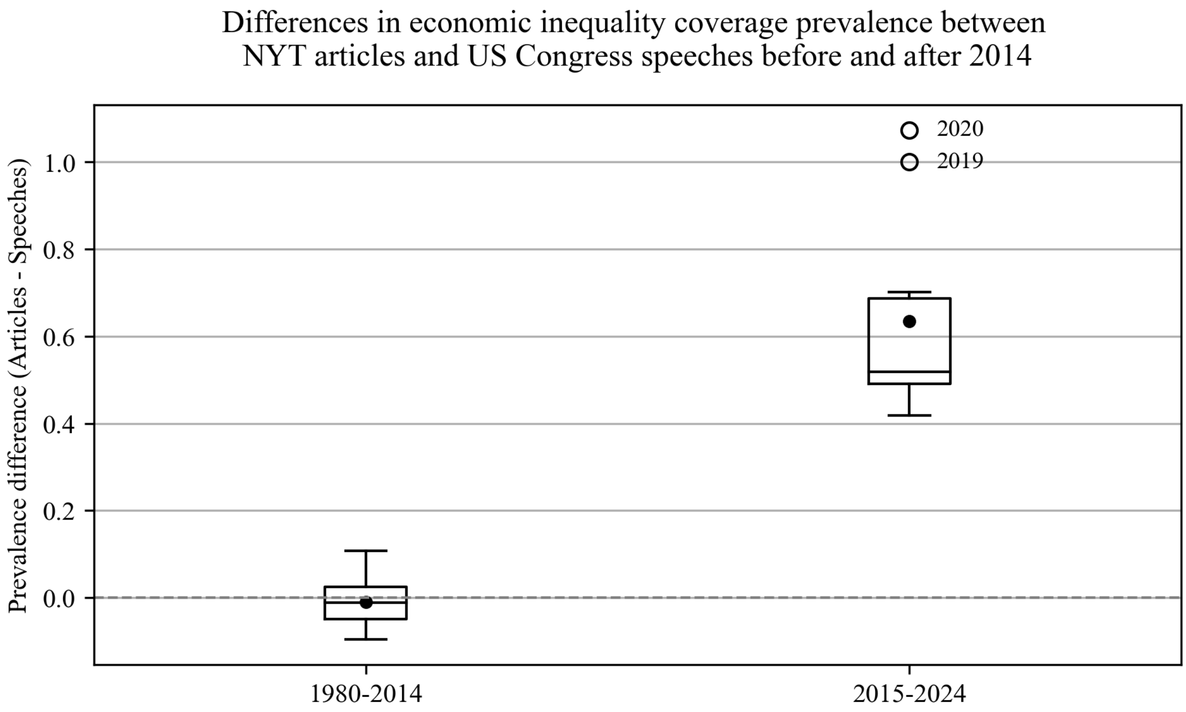

Figure C3: Boxplots for the prevalence difference between NYT articles and US Congress speeches from 1980-2014 and 2015-2024 (mean difference: 0.64% (95% CI [0.48, 0.81]).

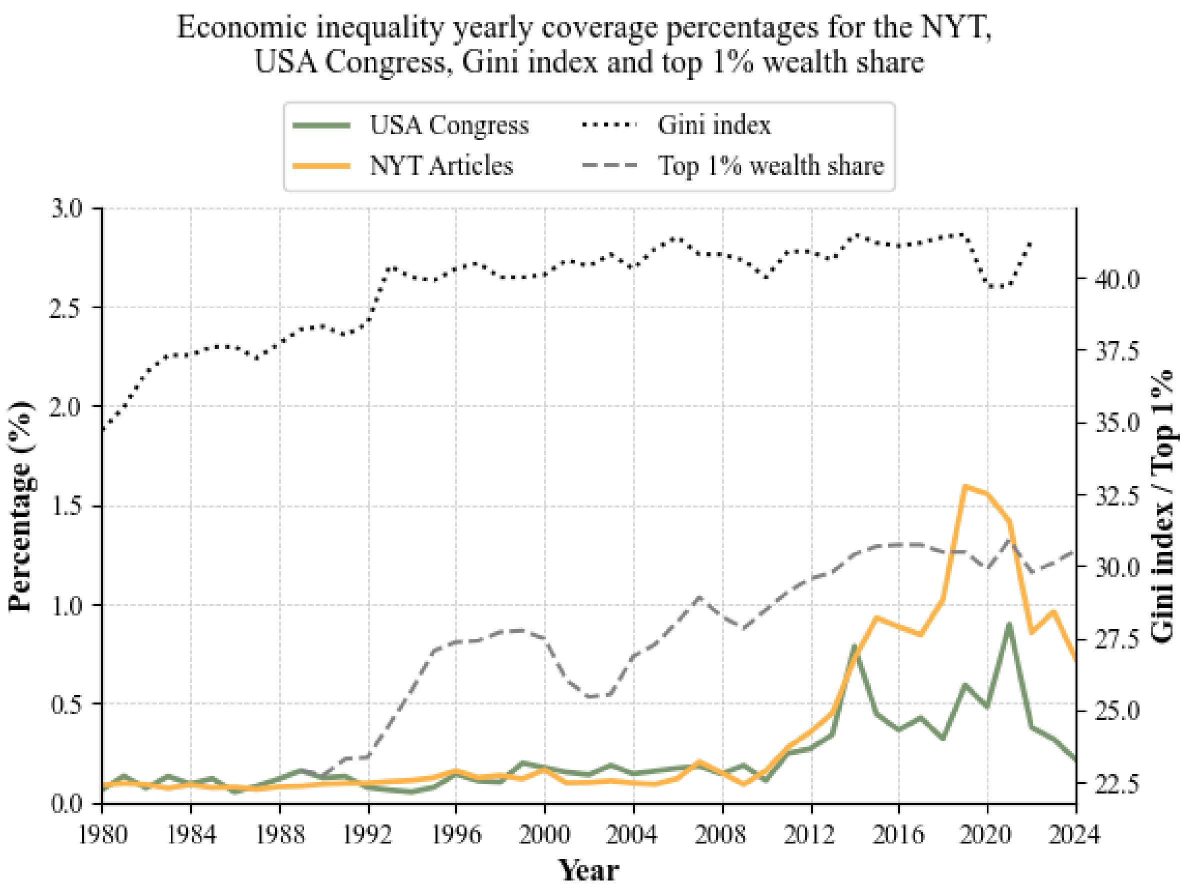

Figure C4: Yearly coverage percentages for the economic inequality subject for the US Congress speeches and the NYT articles and economic inequality objective measures

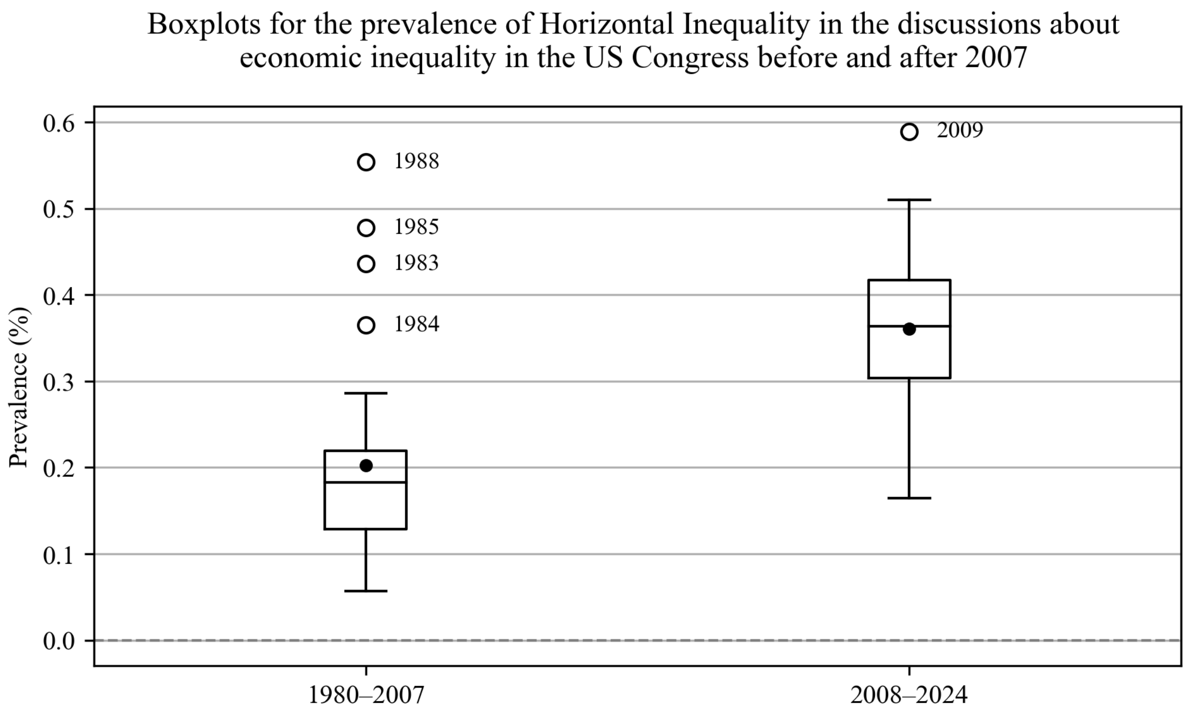

Figure C5: Boxplots for the prevalence of Horizontal Inequality in the discussions about economic inequality in the US Congress before and after 2007 (mean difference: 0.16% (95% CI [0.09, 0.23])

D. Health Prevalence Peaks in US Congress

The two peaks seen in the Health topic yearly prevalence graph (Figure 5.2) are mostly driven by bill discussions held in 1980 and 2002. One of the most present discussions in our data for this category in 1980 is that of the Rural Health Care Viability Act (S. 1438), which aimed to reduce inequalities in the Medicare insurance between urban and rural contexts. In the case of 2002 most speeches are related to the Medicare Modernization and Prescription Drug Act (H.R 4954), a bill that aimed at broadening drug coverage for seniors.

Acknowledgements

This thesis was supported by the incredibly generous guidance from Prof. David García and his research group. Special thanks to Segun, Peer, Alina, Apeksha, Jula, Adri and Giordano for their quick and clear insights into various methodological steps. Thanks also to Prof. Claudia Diehl for her perspectives on inequality research and for sharing ideas that helped shape this thesis. Their support and passion for social science research kept me curious and committed to maintaining rigour throughout the process.

Thanks to Universität Konstanz for offering an international Master's designed to open up the world of social media data, machine learning and LLMs to young social scientists. Professors like David García, Andreas Spitz and Ralf Brüggemann were key to introducing me to this field through their carefully developed courses. Thanks also to the Politics Department for its graduate courses on Causality and Research Design. And to the program coordinator, Kathrin, for her dedication and warmth from the first to the last week of the program.

The journey toward this thesis and Master's would have been unthinkable without the loved ones, colleagues and mentors who have left a lasting imprint on me.

Pursuing this program was possible due to the study opportunities provided by Universidad de Buenos Aires institutions, such as the FCE, FPsico and FCEN. Thanks to my fellow students, who shared their passion for the field and helped me see a research career as a possibility back then (1). Thanks to the research groups who welcomed a curious student and to the students who shared their enthusiasm for Social Psychology. Thanks also to the international psychology researchers who made their syllabi and classes publicly available. I am also grateful to the DAAD, the Ministerio de Educación de Argentina and the Prejudice Research Group at FPsico for supporting my participation in this Master's program.

Thanks to my family, who never gave up on my distracted ways of approaching school and life since I was a kid (2). To my Buenos Aires friends, who have always strived to stay close, especially to those who managed to visit and chat by the Bodensee (3). To Pedro, for being an invaluable support during my move abroad and in my daily challenges. To my KN friends, who made this Master's incredibly enriching and fun. To all the friends from Latin America who kept me company with their advice, jokes and hughs (4). And to my emerging Köln friends, who were there throughout the thesis semester, making it possible to envision a new start (5). [5]

[5] (1) cata, mati, orne, gabi, paz, ivan, shei, feed, joaco m.; (2) mum, dad, sister, abuela; (3) anita, martu, ari, cami, oli, martin, fran, ovi, loló, mini, flia. desteract, flia. ramos; (4) giovis, juli, guido, leo, santi, trini, david cañón; (5) clara, feli, tomi, husserl archiv, emi, freddy; (6) martu and sofi for reading this thesis.