Transition or Tradition? Analyzing 17 Years of Presidential Reports in Mexico Through Leadership Changes

Candidate ID: 27529

2024-04-17

word count: 2995

Contents

1. Abstract

This work analyzes 17 years of presidential reports and speeches from three Mexican administrations, focusing on the period when power shifted among three different parties after 70 years of single-party rule. Web scraping and quantitative text analysis techniques were used to collect and explore data from presidential reports and speeches. Structural Topic Modeling was then applied to detect key themes and assess how political agendas from each governmental change were reflected in these discursive narratives.

I conducted a correlation analysis between the identified topics and economic indicators focusing particularly on expenditure. Additionally, based on the prevalence of themes identified through the application of a thematic dictionary, I developed an indicator to mesure gender and care work themes within the documents and correlated as well it with economic indicators. The analysis first revealed a high degree of similarity across the topics discussed in the speeches, with overlapping themes that may indicate persistent political agendas. However, three topics—Education, Security, and Health and Wellbeing—stood out for their varying prominence depending on the ruling political parties. A strong correlation was also observed between the prominence of gender-related themes in the documents and expenditures on related issues.

2. Introduction

With this work I seek to answer the question: *What, if any, are the most prevalent topics in each presidential period, during the last 3 six-year terms?* The goal is to identify the most prevalent topics during each term and correlate them with related economic indicators such as female labor participation, unemployment rate, women's representation in parliament and budget allocations for health, economic development, social development, and gender.

The methodology for this purpose included:

1. Collecting speeches and reports via web scraping, (initially only speeches were considered but given the small amount of data after processing, it was decided to extend it).
2. Acquiring macroeconomic data from the World Bank and Ministry of Finance ensuring comparability and deflation to 2023 prices when necessary.
3. Text processing, including a detailed cleaning, corpus creation, tokenization, and removal of recurring irrelevant words. The document-feature matrix ended up with 17 documents and 3,262 features.
4. Exploratory Data Analysis (EDA): Quantitative text analysis techniques were employed, including word clouds, tf-idf metrics, dictionary applications, lexical diversity, and sentiment analysis.
5. Topic Identification: Structural Topic Modeling revealed recurring themes indicative of persistence political agendas. Variations in emphasis on health and well-being, education, and security were noted across different political parties. Trends in gender and care-related discussions over time were also analyzed.
6. Correlation analysis between theta values for each topic in each document and diferent macroeconomic variables.
7. Use of visualization tools such as sankey diagram and ToolDAvis for interactive visualization hosted on https://michellepapadakis.github.io/PSFMY459/

The top topics founded were "Health and wellbeing", which showed a positive and strong correlation with gender-focused spending and women's political representation and a surprising weak correlation with health spending despite the concurrent COVID pandemic. The "Security" topic correlated negatively with unemployment rates and the "Education and Wellbeing" topic showed a positive correlation with the unemployment rate and the prevalence of the care issue in the text, and a negative relationship with spending on social development but a positive and high correlation with budget participation in economic development. The prevalence of topic aligns with specific policy focuses which may reflect both long-term national objectives and responses to period-specific challenges, demonstrating the intersection of policy, politics, and historical context in presidential narratives.

3. Motivation

Mexico's political landscape has undergone significant shifts, with the three major parties exchanging power over the last 18 years. In 2000, presidential candidate Vicente Fox ended 70 years of uninterrupted one-party rule in Mexico by the Institutional Revolutionary Party (PRI). Currently, the party's principles include nationalism, freedoms, democracy, and social justice, and according to its statutes, it promotes the exercise of power towards the economic, political, social, and cultural development of Mexico, maintaining an ideological tendency that links it to the social democratic current. (PRI, 1978). After a series of conflicts, corruption, and electoral fraud, there was a political turnover in 2000 to the conservative National Action Party (PAN), which aligns with Christian democracy and a humanist doctrine according to its official documents (PAN, 1939). PAN ruled from 2000 to 2012, with Felipe Calderón Hinojosa elected in 2006, giving continuity to the political project of the PAN (Lawson, Chappell. 2007). However, Calderón's term was marked by a war against drug trafficking that led to extreme violence, resulting in the PRI's return to power in 2012 with former President Enrique Peña Nieto. Due to high crime rates, violence, unsolved massacres, clandestine graves, assassinations of journalists, and general insecurity, coupled with underperforming structural reforms, notably in education and energy, there was significant social discontent. This paved the way for the National Regeneration Movement (MORENA) to win in 2018, with the party's founder, Andrés Manuel López Obrador, leading the "Together We Will Make History" coalition. His administration promised better living standards and the end of political privileges to improve the population wellbeing.(Del Castillo, 2018). In Mexico, the three main

sources of discontent that led to López Obrador's victory, according to Greene, K. and Sánchez-Talanquer, M. (2018), were firstly, the transition to democracy and the free market did not meet expectations for prosperity. Secondly, drug-related violence has dramatically increased since 2006, with 2017 being the deadliest year on record. Finally, structural reforms, notably the educational reform in 2013, were quickly implemented but did not produce the expected results.

This work analyzes the period from 2007 to 2023, covering three presidential terms with three different parties in power, aiming to offer a longitudinal perspective to discern policy focus trends and shifts, if any. Background research can be found in Torres et al. (2020), which includes a comparative analysis from 1983 to 2013 of presidential reports (one per year), an analysis of President Felipe Calderón's speeches to legitimize the so-called "War on Drugs" (Vazquez Moyers, 2014), as well as cases of statistical text analysis for various Latin American countries like Uruguay (Vernazza Mañana & Vicente Villardón, 2021) and Peru (Chung Pinzás & Inche Mitma, 2024).

4. Description of the corpus

The corpus is composed of 17 documents, corresponding to 17 years of presidential reports, each document containing the presidential report and the speech issued by the president on the occasion of the delivery of the report to the congress. Derived from an initial analysis which revealed insufficient information for analysis, it was decided to also include the content of the presidential report within the same variable. The data is from 2007 to 2023, since the 2024 presidential report will be made on September 1 of this year. Therefore, 6 documents for President Felipe Calderón are integrated, corresponding to the speeches from 2007 to 2012, 6 for President Enrique Peña Nieto, corresponding to 2013 to 2018 reports and 5 documents for President López Obrador, corresponding to 2019 to 2023.

Text data was retrieved mainly through the Center for the Study of Democracy and Elections of the Metropolitan Autonomous University of Mexico (CEDE,2024), the mexican government website and one from the national newspaper "El Universal". Finally, due to a change in the domain of the official website of the government of the republic starting in 2012, the presidential reports of the government of Felipe Calderón were recovered from his personal page. The texts were retrieved using scrapping techniques as well as processes for the automated downloading and extraction of data from pdf documents.

Regarding economic data, the female labor force participation rate (percentage of female population aged 15-64), the unemployment rate (percentage of total labor force), and the percentage of women in parliamentary seats (percentage of total seats) were obtained from the World Bank, using their wbstats package. For the expenditure analysis, data were extracted from the Mexican Ministry of Finance. This data, classified according to the functional classification of expenditure (SHCP, 2008), included "economic development" expenditures which cover spending on commercial affairs, and sectors such as agriculture, science and innovation, tourism, transportation, and mining. Expenditures for "social development" —encompassing education, health, housing, and social and environmental protection— were also retrieved, along with "gender-related" expenditures, a budgetary item initiated in 2006 to encourage gender equality, as well as health sector spending. All budget data were presented as a percentage of the total annual expenditure, adjusted for inflation where necessary for comparisson.

The text processing involved several steps such as: 1. Cleaning up the raw text, including the removal of numbers, links and symbols, lowercase it, elimination of common words and proper nouns. 2. Creation of the corpus and tokenization of it, to which the elimination of stopwords, numbers and punctuation was extended. This step was reinforced through a manual removal of a list of Spanish pronouns, as well as a number of words irrelevant to the context that showed high frequency. 3. The data feature matrix was created, where again, taking care of the order of the rules, a general cleaning of the terms was done. The matrix came out at the end with 17 documents, and 3,262 features.

Before undertaking the Structural Topic Modeling analysis detailed in the following section, I conducted an exploratory data analysis. This initial analysis included assessing the total word count per document (Figure 1), identifying the most significant words in the text using Term Frequency-Inverse Document Frequency (Figure 2), and examining the variation in word usage across different presidencies (Figure 3). Additionally,

the lexical diversity within each discourse was calculated (Figure 4), and sentiment analysis was performed on the text from each year (Figure 5).

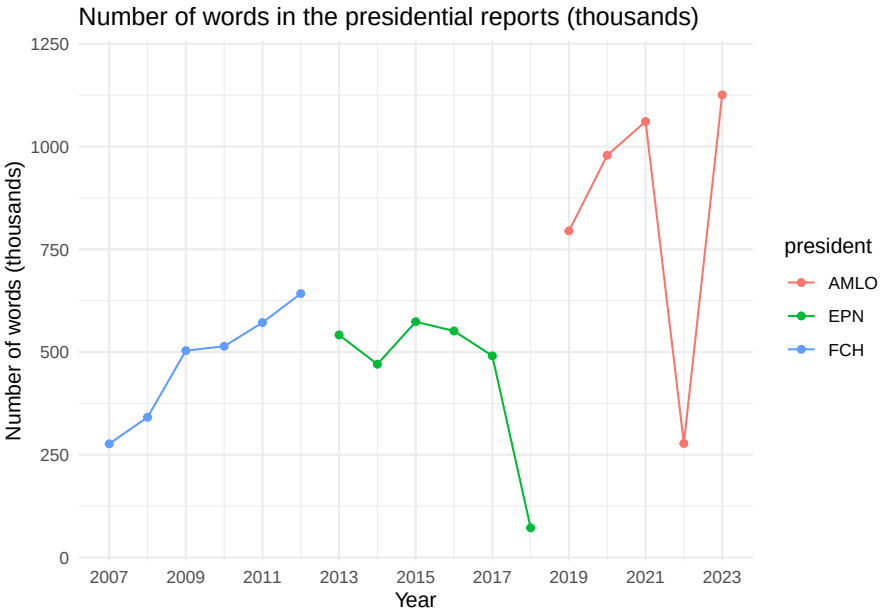


Figure 1: Total number of words per document

Figure 1 shows an increasing tren in the number of words in the texts, with a notable rise in López Obrador's 2023 and a couple of declines in Peña Nieto's 2007 and López Obrador's 2022.

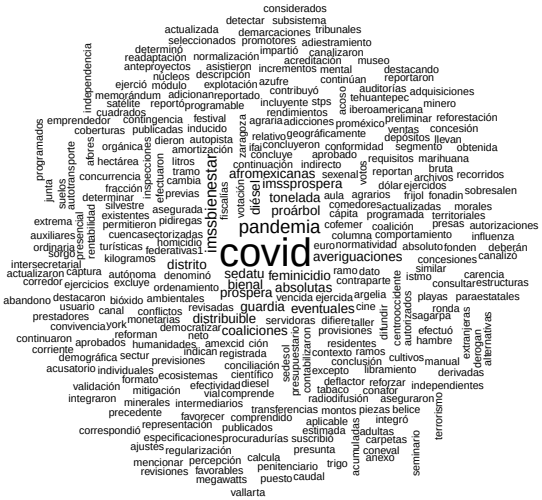


Figure 2: Most relevant words weighted by tf-idf

Figure 2 highlights key terms from presidential speeches, with terms like "COVID," "pandemic," "well-being programs" (implied by "IMSSBIENESTAR"), and legal and safety-related terms such as "prosecution," "homicides," and "investigations" featuring prominently. The prevalence of these terms, particularly those related to health and safety, likely reflects the socio-political challenges and priorities of the times they were spoken.

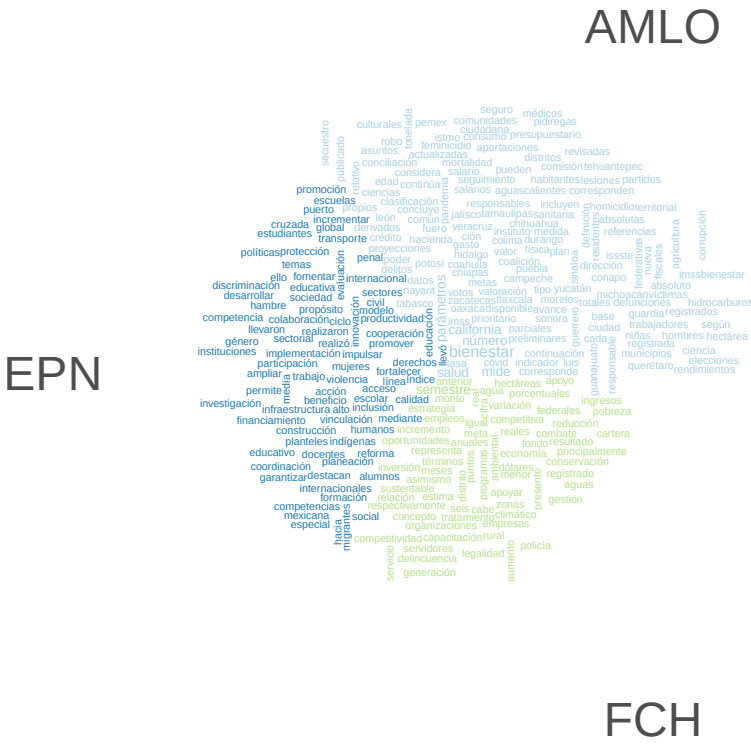


Figure 3: Most relevant words weighted by tf-idf by president

When analyzing by president in figure 3, there's a distinct trend of recurring themes within the speeches of each leader. For instance, López Obrador's speeches are characterized by a high frequency of terms related to health, welfare programs, and extensive mention of country states. In Peña Nieto's case, there is a notable presence of terms linked to education and economic productivity. Lastly, for Calderón, there's a significant recurrence of terms tied to the economy as well as policing and crime but also the used of words related to vulnerable groups such as indigenous and women.

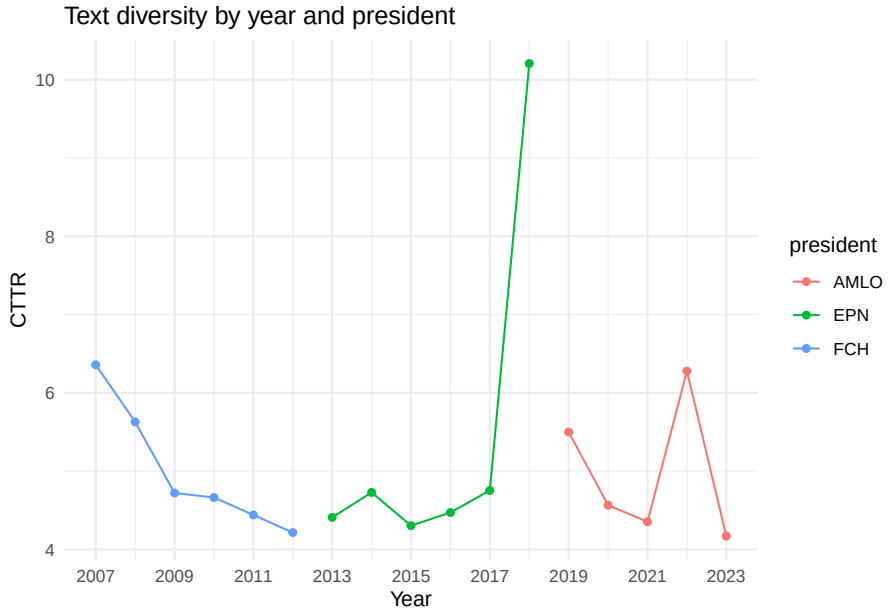


Figure 4: Lexical diversity by year

The lexical diversity plot utilizes the Corrected Type-Token Ratio to adjusts for text length differences,

which is particularly valuable here as Figure 1 revealed significant variations in the length of reports and speeches. The graph exhibits exceptionally high lexical diversity in Peña Nieto´s 2018 report, contrasting with the lower diversity observed in López Obrador's 2023 speech, similar in indicator to Felipe Calderón's 2012 report.

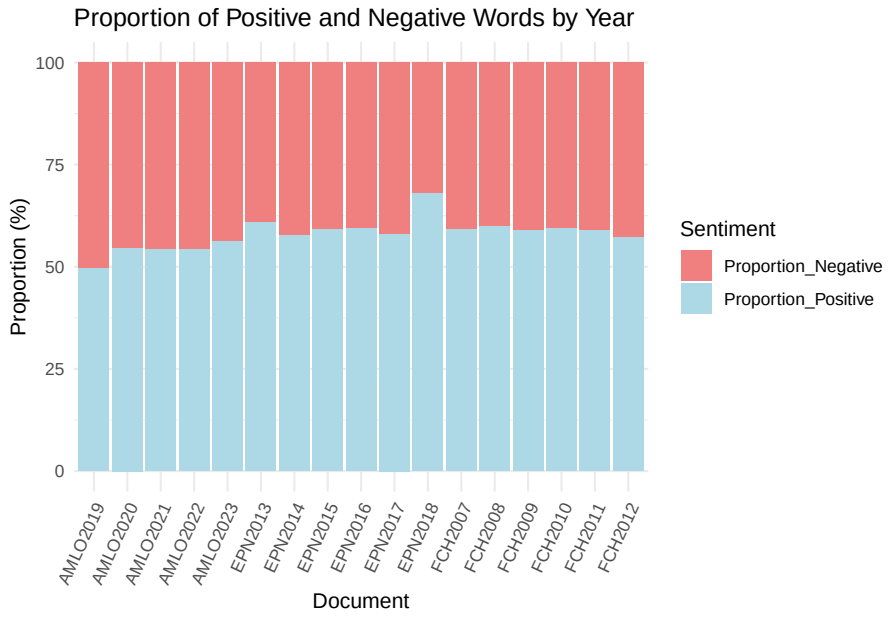


Figure 5: Sentiment analysis: proportion of positive and negative terms

The sentiment analysis of presidential speeches, as seen in figure 5, revealed a tendency towards positive language usage. Interestingly, Peña Nieto's speech in 2018 had the most positive language, which contrasted sharply with next year López Obrador's 2019 speech that recorded the least positive terms.

5. Description of the methods

To refine the textual data analysis, I used an unsupervised machine learning algorithm well-suited for identifying structural topics in texts. I applied the Structural Topic Model (STM), because, among other advantages, it allows for the incorporation of metadata covariates, which can help to understand how topics relate to different variables, which is the main objective of this work. It also helps to identify the most relevant topics in the corpus, which can be particularly useful when dealing with large amounts of text data or data that overlaps in content, like presidential reports (See Mostafa, M.M., 2022 and Roberts et al., 2013, 2014). Given that choosing the right K is crucial—as too low a K may lead to overly broad categories, while too high a K can result in many insignificant topics—the ideal K was inlfuenced by using the *searchK* method, as shown in Figure 6, and further refined by the *selectModel* method. Here, I took the optimal K result from obtaining the heldout values calculated for each of the estimated STM models with different values of K. The optimal K is the number associated with the maximum value in the list of heldout values. Initially, the optimal value found was 10 K; however, after further manual testing and adjustments, K was adjusted to 3 for better topic coherence.

Further steps included estimating how these topics related to 'party' and 'year' covariates, enabling comparisons across political parties and years. To address two themes that were not initially captured in the topic analysis, I created dictionaries with key terms for each theme. These dictionaries helped track relevant word occurrences in the texts.

I exracted theta values for each resulting topic across documents to describe the thematic composition of the documents and correlate them with various economic indicators, summarized in a correlation matrix at the end.

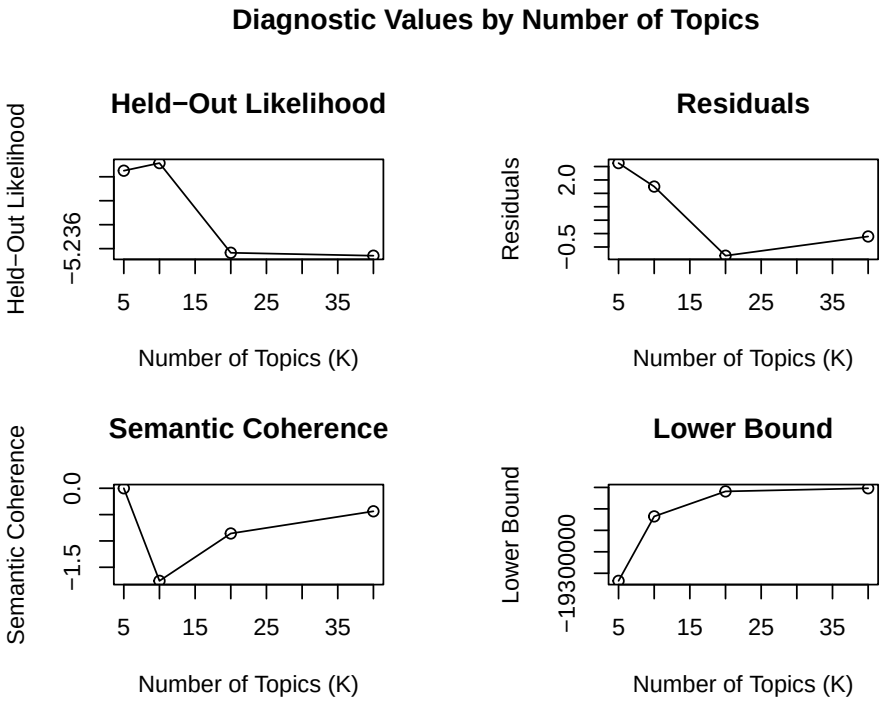


Figure 6. Finding the optimal range of latent topics through model-fit statistics

6. Results.

6.1 Structural Topic Model

When conducting the aforementioned STM analysis with 3 K, I founded, as seen in figure 7, that Topic 2, encompassing education, is the most predominant as it represents 40% of the reports, closely followed by Topic 3, which focuses on security. Topic 1, centered on health and well-being, has the smallest proportion. These results suggest that, on average, topics of education and security and Health and wellbeing are the most discussed in the texts analyzed.

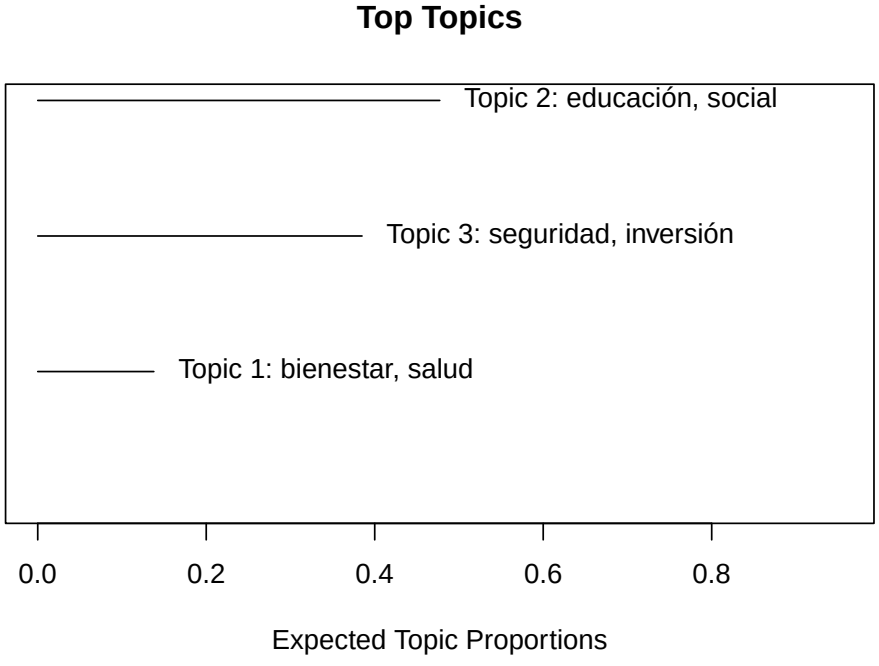


Figure 7: Top Topics and their proportion in the documents

6.2 Topic prevalence over time

Once the effects of covariates on topic prevalence were estimated, the prevalence of each topic over time was determined. For Topic 1, health and wellbeing, an increase in prevalence over time is observed (see Figure 8), particularly notable in 2018 with a sharp decline in 2022 and a resurgence in 2023.

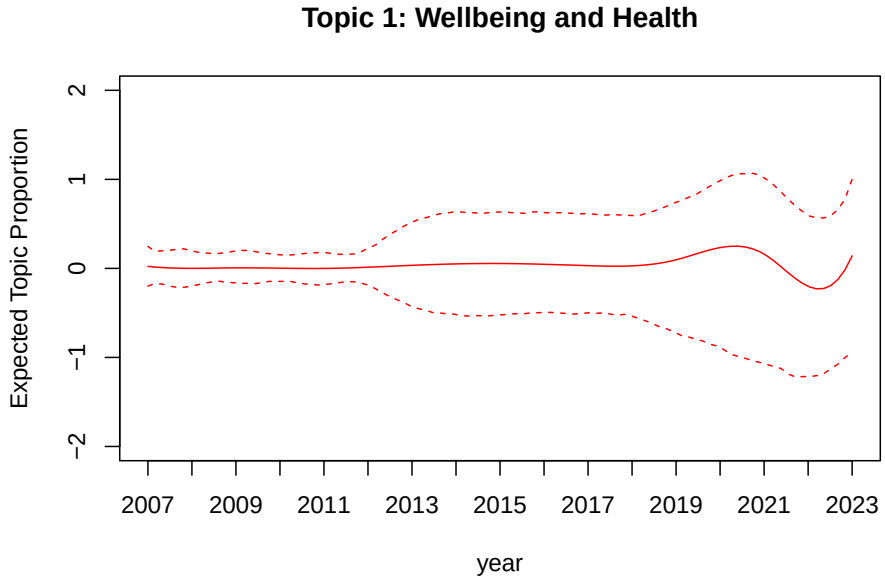


Figure 8: Topic 1: Health and well-being over time

On the other hand, Figure 9 shows that education saw an increase from 2012 to around 2018, followed by a drop in 2020 and a recover in 2022. Finally, Topic 3, reached its peak from 2007 to 2013 and then stabilized. This suggests that the prevalence of the topics has varied over time with marked emphases in the years coinciding with government changes.

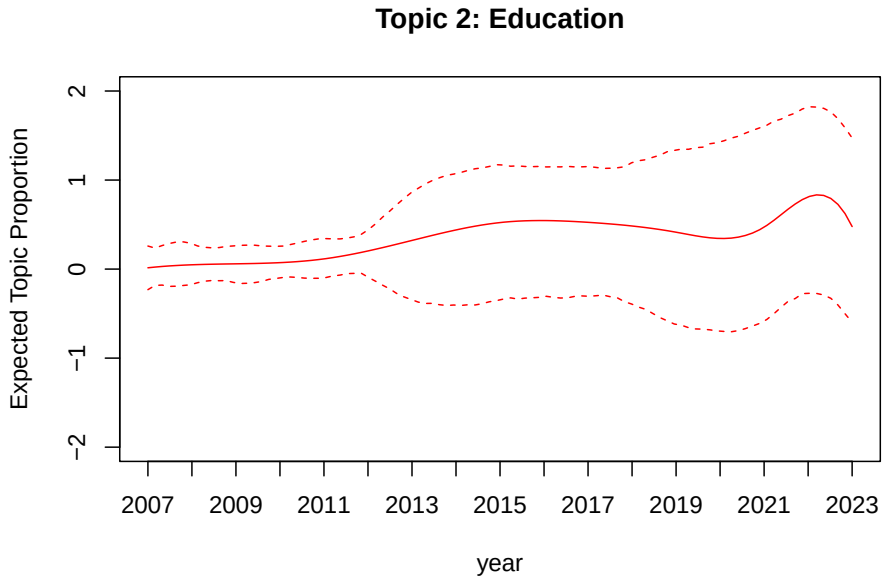


Figure 9: Topic 2: Education over time

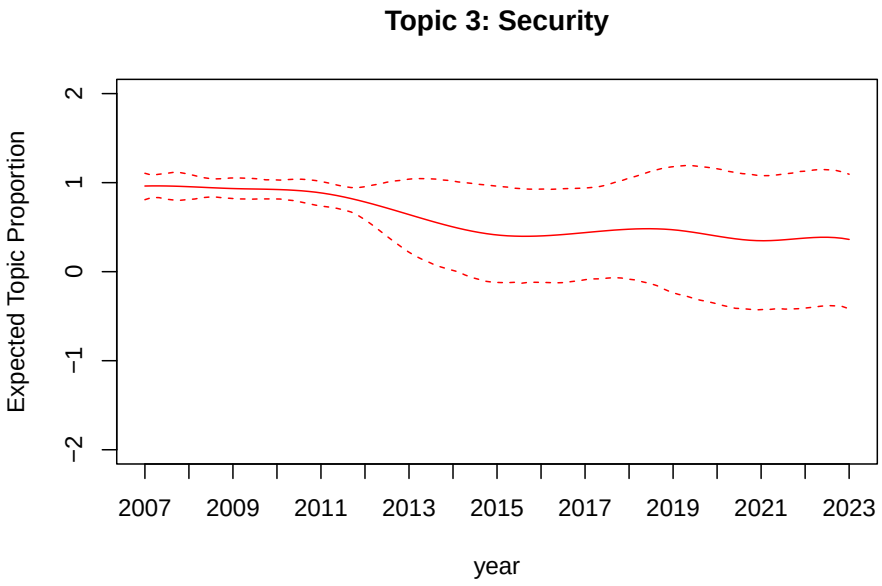


Figure 10: Topic 3: Security over time

6.3 Topic prevalence by political party

The analysis explored how the ruling party influences topic prevalence, with results depicted in Figures 11 to 13. Figure 11 indicates that the 'Health and Wellbeing' topic is most associated with MORENA, less so with PRI and PAN, where the association is not statistically significant. Figure 12 suggests all parties positively align with the Education theme, strongest with PRI. Figure 13 shows Security is more associated with PAN, then PRI and MORENA. Additional graphs in the annex compare topic prevalence across parties.

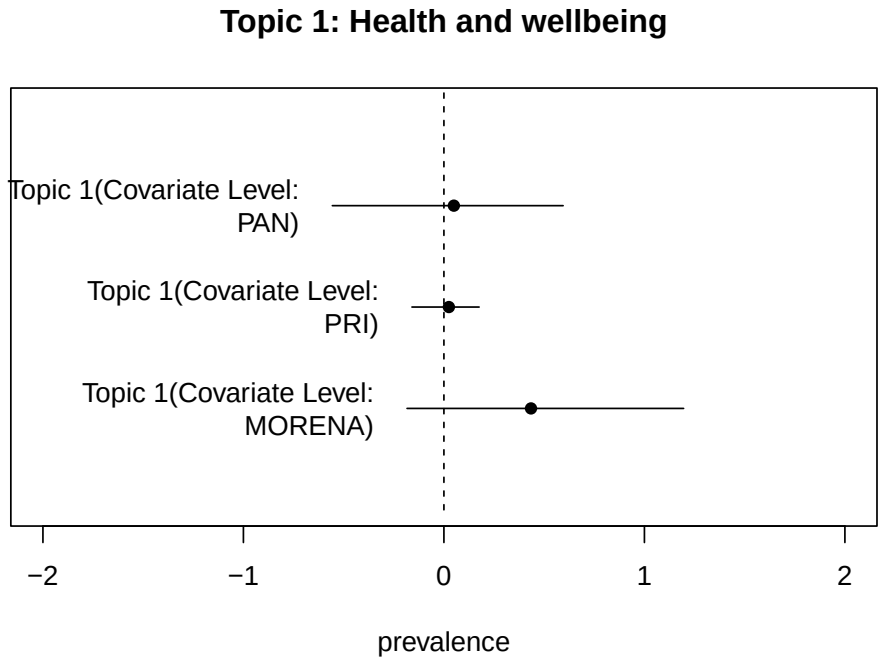


Figure 11: Topic 1: Health and well-being by party

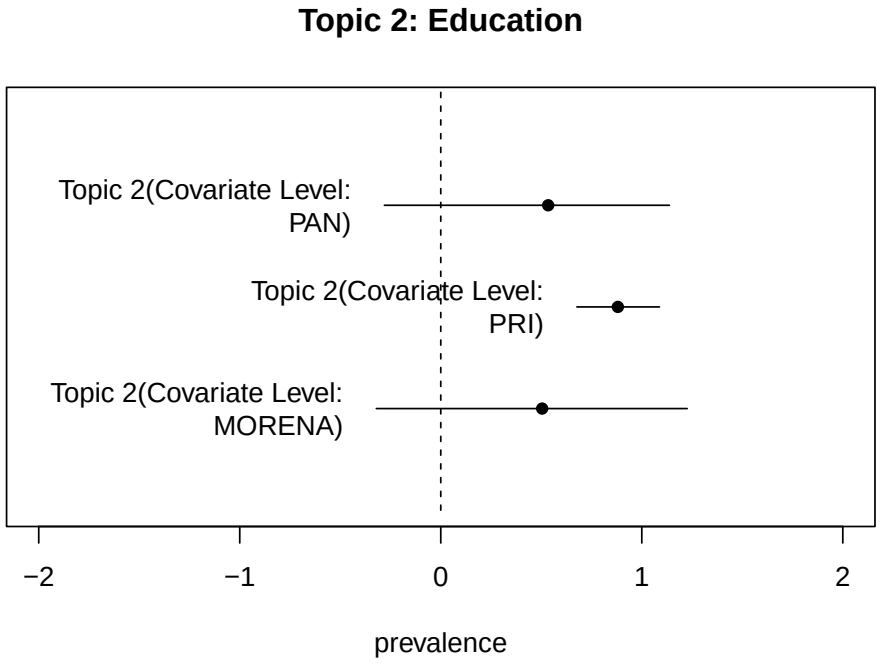


Figure 12: Topic 2: Education by party

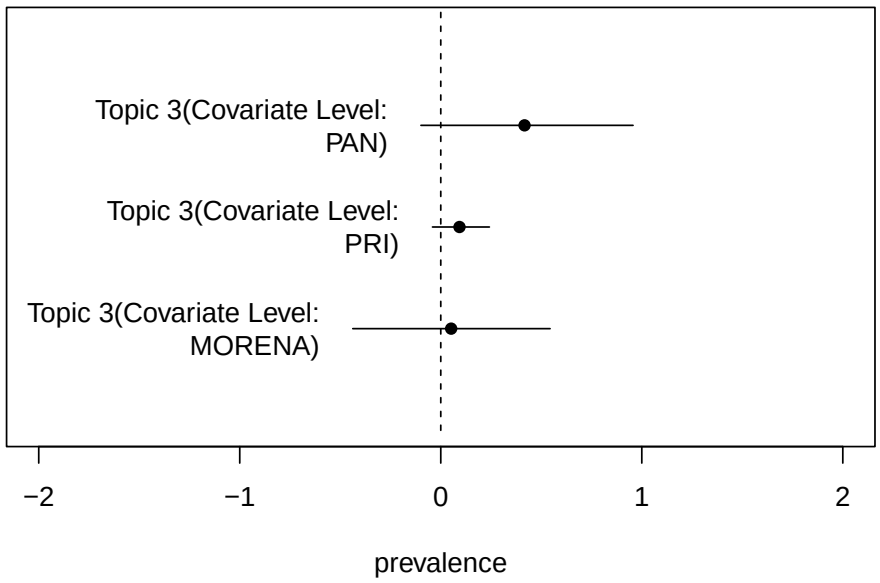


Figure 13: Topic 3: Security by party

Given Mexico's significant gender inequality (IMF, 2024), further analysis examined gender and care-work themes. I created dictionaries to track these terms in presidential speeches, analyzing their occurrence in a term matrix weighted by frequency. Figures 14 and 15 chart the annual variation of these themes in the political arena.

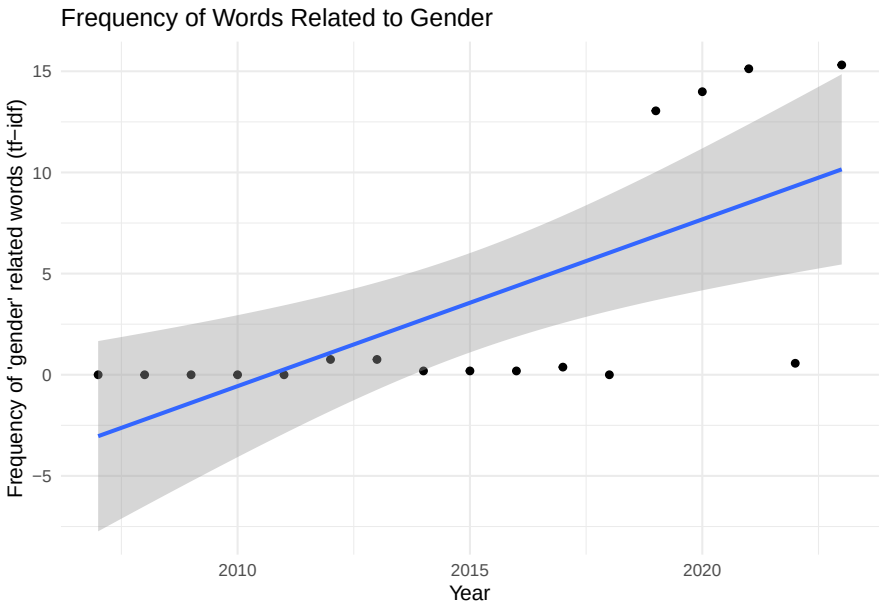


Figure 14: Gender mentions over time

Figure 14 indicates an increasing focus on gender-related terms over time. With the frequency rising, particularly from around 2010 onwards, it suggests heightened awareness on gender issues within the political discourse. In contrast, the frequency of care-related terms has reduced over time, as shown in Figure 15. Suggesting that the topic of care has not been a central focus in the political discourse, with the frequency of care-related terms decreasing over time.

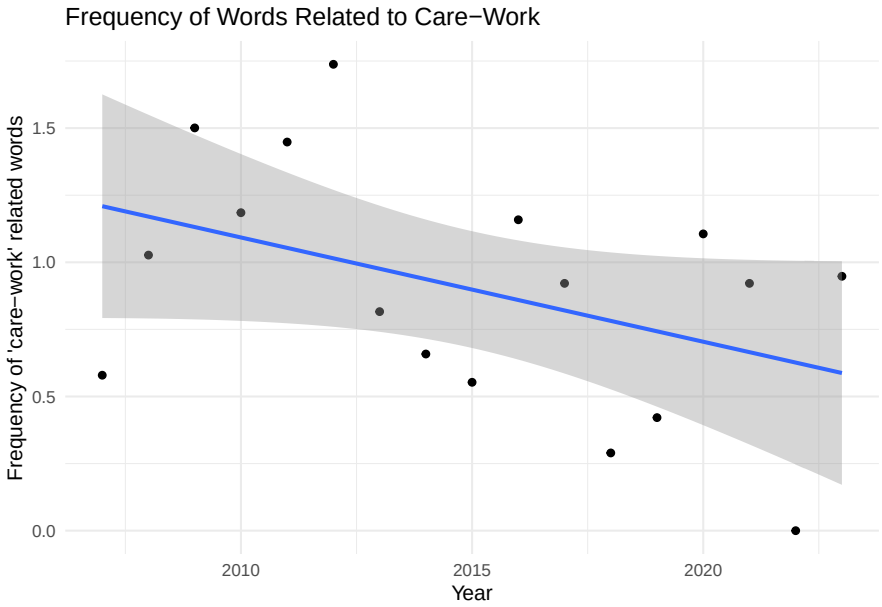


Figure 15: Care mentions over time

6.4 Correlation analysis

Finally, a correlation analysis was conducted on the resulting topics with economic and social indicators. The results are presented in Figure 16. There, a high and positive correlation is observed between Topic 1 and federal spending on gender actions, the proportion of women in parliament, and the budget for social development. Likewise, there is a high and positive correlation between Topic 1 and the prevalence of

care-related themes. However, there is a weak correlation between Topic 1 and health spending, which is surprising given that it is not only a prevalent theme in presidential reports but also coincides with the COVID pandemic. Regarding Topic 2, there is a high negative correlation with the unemployment rate and the prevalence of care and a positive correlation with women in parliament. Finally, Topic 3 has a positive correlation with the unemployment rate and the prevalence of the care theme in the speeches, and a negative relationship with social development spending but a positive and high correlation with the economic development budget share.

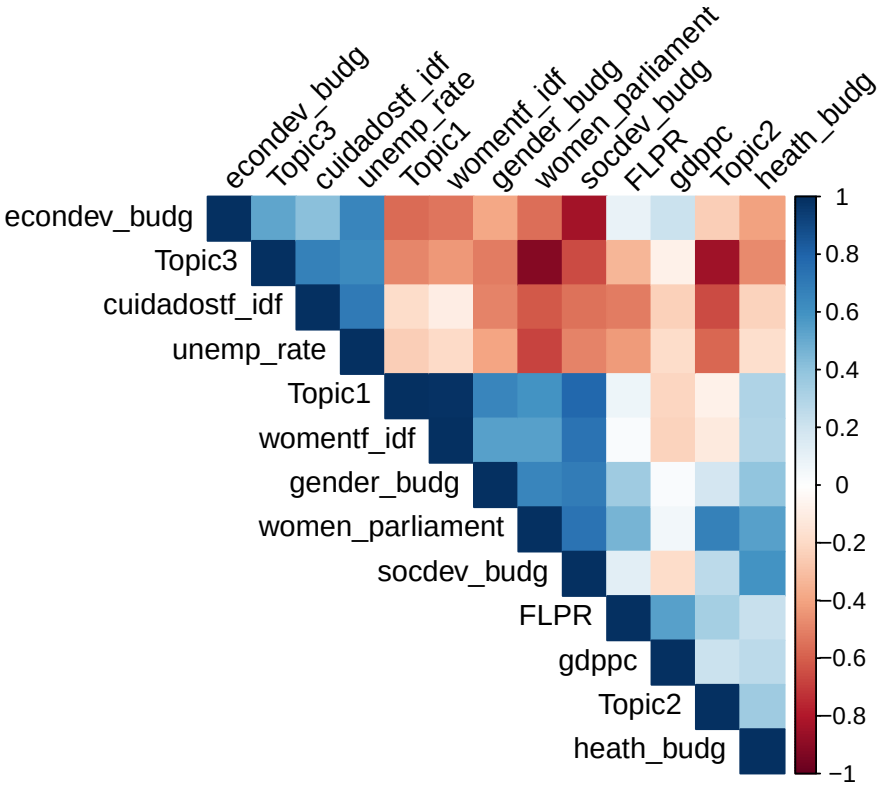


Figure 16: Correlation matrix of topics and economic and social indicators

Final visualization techniques were utilized to depict the topic distribution in presidential speeches. A Sankey diagram and an interactive topic visualization tool, hoosted here: https://michellepapadakis.github.io/PSFMY459/. In Figure 17, the diagram clearly shows a dominance of health and well-being themes in the speeches of the president from the MORENA party, which have increased over the years. This strongly aligns with the party's political discourse on welfare and a dignified life. Conversely, the PRI party has consistently emphasized education in its speeches, aligning with its national project and educational reform. Meanwhile, the PAN party has demonstrated a strong focus on security themes.

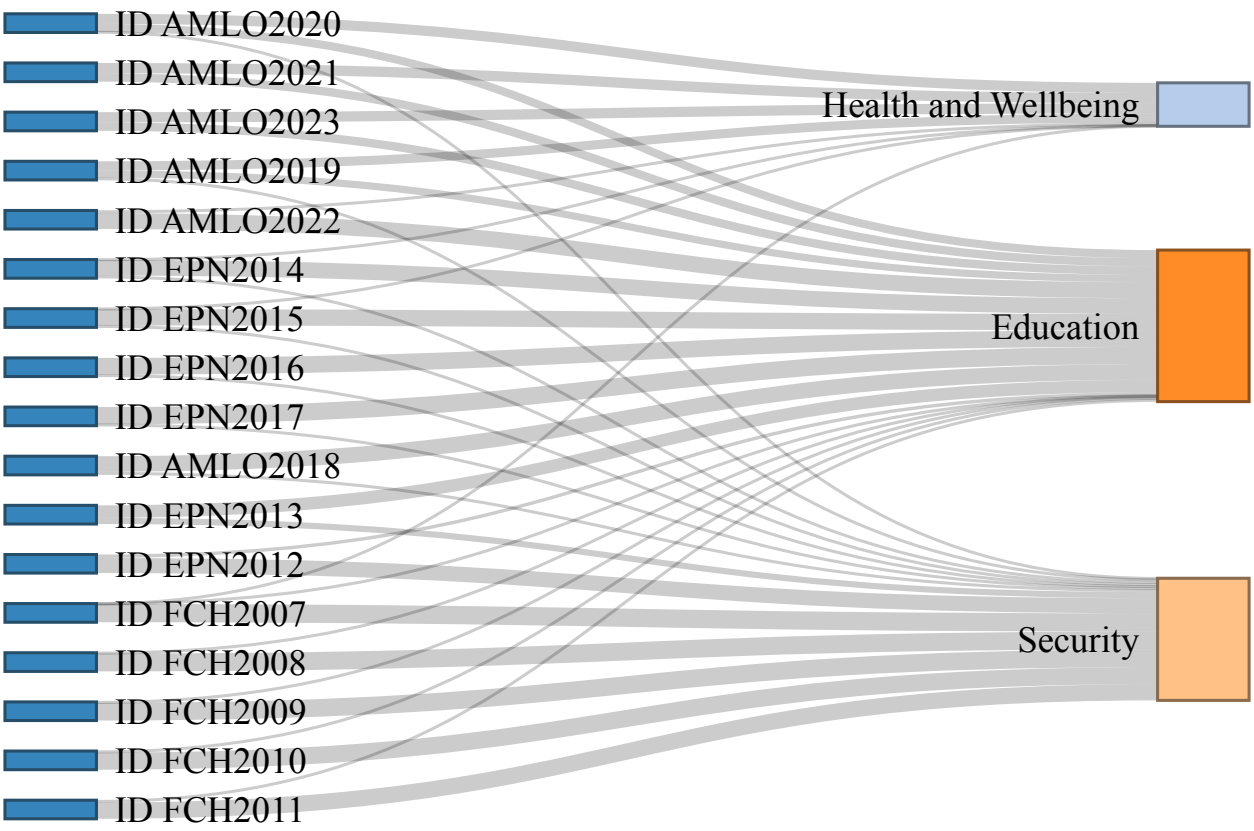


Figure 17: Sankey diagram of topic distribution by party

7. Conclusion

Education, Security, and Health and Wellbeing topics consistently appear, influenced by the party in power. Despite the overlapping political agendas, the analysis successfully identifies specific structural themes predominant in each party. This pattern suggests that while political leadership changes, certain themes persist, reflecting a complex interplay of tradition and transition in the political narrative. The difficulty in distinguishing more top topics from stm likely stems from these overlapping agendas; however, this works still manages to highlight the predominant themes for each party. These findings may indicate that shifts in government don't always lead to drastic changes in thematic focus but rather reflect a gradual evolution in addressing both longstanding and emergent societal needs.

8. References

- Calderón, Felipe. "Informe de Gobierno 2007-2012." Accessed April 2024. https://felipe.mx/.
- Centro de Estudios de la Democracia y Elecciones (CEDE). "Discursos presidenciales" [Presidential speeches]. Metropolitan Autonomous University of Mexico (UAM). Accessed April 2024. https://cede.izt.uam.mx/?page_id=2999
- Chung Pinzás, A. R., and Inche Mitma, J. L. "Minería de datos aplicada a los discursos presidenciales de Pedro Castillo Terrones en Perú" [Data mining applied to the presidential speeches of Pedro Castillo Terrones in Peru]. Revista De Comunicación 23, no. 1 (2024): 141–156. https://doi.org/10.26441/RC23.1-2024-3417.
- Del Castillo, Graciana. "El Plan Nacional de AMLO para México es Motivo de Optimismo" [AMLO's National Plan for Mexico is a cause for optimism]. LSE Latin America and Caribbean Blog. July 06, 2018 https://blogs.lse.ac.uk/latamcaribbean/2018/07/06/el-plan-nacional-de-amlo-para-mexico-es-motivo-de-optimismo/.
- El Universal. "Texto íntegro del mensaje de Felipe Calderón en su sexto informe de gobierno" [Full text of Felipe Calderón's message in his sixth government report], El Universal, September 2012. https://archivo.eluniversal.com.mx/notas/868097.html
- Greene, Kenneth, and Mariano Sánchez-Talanquer. "Latin America's Shifting Politics: Mexico's Party System Under Stress." Journal of Democracy 29, no. 4 (October 2018): 31-42. https://www.journalofdemocracy.org/articles/latin-americas-shifting-politics-mexicos-party-system-under-stress/.
- International Monetary Fund. Western Hemisphere Dept. "Mexico: Selected Issues." IMF Staff Country Reports, 2023, A001. Accessed April 2024. https://www.elibrary.imf.org/view/journals/002/2023/357/002.2023.issue-357-en.xml .
- Lawson, Chappell. "How Did We Get Here? Mexican Democracy After the 2006 Elections." PS: Political Science & Politics 40, no. 1 (2007):45-48 https://www.jstor.org/stable/20451890 .
- Moyers, A. V. "La guerra contra el narcotráfico en el sexenio de Felipe Calderón: Análisis del discurso" [The war against drug trafficking during the six-year term of Felipe Calderón: Analysis of the speech], (2014) https://ri-ng.uaq.mx/handle/123456789/745.
- Mostafa, M.M. "A one-hundred-year structural topic modeling analysis of the knowledge structure of international management research." Qual Quant 57 (2023): 3905–3935. https://doi.org/10.1007/s11135-022-01548-w.
- Partido Acción Nacional. "Principios de Doctrina.", [Principles of Doctrine], (1939). https://www.pan.org.mx/documentos/principios-de-doctrina.
- Partido Revolucionario Institucional. "Declaración de Principios", [Declaration of Principles], (1978) https://pri.org.mx/ElPartidoDeMexico/NuestroPartido/Documentos.aspx.
- Presidencia de la República. "Mensajes con Motivo del Informe de Gobierno 2012-2018", [Messages on the occasion of the Government Report, 2012-2018], Gobierno de México. Accessed April 2024. https://www.gob.mx/epn.
- Presidencia de la República. "Versión Estenográfica del Mensaje con Motivo del Informe Gobierno, 2019-2023" [Stenographic Version of the Message on the occasion of the Government Report, 2019-2023], Gobierno de México. Accessed April 2024. https://www.gob.mx/presidencia.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. "Structural Topic Models for Open-Ended Survey Responses." American Journal of Political Science 58, no. 4 (2014): 1064-1082. https://doi.org/10.1111/ajps.12103.
- Torres, I. B., Urrutia, J. A. V., María de Jesús, O., and Armas, H. J. H. "Minería de datos aplicada al análisis de los discursos presidenciales de México.", [Data mining applied to the analysis of Mexican

presidential speeches], Tla-melaua: revista de ciencias sociales, no. 48 (2020): 89-105. http://portal.amelica.org/ameli/jatsRepo/47/471772005/index.html.

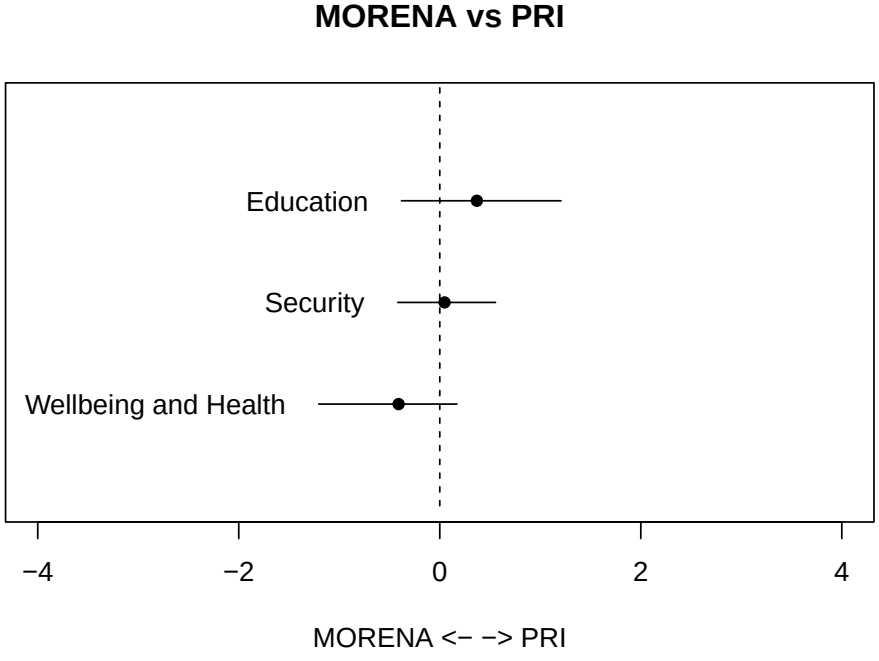
- Vernazza Mañana, E., and Vicente Villardón, J. "Discursos presidenciales en Uruguay: enfoque desde el análisis estadístico de texto" [Presidential speeches in Uruguay: approach from statistical text analysis], Cuadernos Del CIMBAGE 1, no. 23 (2021): 21-46. https://ojs.econ.uba.ar/index.php/CIMBAGE/article/view/2054.
- Warin, Thierry. "Structural Topic Modeling." Shiny. Accessed April 2024. https://warin.ca/shiny/stm/#section-evaluate.

9. Annex

9.1 Top words by top topic

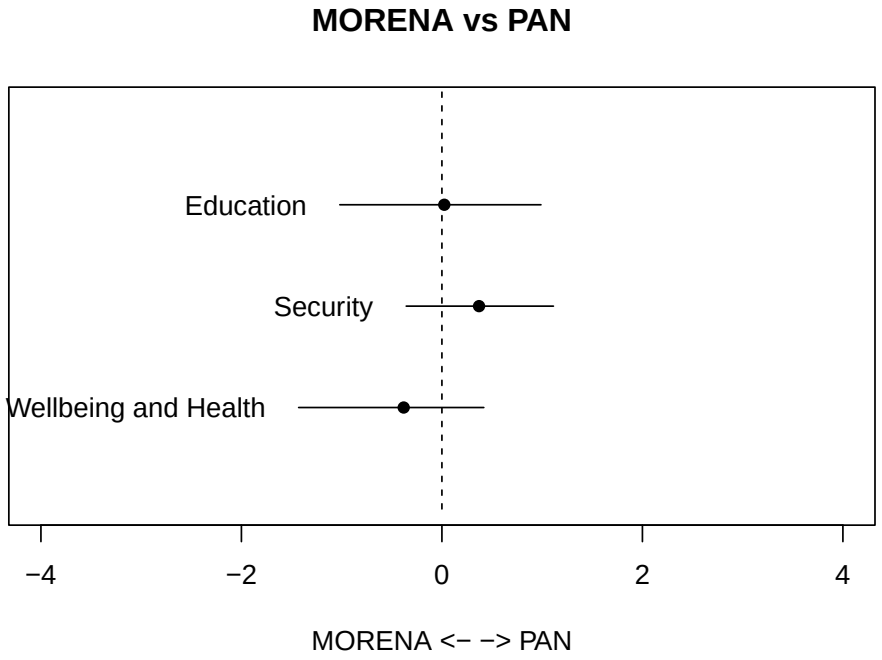
```
##   Topic Highest_Prob.1 Highest_Prob.2 Highest_Prob.3 Highest_Prob.4
## 1     1                     bienestar          salud         número
## 2     2                     educación         social          salud
## 3     3                     seguridad      inversión         social
##   Highest_Prob.5 Highest_Prob.6 Highest_Prob.7 Highest_Prob.8  Highest_Prob.9
## 1           mide         social      seguridad     california     federativas
## 2      seguridad       derechos       mediante        mujeres       inversión
## 3      programas      educación          salud    federativas infraestructura
##   Highest_Prob.10    FREX.1     FREX.2        FREX.3         FREX.4   FREX.5
## 1           datos     covid  bienestar    parámetros    prioritario     mide
## 2       programas inclusión innovación     violencia implementación impulsar
## 3        anterior  semestre    empleos oportunidades           real    igual
##          FREX.6     FREX.7       FREX.8     FREX.9    FREX.10      Lift.1
## 1  continuación valoración    parciales disponible  coalición coaliciones
## 2 productividad   docentes  vinculación     acceso planeación     sexenio
## 3    representa   distrito porcentuales     puntos    anuales    proárbol
##          Lift.2       Lift.3       Lift.4    Lift.5      Lift.6       Lift.7
## 1 imssbienestar        covid     tonelada  pandemia feminicidio sectorizadas
## 2      prospera federativas1 imssprospera    hambre   comedores aprendizajes
## 3   competitiva         ifai   superación megawatts    programó       diesel
##      Lift.8     Lift.9         Lift.10 Score
## 1 coalición      votos   afromexicanas     1
## 2   cruzada incluyente   desaparecidas     2
## 3 acuíferos criminales centrooccidente     3
```

Output 1: Top words in each topic

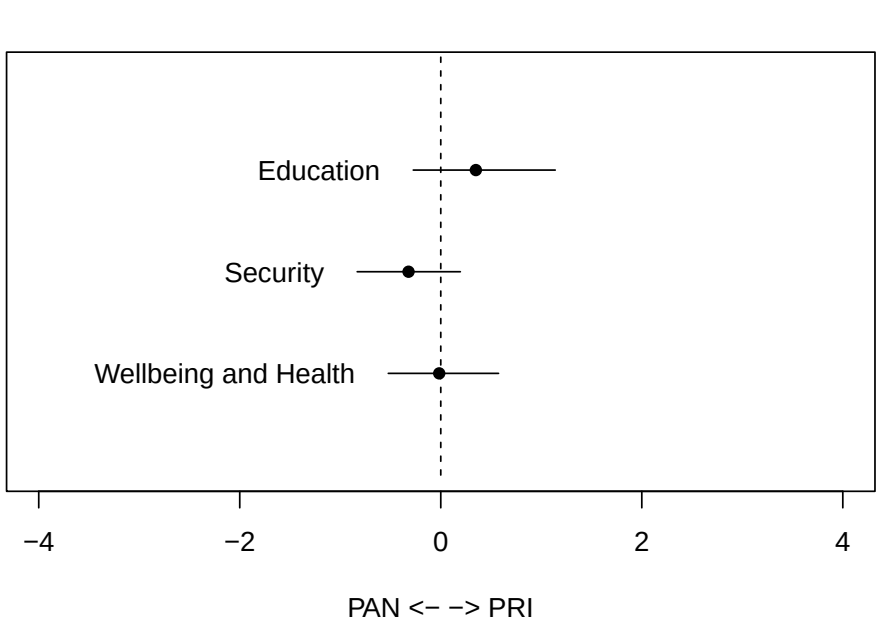


9.2 Topic Prevalence by Party

Annex Figure 1: Comparison of MORENA and PRI on each topic



Annex Figure 2: Comparison of MORENA and PAN on each topic



Annex Figure 3: Comparison of PAN and PRI on each topic