”Mañaneras” Data Analysis: Trending topics in México

1st Alma Rosa Cuevas Romero
School of Engineering and Sciences
Tecnológico de Monterrey
Monterrey, México
A00825413@tec.mx

2nd Luis Kevin Cepeda Zapata
School of Engineering and Sciences
Tecnológico de Monterrey
Monterrey, México
A00824840@tec.mx

***Abstract*—The ”Mañaneras” are conferences that the president of Mexico holds each weekday to answer various questions, explain policies, give directions, and impart historical lessons. This article analyzes the dialogues in the ”morning” transcripts, in which the most relevant themes and their evolution over time are examined. Through this analysis, we seek to understand the trending topics in the interactions between Andrés Manuel López Obrador, President of Mexico, and the participants, primarily representatives of the Mexican press, spanning from December 7, 2018, to July 10. 2023. This study offers a detailed view of current and changing concerns in the country, as well as how these conversations influence political dynamics and decision-making in Mexico.**

***Index Terms*—Mañaneras, AMLO, trending topics, México**

I. Introduction

In the age of information and instant communication, the role of political leaders as communicators has taken on a new dimension. In this context, the president of Mexico, Andrés Manuel López Obrador, has adopted a unique practice that distinguishes him from his predecessors: the morning conferences popularly known as ”Las Mañaneras”. This is opposite to what has been found in [1], as they state that populist leaders tend to use more social networks for communications instead of show talks. These daily sessions offer an unfiltered window into how the country’s leader addresses current affairs while interacting with the press and some citizens.

In a country facing a diverse range of challenges ranging from security, economy, public health, and corruption, the Mañaneras have become an essential source of information and analysis for Mexicans. These conferences have proven to be a powerful tool for understanding the priorities and strategies of the current government. One way to explore and unravel the most pressing concerns of the Mexican people is by analyzing the trending topics that emerge from the dialogues held in these sessions.

Although there are many benefits from having a direct dialogue with the elected head of the republic, there are also many concerns about having these regular meetings. Some opposites criticize these events as they are biased towards merely the president’s desires. Also, the rival parties are always a target for the president, who constantly assures that his government is better in all affairs. These claims usually are unsupported by the data and are referred instead to other unknown data.

This article will delve into the analysis of the dialogues that have taken place in López Obrador’s Mañaneras, focusing on identifying and understanding the most outstanding and recurring themes that reflect the current concerns of Mexico under the president’s eyes. From fighting crime to environmental issues and economic challenges, Mañaneras offer real-time insight into government priorities and how civil society voices influence the political agenda.

Bidirectional Encoder Representations from Transformers (BERT) [2] is an influential Natural Language Processing (NLP) technique that has gained a lot of popularity for its groundbreaking advancements in natural language understanding and processing. Among the applications where it has been used are: Improving Search Engines, Enhancing Chatbots and Virtual Assistants, and Transforming Sentiment Analysis. Other NLP methods may be inadequate due to language barriers, however there are multilingual transfer models of BERT [3] that can be adapted. BETO is the spanish version and it has been trained with a database as big as BERT and its effectiveness has been proven by [4].

Text similarity analysis is also a NLP technique that aims to quantify how similar two pieces of text are to each other. It is a valuable tool for applications including information retrieval, document clustering, recommendation systems, and plagiarism detection, among others. In this case, it could be used to assess how similar are the conferences of president and to obtain a clusted of the most important topics [5]. There are different methods and metrics, including Beto-based Embeddings.

BERT has been applied previously to identify political fake news in social networks in Romania [6]; for detecting political leaning in social media [7], and for decoding the politcal message of Korean presidents and their ideologies [8].

Through this analysis, we seek to shed light on how political communication in this unique format can shape public perception, generate discussion, and catalyze actions in Mexican society. By examining emerging trends in Mañaneras, we can discover not only the country’s evolving concerns but also how these conversations influence decision-making and shape Mexico’s future.

Thanks to Tec de Monterrey for the scholarship tuition.

II. Materials and Methods

A. Data to analyze

The repository used is called "Morning Conferences of President Andres Manuel López Obrador (Mañaneras AMLO)" [9], which contains the transcriptions of the stenographic versions of the morning conferences of President Andres Manuel López Obrador since December 2018. It is contained in a CSV file that contains the person that spoke, the text said, the sentiment, number of words, day, month, and year.

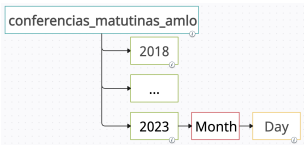

Fig. 1. Folder organization of the "Morning Conferences of President Andres Manuel López Obrador (Mañaneras AMLO)"

B. Text preprocessing

First, all the CSV files were concatenated in a single dataframe as they already have a timestamp. Then, all the transcripts from the "Mañaneras" underwent a text preprocessing process to remove punctuation, special characters, conjunctions, and stop words.

III. Method and Data

A. Word count

Each participant can intervene; each intervention has a word count averaged per conference. Thus, per year, we can see how long each intervention lasts. In other words, how long a comment turn lasts.

B. Participation count

After grouping the main characters into three groups, they counted how many times that group appeared. In this way we can know if there is a balance between actors or if one of the groups is the protagonist of the conferences.

C. The 50 most repeated words

Only one ngram was chosen, that is, only one word. We then classified these words according to their Part-of-Speech tag using Spacy's 'es_core_news_lg' model, a linguistic model in Spanish that is based on written text (such as news and media). We only kept the words that belong to the 'NOUN' group. We consider words like "good morning" or "inaudible" as noise; they are part of the conversation but do not contain any relevant information. The reason for using only the top 50 is that as more words are added, by eliminating repetitions based on lemmatization, relevant topics that are repeated per year begin to be eliminated. The results are the origin term, not its lemma.

To identify the most frequently used significant words, we utilized Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer, which weighs the word counts based on how often they appear in the documents.

TF-IDF is a tool used to convert a collection of text documents into numerical vectors. The purpose of TF-IDF is to highlight the importance of words in a document relative to a collection of documents (corpus). The terms are divided in two parts. Term Frequency Measures how often a term appears in a document. The equation is shown in (1) which shows the ratio of the apperances of a term in a document to the total of terms [10] where t is the term and d the document in the corpus D.

$$\text{TF}(t,d) = \frac{\text{\# of times term } t \text{ appears in doc. } d}{\text{Total \# of terms in doc. } d} \tag{1}$$

The Inverse Document Frequency measures the importance of a term across a corpus. Rarer terms receibe higher score. This term is calculated as the logarithm of the ratio of the total number of documents to the number of documents containing the term as shown in equation (2):

$$\text{IDF}(t,D) = \log\left(\frac{\text{Total \# of docs. in corpus } D}{\text{\# of docs. containing term } t+1}\right) \tag{2}$$

Once the two parts are considered, the TF-IDF is calculated as the product of equation (1) and (2) as shown in equation (3):

$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) \times \text{IDF}(t,D) \tag{3}$$

D. Clustering by topic modeling (using LDA)

The first task that would be done is select the most important topics in the data. Clustering by topic modeling, particularly using Latent Dirichlet Allocation (LDA), can be a powerful technique to determine the concerns of México in each year. LDA is a probabilistic model that assumes documents are mixtures of topics, and topics are mixtures of words. LDA can uncover the underlying themes in the text. This approach leverages natural language processing and machine learning to identify and categorize the key topics and concerns that dominate public discourse and media coverage.

To apply an unsupervised method for topic clusterization, we needed to:

- Take the transcripts and join them in a single .CSV, instead of multiple files per conference. Divide the information in years, ensuring a comprehensive snapshot of the country's concerns over time.
- The collected data was preprocessed to remove noise, such as stopwords, punctuation, and special characters. It was tokenized and lemmatized to standardize the text for analysis. For that, SnowballStemmer [11] in its Spanish version was used. Stemming maps different forms of the same word to a common "stem" - for example, the English stemmer maps connection, connections, connective, connected, and connecting to connect.
- To determine the country's concerns, we clustered similar topics.

IV. Results

A. Word Count

The word count considers every intervention made at conferences held within a year. The same person can have made multiple interventions repetitively. The first and last years do not represent a full 12 months of panels, but they provide a good indication of the trend over the meetings. From 2018 to 2020, the total number of words in interventions increased progressively. In 2021, the duration of interventions varied significantly, with outliers representing very long or concise interventions. Overall, the number of words per intervention has remained stable at around 35 to 40 words.

2018 had more non-repeated words within the top 50 most repeated words per year because its most repeated terms were more original in the short time the conferences took. If all the available words had been chosen, 2018 would have had the fewest original words.

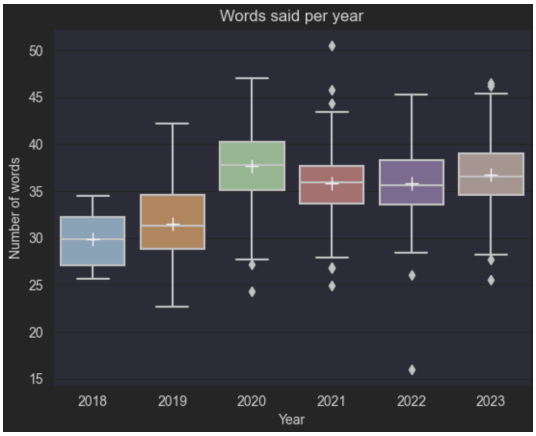

Fig. 2. Word count for each year. The first (2018) and last (2023) years don't represent the whole year.

B. Participation Count

Participants can be classified as 'President,' 'Public,' and 'Cabinet.' The president is Andrés Manuel López Obrador, he is the representative of the country, and he created the conferences. The audience can be interviewers from the media, citizens, or anyone who wants to ask a question at the conference. The Cabinet consists of government representatives who are specialized to answer questions, for example, the secretary of health.

The councils tend to appear more than participants, but their interactions seem more balanced than the president's appearances.

C. Unique words from the 50 most repeated terms

The text has been translated from Spanish to English. Most of the words had only one ngram, but due to translation, there were two exceptions. The word 'fiscalía' translates to 'prosecutor's office', and 'Luna' refers to García Luna, not the moon. Although lemmatization was used, the translation of the words led to the appearance of 'doctors', which corresponds to the Spanish word 'médicos', and 'doctor', which comes from the Spanish 'doctor'.

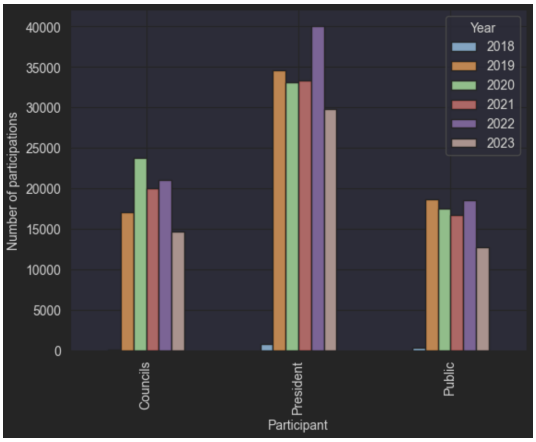

Fig. 3. Number of total participations per participant per year.

Fig. 4. Unique words across the Top 50 most repeated terms per year.

D. Clustering by topic modeling (using LDA)

It should be noted that the analysis was carried out in Spanish and that we are translating and interpreting the meanings of the results to the best of our ability. LDA detected 38 topics. However, LDA didn't mention all of them in the final result due to the low percentage contribution they had.

For the graph, the contribution of the most important topics was illustrated. The fact that other topics were not represented does not mean they did not exist. However, no different theme was repeated enough to become a topic.

1) 2018: The 'Mañaneras' began at the end of 2018; the analysis of LDA topics indicated that the topics were focused on proving that the new government had no obstacles that would make it impossible for them to meet their goals. Appreciating the post-revolutionary advances and comparing how things have changed. This year's topics represent the introduction to the government, while if there were concerns in the country, the conferences focused on the prologue of the future of the new presidency.

2) 2019: In 2019, there was a lot of talk about "stopping." However, it was not a particular topic; the idea of stopping appeared in concerns such as stopping the wave of violence

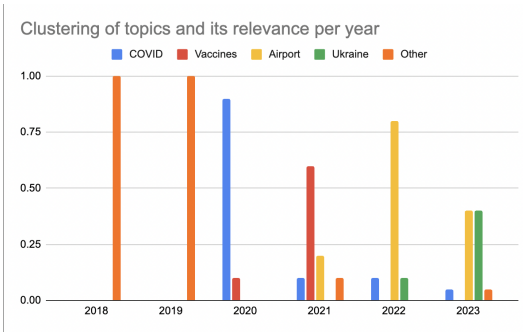

Fig. 5. Relevance of each cluster of topics given by the LDA processing.

in Sonora, stopping arms trafficking and crime waves, stopping officials who earn a lot of money, prevent fuel theft, among others. As it was seen, this year had too much violence and the people demanded strategic plans to stop the events that hurt Mexico.

Other words involved the case of the 80 thousand medical students who were going to aspire to a specialty and the concern that only eight thousand would be able to start with the specialty. In addition, much emphasis was placed on the National Transparency Institute, which ordered information regarding the cars the president used to get around.

3) 2020: In 2020, the coronavirus had a lot of weight during the conversations and the "plateau" in the number of COVID cases was sought. Similarly, Severe Acute Respiratory Syndrome (SARS) was mentioned, which is a severe form of pneumonia, and concern about the "resurgence" of the epidemic in the population.

The rest of the words indicated by the LDA analysis correspond to the names of vaccines such as Pfizer, Astrazeneca, Johnson and Johnson, Sputnik and CanSino, and pharmaceutical companies such as Birmex, Sinovac.

4) 2021: Topics started to shift. Covid was still a popular term but focused more on the vaccines than the disease per se.

The airport issues gained weight. There was speculation about "bondholders" and concerns that the debt to build the airport would increase.

Finally, an issue that gained weight was huachicoleo, which is the theft of fuels —such as gasoline— carried out by drilling the pipes transporting it.

5) 2022: By 2022, the pandemic topic was more wide and included vaccines, pharmaceuticals, and COVID-19; the epidemic ideas continued to a lesser extent.

The war in Ukraine and Russia was widely mentioned, but the airport stole the conversation. Particularly, the talk corresponded to the construction of the elevated viaduct for access at the departure level of the passenger terminal.

6) 2023: Like 2022, 2023 repeated the same ideas about the airport and COVID themes. Notwithstanding, the war between Ukraine-Russia was mentioned more. The consequences of the war could be seen; for example, the world's major oil companies began to sell their refineries and then came the crisis due to Russia's war with Ukraine, and oil and gasoline went through the roof.

It should be noted that for this year, only the transcript of the first quarter was available, and therefore, they do not represent the entire year.

V. Discussion

In the LDA results, 2019 mentioned "stopping" to refer to different types of violence; meanwhile, in 2021, fuel theft emerged. These two events do not share the exact words or semantic concepts. However, it is part of the violence that is experienced in Mexico. Since violence has so many aspects, LDA could not group them, and therefore, they seem invisible in the clustering of topics. That analysis did not represent topics such as missing persons, homicides, and kidnappings but was present in the conferences.

Furthermore, the annual analysis highlights the dominant themes, leaving some topics unexplored. Consequently, despite the occurrence of fuel theft in 2022, it was entirely eclipsed by the airport. If this investigation were to be revisited, a shift from yearly to monthly scrutiny would allow for a more comprehensive understanding of other facets of Mexico's landscape.

LDA groups by topics and offers the most relevant issues; the unique words of the most used terms only consider one word (ngram=1), and therefore, independent words that did not appear in the LDA can appear in this new list, like 'García Luna' in 2023. Some results like 'covid,' 'vaccination,' and 'airport' overlap with LDA results, but the fact that the LDA title is 'airport' does not mean that it appeared in the most used; LDA titles are intended to group or identify the whole topic, not just one word even though it was widely used.

VI. Conclusion

References

[1] N. Ernst, S. Blassnig, S. Engesser, F. Büchel, and F. Esser, "Populists prefer social media over talk shows: An analysis of populist messages and stylistic elements across six countries," *Social Media + Society*, vol. 5, no. 1, p. 2056305118823358, 2019.

[2] S. Alaparthi and M. Mishra, "Bidirectional encoder representations from transformers (bert): A sentiment analysis odyssey," 2020.

[3] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of bert," 2019.

[4] A. Quiñones, "Análisis de la sintaxis aprendida por beto, un modelo de lenguaje en español basado en transformers," 2021.

[5] J. Jia and X. Liu, "Improving systematic literature review based on text similarity analysis," *Journal of Physics: Conference Series*, vol. 1069, p. 012059, aug 2018.

[6] C. Busioc, V. Dumitru, S. Ruseti, S. Terian-Dan, M. Dascalu, and T. Rebedea, "What are the latest fake news in romanian politics? an automated analysis based on bert language models," in *Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education* (Ó. Mealha, M. Dascalu, and T. Di Mascio, eds.), (Singapore), pp. 201–212, Springer Singapore, 2022.

[7] J. Jiang, X. Ren, and E. Ferrara, "Retweet-bert: Political leaning detection using language features and information diffusion on social networks," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, pp. 459–469, Jun. 2023.

[8] C. J. Chung and H. W. Park, "Textual analysis of a political message: the inaugural addresses of two korean presidents," *Social Science Information*, vol. 49, no. 2, pp. 215–239, 2010.

[9] Nostrodata, “Conferencias matutinas del presidente andres manuel l opez obrador (mañaneras amlo). github. retrieved from: https://github.com/nostrodata/conferenciasmatutinasamlo.”
[10] S. Kaur, P. Kumar, and P. Kumaraguru, “Automating fake news detection system using multi-level voting model,” *Soft Computing*, vol. 24, pp. 9049–9069, Nov. 2019.
[11] S. Developers, “snowballstemmer,” *pypi*, 2014.