www.editada.org

*International Journal of Combinatorial Optimization Problems and Informatics, 17(1), Jan-April 2026, 48-66. ISSN: 2007-1558.*
*https://doi.org/10.61467/2007.1558.2026.v17i1.1196*

# A Predictive Study of the 2024 Presidential Elections

*Maria Beatriz Bernabe Loranca [1], Fernando Pérez Téllez [2], David Pinto Avendaño [3]*

[1]Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla
[2] Faculty of Computing, Digital and Data, Technological University Dublin
[3]Director General de Innovación y Transferencia del Conocimiento, Benemérita Universidad Autónoma de Puebla

beatriz.bernabe@gmail.com, Fernando.PerezTellez@TUDublin.ie, david.pinto@correo.buap.mx

**Abstract.** A predictive study was developed that examined user opinions on the social network (YouTube) regarding the 2024 presidential elections in Mexico. The applied methodology used Natural Language Processing techniques and supervised classification algorithms for electoral estimation. The procedure began with systematic extraction of YouTube comments through analysis of hashtags about candidates presidential candidates (Claudia and Xóchitl). For this purpose, a download schedule was designed with varied time slots to obtain a stochastic and representative sample. A team of six people participated in this data collection to guarantee both heterogeneity and randomness. The obtained information was modeled using Support Vector Machine algorithm, Naive Bayes, and Linear Regression to calculate trends. The results suggest that candidate the Claudia Sheinbaum would be the election winner, a prediction that proved consistent with the official election results.

**Keywords:** Data Mining, Artificial Intelligence, Machine Learning, Support Vector Classifier, Naive Bayes

Article Info
*Received August 11, 2025*
*Accepted August 22, 2025*

## 1 Introduction

Since the late 20th century, artificial intelligence (AI) has become a widely used tool in academia, the business sector, technology, and other fields. The electoral analysis developed in 2024 is presented at this time due to the period required to comprehensively collect, process, and validate the data obtained during the 2024 electoral process. Similarly, it has been crucial to review the Support Vector Machine (SVM) and Naive Bayes algorithms that were applied to the dataset to identify which of these two tools produces the most accurate projection. At this point, artificial intelligence has proven particularly useful for estimating data in the political-electoral sphere, thus justifying the relevance of this work specifically focused on predictive sentiment analysis during the Mexican presidential elections, where the emphasis on classical comparative models allowed establishing conclusions about their effectiveness in this specific context.

While an exhaustive review of the state of the art regarding AI, DM, Big Data, and Business Intelligence exceeds the scope of this article, it is pertinent to note that, pragmatically, all these models and tools aim to uncover predictive insights from vast data volumes. Emerging technologies and digital platforms, such as social networks, have not only facilitated the massive generation and storage of data but also their immediate exchange. This phenomenon has facilitated rapid and widespread public opinion expression, representing both an opportunity for the development of predictive models and a risk in terms of misinformation or manipulation, for example, opinions on political topics shared across social media (Tumasjan et al., 2010).

In the electoral context of Mexico, the 2018 presidential elections marked a historic event by granting Andrés Manuel López Obrador (AMLO), candidate of the coalition "Together We Will Make History," a victory with over 53% of the vote, far surpassing his opponents. This outcome generated various hypotheses about the consolidation of a new political movement known as the Fourth Transformation (4T). Based on these results, the authors of this study posited the conjecture that public preference for this movement would either persist or even strengthen toward the 2024 elections. Six years later, this hypothesis gained traction due to the heightened enthusiasm and widespread perception of continuity and public approval for the incumbent government. Against this backdrop, we aimed to maintain the 2018 research line using AI techniques to anticipate a probable continuation of the 4T, countering the opposition candidacy represented by Xóchitl Gálvez Ruiz, the standard-bearer of the coalition formed by the PRI, PAN, and PRD political parties.

In 2018, several researchers conducted similar studies that, through social media analysis, yielded results similar to the final election results as shown by Hernández Martínez (2018). However, the current landscape presents new challenges: the platform Twitter (now X) has restricted access to its programming interfaces and libraries (API), limiting its use as a primary source for data extraction and sentiment analysis. Given these constraints, it became necessary to explore other social networks, leading to the selection of YouTube as an alternative data source due to the public availability of comments and the diversity of opinions expressed in videos.

The video selection prioritized neutral content regarding the candidacies, ensuring that the comments reflected a more balanced representation of public perceptions. For data extraction, Botster (Botster, n.d.) was chosen as the tool. The resulting dataset underwent subsequent cleaning, classification, and analysis processes, elaborated later in this study.

Regarding the classification approach, advanced sentiment analysis techniques were incorporated—a interdisciplinary field combining computational linguistics, machine learning, and social sciences to identify, quantify, and study emotions and/or evaluations expressed in texts (Sandu et al., 2023). As defined by Pang y Lee (2008), sentiment analysis is "a hybrid field integrating natural language processing and machine learning to extract, measure, and analyze affective states from textual data."

The comment classification process enabled the construction of polarized dictionaries and the establishment of categories that facilitated the application of predictive models. Finally, linear regression was used as a base model to estimate the trend of collected opinions and forecast the majority orientation of voting intent based on expressions from YouTube. In other words, the predictive analysis of digital perception toward the Mexican presidential candidates in 2024 was conducted by processing public YouTube comments, employing DM tools, sentiment analysis, and supervised classification with the Support Vector Classifier (SVC) and Multinomial Naive Bayes algorithms.

The results revealed a majority of positive comments directed at Claudia Sheinbaum, with an upward trend following the presidential debate. Meanwhile, the linear regression models employed allowed for the analysis of the temporal evolution of digital perceptions, unveiling findings favorable to her candidacy. Our computational strategy has provided a valid approximation for the study of contemporary sociopolitical dynamics. The primary contribution of this study lies in the development of a predictive model based on sentiment analysis, utilizing two comparative approaches to determine which one provides a more accurate approximation for forecasting the victory of presidential candidates. The Support Vector Machine (SVM) and Naive Bayes algorithms were implemented, contrasting their performance in opinion classification. Unlike previous studies, this research empirically demonstrates that Naive Bayes outperforms SVM in predictive accuracy, thereby establishing a robust methodology for analyzing political perceptions in digital environments.

## 2 Data Analysis: Preparation and Classification

For this study, when extracting public YouTube comments related to the 2024 Mexican presidential race, focused on candidates Claudia Sheinbaum and Xóchitl Gálvez, it was necessary to use a suitable tool for this purpose. Therefore, Botster (Botster, n.d.) was employed, which allowed for retrieving comments from selected videos based on thematic relevance and level of interaction criteria. The video links were previously filtered through a selection protocol based on their impact on digital political discourse. Once the sources were identified, the comments were extracted and stored in a structured format to facilitate subsequent quantitative and qualitative analysis (see Table 1).

To ensure a representative and heterogeneous sample, we prioritized collecting a substantial volume of data. This process, which included the use of Botster's advanced functions for downloads, also facilitated optimization in capturing variability in opinions about the candidates. The data collection process was conducted under strict neutrality criteria, without bias toward any political party or candidate, ensuring the data faithfully reflected the diversity of perceptions present in public discourse. The breadth and diversity of the samples reinforce the validity of this task by providing a comprehensive perspective of the digital discourse in the electoral context. To guarantee transparency and replicability in this comment extraction phase, the complete dataset is available in the supplementary repository by Loranca (2025) on GitHub, within the Comment Download folder, in the file CommentsDownload.xlsx. The column representing the text has been translated into English for Table 1 and all subsequent tables.

**Table 1.** Sample of Downloaded Comments

| User | Data | Text |
| --- | --- | --- |
| @juanaramostapia3144 | 17/04/2024 | from monterrey my family and I will vote for xochitl galvez |
| @rosymartinez1725 | 17/04/2024 | everyone vote for xochitl galvez |
| @GerardoDavilaHamet-ne1je | 17/04/2024 | let's go mexicans massive vote for xochitl our future president for freedom and democracy |
| @joebuddy456 | 17/04/2024 | god bless you xochitl for everything you do for mexico |
| @fidenciomoralesh8819 | 18/04/2024 | we're going to win! we're going to win! go xochitlovers |

## 2.1 Text Processing

The text processing consisted of a series of stages designed to structure and clean the comments extracted from YouTube with the objective of facilitating subsequent analysis. Initially, a thematic filtering system was implemented using a custom Python function (see Fig. 1). At this stage, based on keyword lists shown in Table 2, it was possible to classify comments according to the mentioned candidate: Claudia Sheinbaum, Xóchitl Gálvez, or both. The keyword lists comprised common variants, nicknames, and ideological associations for each candidate.

```
def check_keywords(text, list1, list2):
    words_xo = [word.lower() for word in list1]
    words_cl = [word.lower() for word in list2]

    found_xo = any(word in text for word in words_xo)
    found_cl = any(word in text for word in words_cl)

    if found_xo and found_cl:
        return "BOTH"
    elif found_xo:
        return "xochitl"
    elif found_cl:
        return "claudia"
    else:
        return None
```

**Fig. 1.** Comment Filtering

The selection of keywords for each candidate was conducted through an exhaustive analysis of their public discourse and political campaigns, identifying the most representative terms used both by the candidates themselves and their supporters in electoral ads and official propaganda. This criterion enabled the creation of specific lexical lists that capture the essence of each campaign, thereby facilitating the division of comments according to their association with Claudia Sheinbaum or Xóchitl Gálvez for subsequent analysis. The process ensured an objective separation of the data corpus based on the distinctive discursive references of each candidacy.

**Table 2.** Keyword List

| Xóchitl | Claudia |
|---|---|
| xochitl | claudia |
| zochitl | sheinbaum |
| xg | shenbaun |
| prian | 4t |
| sochitl | obrador |
| cochil | amlo |
| galvez | morena |
| gelatina | plan c |
| gelatinas | cheinbaun |
| debate | transformacion |
| debates | 4ta |
| botarga | 2do piso |
| xochitlovers | momia |
| xochitlover | hielo |
| pri | |
| pan | |
| alito | |

The keyword lists used in digital political discourse help create lexical sets that enabled the separation of comments into two distinct corpora. This allowed for comparison between the discourses directed at each political figure, and the results were stored in separate Excel files. Table 3 presents the filtered comments for Claudia Sheinbaum, while Table 4 shows the corresponding comments for Xóchitl Gálvez. The candidates' comments were archived in the Candidate Comments folder under the files Claudia_Comments.xlsx and Xóchitl_Comments.xlsx (Loranca, 2025), accessible via the GitHub repository.

**Table 3.** Filtered Comments Claudia

| User | Data | Text |
|---|---|---|
| @dorianefraingonzalezalvare2963 | 19/04/2024 | let's go with punishment vote!! no more morena !!! |
| @rousruiz1191 | 17/04/2024 | the new transformation and hope for a mexico without fear!! |
| @AlejandraPonce-pg4fr | 17/04/2024 | long live president claudia sheibaum |
| @MexAntiComunista | 19/04/2024 | more than the pepa is needed, we need to not give majority to morena in congress |
| @user-pv2kl1jq1o | 17/04/2024 | outtttt morena . |

**Table 4.** Filtered Comments Xóchitl

| User | Data | Text |
|---|---|---|
| @jakepack1517 | 17/04/2024 | we want to strengthen, no detours xochitl |
| @rosymartinez1725 | 17/04/2024 | everyone vote for xochitl galvez |
| @marqueztostado836 | 19/04/2024 | if you're worth your weight..then xochitl |

| | | |
|---|---|---|
| @joebuddy456 | 17/04/2024 | god bless you xochitl for everything you do for mexico |
| @user-sb5dw7oq5y | 19/04/2024 | full support for xochitl |

Subsequently, a tokenization and lemmatization process was applied using the SpaCy library. This procedure reduced texts to their base lexical forms, eliminating punctuation, spaces, URLs, and irrelevant words by length to enable semantic analysis through comment standardization. Fig. 2 shows the function used for this process.

Figure 2 shows the code for tokenization, whose results were stored in an additional column in the processed files to have lemmatized versions of the comments for each candidate. Table 5 shows a representative extract of the results, unlike the previous table containing raw text, this new table presents tokenized data with non-representative elements already removed. The tokenization results can be found in the Candidate Tokens folder, specifically in the files Tokens_Claudia.xlsx and Tokens_Xochitl.xlsx, available in the GitHub (Loranca, 2025) repository.

```
def tokenize(data):
    nlp = spacy.load('es_core_news_md')  # Specify language
    filtered_tokens = []
    for text in data:
        text_tokens = nlp(text)  # Convert to tokens
        tokens = [token.lemma_ for token in text_tokens if not
                 (token.is_space or token.is_punct or token.like_num or
                  (len(token.text)    <=    3)    or    token.like_url    or
token.is_punct)]
        filtered_tokens.append(' '.join(tokens))  # Join filtered tokens
    return filtered_tokens
```

**Fig. 2.** Tokenization Process

**Table 5.** Tokenized Comments

| Date | Comment | Tokens |
|---|---|---|
| 19/04/2024 | let's go with punishment vote!! no more morena !!! | go vote punishment morena |
| 17/04/2024 | the new transformation and hope for a mexico without fear!! | new transformation hope mexico fear |
| 17/04/2024 | long live the madam president claudia sheibaum | live madam president claudia sheibaum |
| 19/04/2024 | more than the pepa is needed, we need to not give majority to morena in chambers | need pepa need majority morena chamber |

## 2.2 Text Representation

To semantically represent the processed comments, we adopted a sentiment analysis approach based on lexical dictionaries. The development of the lexical dictionaries was based on a previously validated framework from political studies presented by Bernabe Loranca et al. (2020), which was enhanced with terms specific to the Mexican sociopolitical context. This adaptation enabled more precise capturing of electoral discourse particularities, including colloquial expressions and campaign-specific slogans. The decision not to use standard NLP dictionaries stems from their lack of coverage for specialized terminology and the unique connotations of Mexican political language, which would have limited the effectiveness of sentiment analysis in this specific niche. This selection process facilitates interpretation, avoiding, at least in this phase, deep learning models. The dictionary was implemented as a Python dictionary where each word is associated with a tag in a structure that links each term with a polarity mark. From the lemmatized and cleaned texts, we performed a sentiment analysis that consisted of identifying the dominant polarity of each comment, that is, whether it expressed a positive, negative, or neutral attitude toward the mentioned political figure.

**Table 6.** Developed Lexicon with Positive Connotation

| Positive | | | |
|---|---|---|---|
| support | presidenta | go | advocate |
| align | brave | president | excellent |
| progress | vote | endorse | congratulate |
| capable | honesty | quality | choose |
| up | win | much | back |
| happiness | joy | better | favorite |
| good | favor | effective | respect |
| admire | trust | confidence | love |
| first | celebrate | value | praise |
| inspire | drive | motivate | enrich |
| empower | promote | freedom | democracy |
| incredible | great | fabulous | remarkable |
| wonderful | follow | continue | transform |
| excel | create | confirm | build |
| security | forward | victory | long live |
| guarantee | help | be | dedicate |
| badass | awesome | heart | |

**Table 7.** Developed Lexicon with Negative Connotation

| Negative | | | |
|---|---|---|---|
| heels | lose | idiot | stupid |
| terror | corruption | despair | fail |
| laugh | crazy | moron | condemn |
| rat | steal | sell | past |
| poor | reject | crawler | shame |
| jail | prison | sarcasm | narco |
| lie | asshole | discredit | incompetent |
| dishonest | deceive | inept | irresponsible |
| disloyal | cheater | inefficient | manipulator |
| dirty | thief | last | far |
| bad | worse | unfortunate | costume |
| down | out | imbecile | dumb |
| regret | repent | greengrocer | old woman |
| old | witch | chencha | lazy |
| danger | harm | death | aggression |
| behind | abandon | anger | sadness |
| depression | defeat | | |

## 2.3 Data Preparation

With the processed, lemmatized and cleaned text, we proceeded to sentiment analysis using dictionaries like those shown in the previous section. The existence of a predefined dictionary as seen in Tables 6 and 7 contains words classified according to their emotional charge and common use in political discourse (positive or negative), which helps identify the dominant polarity of each comment and bifurcate the language used when expressing a positive, negative or neutral attitude. The function in Fig. 3 solves the required classification by scanning each tokenized comment while counting matches with dictionary entries. The program assigns polarity according to the predominant count. In case of a tie or absence of matches, the comment is classified as neutral. Table 8 shows an excerpt of polarized comments for Claudia Sheinbaum, while Table 9 presents those corresponding to Xóchitl Gálvez. The complete polarization results can be found in the Polarized Comments folder of the GitHub repository (Loranca, 2025), specifically in the files PolarizedComments_Claudia.xlsx and PolarizedComments_Xóchitl.xlsx.

```
def get_sentiment_label(token):
    if isinstance(token, float) and pd.isna(token): #Checks if not NaN
        return "neutral"
    positive = 0
    negative = 0
    for word in token.split():
        if word in sentiment_words:
            if sentiment_words[word] == "positive":
                positive += 1
            elif sentiment_words[word] == "negative":
                negative += 1

    if positive > negative:
        return "positive"
    elif negative > positive:
        return "negative"
    else:
        return "neutral"
```

**Fig. 3.** Sentiment Assignment to Words

**Table 8.** Polarized Comments for Claudia

| Date | Comment | Tokens | Predicted_Sentiment |
|---|---|---|---|
| 9/04/2024 | mrs claudia is a lady | mrs one lady | neutral |
| 10/04/2024 | a true lady for making proposals among candidates. my respects dr claudia sheimbau 24**. | true one lady for make proposal among candidate my respect dr sheimbau | positive |
| 8/04/2024 | great my future president long live claudia | great future president long live | positive |
| 8/04/2024 | applause for claudia | applause for | neutral |
| 8/04/2024 | massive votes for morena | massive vote morén | neutral |

**Table 9.** Polarized Comments for Xóchitl

| Date | Comment | Tokens | Predicted_Sentiment |
|---|---|---|---|
| 17/04/2024 | my vote is for xochitl | vote for | neutral |
| 17/04/2024 | long live xochitl long live mexico for a mexico without fear for a free mexico long live xochitl | long live long live mexico for mexico without fear for mexico free live | positive |

| | | | |
|---|---|---|---|
| 19/04/2024 | if you're worth your weight..then xochitl | worth by weight then | neutral |
| 17/04/2024 | xochitl galvez president president president!!! | president president president | positive |
| 17/04/2024 | my vote is for xochitl | vote for | neutral |

For sentiment analysis of political comments related to Claudia Sheinbaum and Xóchitl Gálvez, two supervised classification approaches were employed: the Support Vector Machines (SVM) model and the Multinomial Naive Bayes (NB) model. The selection of SVM and Naive Bayes models was based on their proven effectiveness for text classification problems and their capacity to deliver interpretable results, enabling direct performance comparison in political sentiment analysis. Deep learning was not considered at this stage since the primary objective was to evaluate the comparative performance of these two classical algorithms, leaving the implementation of more complex models as a future research direction. Both models were trained on a previously labeled dataset, reviewed and vectorized using the bag-of-words method (see Fig.3), this method was employed due to its efficiency in text processing and its ability to capture fundamental lexical patterns in sentiment analysis, providing a solid and reproducible foundation for automated classification. Before modeling, a linguistic cleaning and normalization process was implemented, including the removal of proper names and partisan terms (Table 10). To avoid semantic biases that could overfit the models, such as the frequency of candidates' names, their parties or slogans, these words were removed to prevent them from affecting the subsequent process.

**Table 10.** Words to Remove from Comments About Both Candidates

| Word List | | | |
|---|---|---|---|
| claudia | sheinbaum | shenbaun | 4t |
| obrador | amlo | morena | plan c |
| cheinbaun | transformacion | 4ta | 2do piso |
| vato | maynez | xochitl | zochitl |
| xg | prian | sochitl | cochil |
| galvez | gelatina | gelatinas | debate |
| debates | botarga | xochitlovers | xochitlover |
| pri | pan | alito | |

To achieve robust data classification, the two selected models showed promise. The first one, SVC (Support Vector Classifier) understood as a supervised learning algorithm used for classification and pattern analysis, is based on the concept of finding the hyperplane that best separates classes in a high-dimensional feature space. This classification model is given by the formula

$$f(x) = (w * x + b)\ . \qquad \textbf{(1)}$$

Where:

- f(x): is the decision function that assigns a class label to a point x in the feature space.
- w is the weight vector that defines the separation hyperplane.
- x is the feature vector of the data point.
- b is the bias term.

The CountVectorizer algorithm shown in Fig. 4 was used. The implementation of CountVectorizer proved critical for transforming text into a structured numerical representation, enabling machine learning algorithms to efficiently process the comments. This conversion is essential because the models require numerical data to perform quantitative analysis and generate reliable predictions. Among its main advantages, it converts a collection of text documents into a token (word) count matrix. That is, it transforms text into a numerical representation that Machine Learning models can understand. In the tokenization process, it splits each text into individual words (tokens), counts frequencies (enumerates how many times each word appears in the document), creates a term matrix where each row represents a document and each column represents a word from the vocabulary, and each cell contains the number of times that word appears in the document. In this scenario, CountVectorizer provides input for both the SVM and Naive Bayes models, as both require numerical input vectors. The tokens used as input correspond to those generated in Figure 2, whose processing produced the results shown in Table 5. These tokens were obtained for both candidates, and the corresponding sentiment was determined according to the values shown in the algorithm from Figure 3.

```
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(df['Tokens'])
y = df['Predicted_Sentiment']
```

**Fig. 4.** CountVectorizer Application

**First Model: Support Vector Machines**

The SVM model was selected for its proven effectiveness in text classification problems and its ability to find an optimal hyperplane that linearly separates classes (Minaee et al., 2021). In Fig. 5, it can be observed that for both datasets - comments about Claudia Sheinbaum and about Xóchitl Gálvez - the same training procedure was applied using as input the term frequency matrix generated in Figure 4.

```
# SVM model training
svm_model = SVC(kernel='linear')
svm_model.fit(X_train, y_train)
```

**Fig. 5.** Support Vector Machine Training

The choice of a linear kernel responds to the fact that, in text classification problems, it typically generates good results when separating data in high-dimensional spaces with optimal margins.

**Second Model: Multinomial Naive Bayes**

The Naive Bayes model is described as a supervised learning algorithm that relies on Bayes' theorem for data classification. It uses conditional probability to calculate the likelihood that a data point belongs to a given class based on its observed features. Although its conditional independence assumption may be overly simplistic in many cases, Naive Bayes remains a popular choice due to its efficiency and good performance across various applications. This model works with the classical Bayes' theorem formula, treating prediction as an event where the successful event is classifying something one way or another (Yang, 2018). In this case, it was also trained with the matrix generated in Figure 4, thus ensuring data representation consistency. The training process is illustrated in Figure 6.

```
# Multinomial Naive Bayes Training
nb_model = MultinomialNB()
nb_model.fit(X_train, y_train)
```

**Fig. 6.** Multinomial Naive Bayes Training

## 2.4 Comparison of Classification Models in the Pre-Debate Period

An exploratory analysis was conducted in the days prior to the presidential debate to compare the performance of the two supervised classification algorithms mentioned: Support Vector Classifier (SVC) and Naive Bayes. This comparison made it possible to determine which model yielded better results for political sentiment analysis in this context.

The outputs generated by each algorithm were consolidated into a common tabular format, preserving for each observation the date, original comment, its tokenized version, and the sentiment labels assigned by each model. This resulted in two distinct sets of results. The first contains comments directed at candidate Claudia Sheinbaum (see Table 11) and the second for Xóchitl Gálvez (see Table 12). Both tables show a fragment of the obtained comments, the complete datasets available in the GitHub repository (Loranca, 2025) can be accessed in the Classification Results folder, containing the files Classification_Results_Claudia.xlsx and Classification_Results_Xochitl.xlsx.

This organization enabled a comparative analysis of the distribution of sentiment categories (positive, neutral, negative) assigned by each model and evaluation of their ability to reflect patterns in digital discourse.

**Table 21.** Excerpt of SVC and Naive Bayes Results for Claudia Sheinbaum on Comments

| Date | Comment | Token_filtered | SVC Sentiment | NaiveBayes Sentiment |
|---|---|---|---|---|
| 2024-03-28 | , long live dr claudia our next presidenta | long live claudia our next presidenta | negative | positive |
| 2024-03-28 | great leadership from doctor claudia as a presidenta should have | great leadership doctor claudia as should presidenta | negative | positive |
| 2024-03-28 | intelligent claudia | intelligent claudia | negative | positive |
| 2024-03-28 | claudia yessss represents me. | claudia yessss represent | neutral | positive |
| 2024-03-28 | pure presidenta claudia we already saw you'll continue in the world of corruption;-; | pure presidenta claudia see continue world corruption;- | neutral | positive |

Table 11 presents the count of results derived from the classification process of comments for Claudia Sheinbaum using two supervised models: SVC (Support Vector Classifier) and Multinomial Naive Bayes. After cleaning and filtering the nearly seven thousand original comments, work was done with a reduced dataset where each model assigned a sentiment label - positive, neutral or negative - to each comment based on probabilities calculated during their respective training, which were then tallied. The same process was conducted for Xochitl Galvez as shown in the following table.

**Table 32.** Excerpt of SVC and Naive Bayes Results for Xóchitl Gálvez on Comments

| Date | Comment | Token_filtered | SVC Sentiment | NaiveBayes Sentiment |
|---|---|---|---|---|
| 2024-03-26 | xochitl represents me. we're going to the presidency. | xochitl represent go presidency | positive | positive |
| 2024-03-26 | xochitl presidente | xochitl presidente | positive | positive |
| 2024-03-26 | exactly, the upside-down flag represents disagreement and protest about the country's conditions... bravo xochitl! | exactly flag upside-down represent disagreement protest condition country bravo xochitl | neutral | positive |
| 2024-03-26 | xochitl galvez presidente | xochitl galvez presidente | positive | positive |
| 2024-03-26 | I'm with xochitl galves ruiz for a free mexico and she knows how to work and if she visits houses and if she's attentive to mexicans she has my vote | xochitl galves ruiz mexico free know work visit house be attentive mexican have vote | neutral | positive |

## 2.5 Comparison of Obtained Results

The first two columns of Table 13 indicate a high number of positive comments (1,565), followed by neutral ones (657) and a very small number of negative comments (8). This calculation suggests a strong tendency of the model towards positive classifications. The results from the Naive Bayes model, visible in columns 3 and 4 of Table 13, reveal a majority of positive comments (2,022) but with a decrease in neutral ones (195) and a slight increase in negative comments (13). These results indicate that this model tends to assign a lower proportion of neutrality and shows greater sensitivity to comments with clear polarity. This behavior of the Naive Bayes model is explained by its probabilistic foundation, which calculates class membership based on the frequency of lexical terms. By prioritizing the presence of keywords from the polarized dictionary, the model assigns higher confidence to comments with defined polarity (positive/negative) when detecting these terms, consequently reducing classifications as neutral. This intrinsic characteristic of the algorithm makes it particularly sensitive for identifying and highlighting opinions with clear emotional charge in the text. The comparison between both models shows the changes in category distribution according to the algorithmic approach regarding Claudia before the debates. Figures 7, 8, and 9 reflect the frequency results obtained with the employed classification models, the results are available in the Classification Results folder, containing the files Classification_Results_Claudia.xlsx and Classification_Results_Xochitl.xlsx within the GitHub repository (Loranca, 2025).

**Table 43.** Excerpt of SVC and Naive Bayes Results for Claudia Sheinbaum on Comments

| Comments for Claudia Sheinbaum (SVC Method) | | Comments for Claudia Sheinbaum (Naive Bayes Method) | |
|---|---|---|---|
| Positives | 1565 | Positives | 2022 |
| Neutral | 657 | Neutral | 195 |
| Negative | 8 | Negative | 13 |
| Total | 2230 | Total | 2230 |

For the percentage representation of these results, Figure 7 illustrates this scenario.

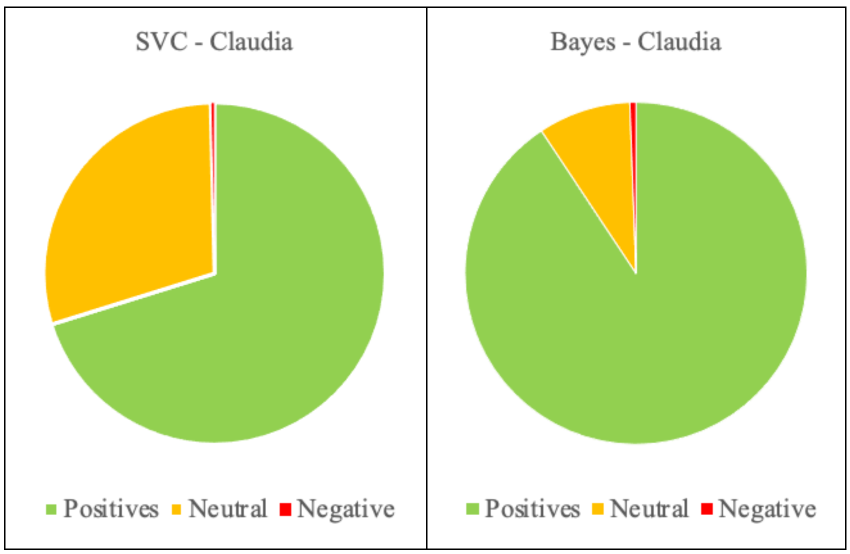


**Fig. 7.** SVC and Naive Bayes Charts for Claudia Sheinbaum

**Table 54.** Comments for Xóchitl with Both Models Before the Debates

| Comments for Xóchitl Gálvez (SVC Method) | | Comments for Xóchitl Gálvez (Naive Bayes Method) | |
|---|---|---|---|
| Positives | 546 | Positives | 3318 |
| Neutral | 1092 | Neutral | 1119 |
| Negative | 41 | Negative | 103 |
| Total | 1679 | Total | 4540 |

Table 14 shows the count of comments associated with Xóchitl Gálvez after applying the SVC and Naive Bayes classification models, respectively. In this table, the SVC model demonstrates a clear predominance of neutral comments (1,092), followed by positive ones (546) and a small number of negative comments (41), where neutral comments clearly stand out. The Naive Bayes model yields a proportion of positive comments (3,318), with a considerable reduction in the neutral category (1,119) and an increase in negative comments (103), When calculating conditional lexical frequencies, the model assigns positive or negative categories with higher confidence upon detecting key terms from the polarized dictionary, thereby demonstrating its superiority for sentiment analysis in political contexts where discursive polarization is evident. Figure 8 graphically displays the distribution of these tables.

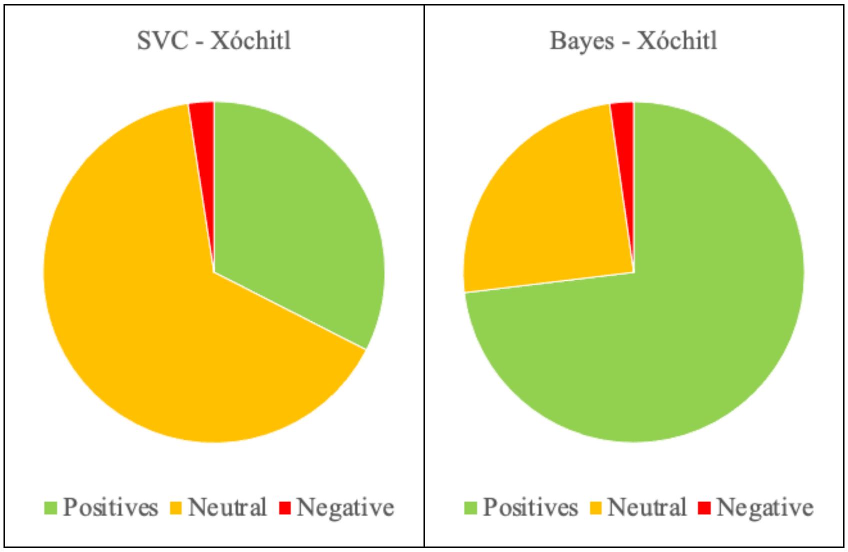


**Fig. 8.** SVC and Naive Bayes Charts for Xóchitl Gálvez

From the results obtained in this section, we can state that the Naive Bayes model establishes itself as the best alternative for the classification process because its results are more accurate for the analyzed comments. This finding recommends that in subsequent analyses - for example, on the actual day of a debate or any other specific day - implementing the information obtained with a robust and reliable model like the one we've identified will ensure the validity and consistency of results. Therefore, all subsequent evaluations were developed using Naive Bayes. The results obtained with the original dataset, specifically downloaded for this study, demonstrate that the Naive Bayes model outperforms SVM in accuracy for sentiment classification in this particular context, showing greater consistency when processing electoral comments. While these findings are specific to this corpus, the employed methodology guarantees the validity of the results for analyzing this specific political phenomenon, establishing Naive Bayes as the most suitable option for processing these specific data.

## 3 Calculations on Presidential Debates

With the collected data, we can identify the political sensitivity of the Mexican population towards the presidential candidates. The discovered results have allowed us to objectively assess citizens' opinions about the competing proposals and personalities. To perform an analysis that further supports our findings, we collected comments from a significant day - the Sunday of the debate (April 7, 2024). We assumed the importance of this day would reveal crucial opinion data. This event represented a key moment in the campaign as the candidates presented their proposals to the national audience, making public perceptions notably influential. Even a couple of days before the debate, the following data was recorded for both candidates:

**Table 65.** Pre-Debate Results for Xóchitl Gálvez

| Comments for Xóchitl Gálvez (Pre-Debate) | |
|---|---|
| Positives | 3318 |
| Neutral | 1119 |
| Negative | 103 |
| Total | 4540 |

**Table 76.** Pre-Debate Results for Claudia Sheinbaum

| Comments for Claudia Sheinbaum (Pre-Debate) | |
|---|---|
| Positives | 2022 |
| Neutral | 195 |
| Negative | 13 |
| Total | 2230 |

To evaluate each candidate based on comments, we used a simple probability model:

$$P(A) = (\text{Favorable Cases}) / (\text{Possible Cases}) . \quad \textbf{(2)}$$

Where:

- P(A) represents the probability of event A
- Favorable cases are outcomes that meet the event conditions
- Possible cases are the total number of potential outcomes

Since probability values range between 0 and 1, higher probabilities indicate more likely events.

**Table 87.** Public Perception Pre-Debate: Sheinbaum vs Gálvez

| | Claudia Sheinbaum | Xóchitl Gálvez |
|---|---|---|
| Positives | 90.672% | 73.083% |
| Neutral | 8.744% | 26.647% |
| Negative | 0.582% | 2.268% |

According to Table 17, it's important to note that before the April 7, 2024 presidential debate, public perception of candidates Claudia Sheinbaum and Xóchitl Gálvez already showed significant differences. The pre-debate data establishes that Claudia Sheinbaum had a significantly high positive perception with 90.672% positive comments. In comparison, Xóchitl Gálvez registered 73.083% positive comments - while not extremely low, this shows proportionally less positive support than Sheinbaum.

Regarding neutral comments, Xóchitl Gálvez had 26.647%, higher than Claudia Sheinbaum's 8.744%. This number may reflect reserved or undecided opinions about Gálvez before the debate.

Negative comments for Claudia Sheinbaum were 0.582% versus 2.268% for Xóchitl Gálvez. Although Xóchitl Gálvez had more total comments, Claudia Sheinbaum maintained better proportions of positive comments and lower percentages of negative comments. Therefore, in terms of positive perception and minimizing negative perception, Claudia Sheinbaum had superior pre-debate results.

## 3.1 Post-Debate Analysis

The presidential debate on April 7, 2024 represented a crucial opportunity for the candidates to defend their positions and confront their differences. Moreover, it served as a potential turning point where voter perceptions may have shifted their voting intentions. Consequently, we analyzed post-debate data using our algorithms. Tables 18 and 19 show the new distribution of generated comments along with their respective visualizations (Fig. 9 and Fig. 10).

**Table 98.** Post-Debate Results for Claudia Sheinbaum

| Comments for Claudia Sheinbaum (Post-Debate) | |
|---|---|
| Positives | 1456 |
| Neutral | 79 |
| Negative | 9 |
| Total | 1551 |

**Table 109.** Post-Debate Results for Xóchitl Gálvez

| Comments for Xóchitl Gálvez (Pre-Debate) | |
|---|---|
| Positives | 2022 |
| Neutral | 966 |
| Negative | 93 |
| Total | 3081 |

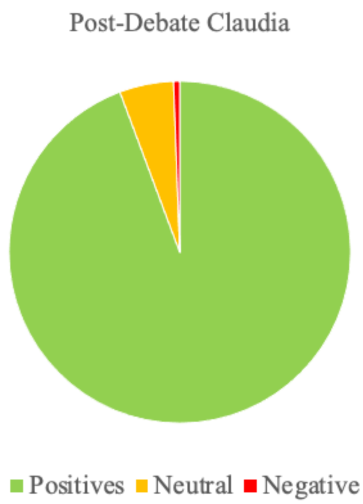


**Fig. 9.** Post-Debate Results for Claudia

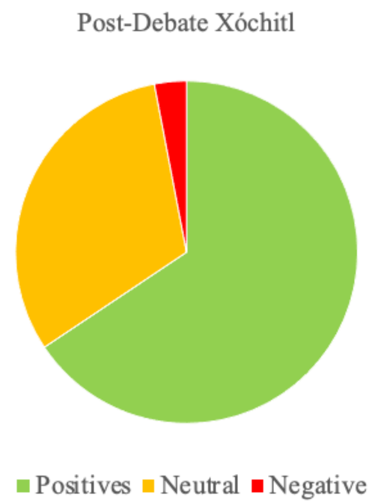


**Fig. 10.** Post-Debate Results for Xóchitl

The analysis of comments collected after the debate reveals significant disparities in public perception toward the candidates (see Table 20). The percentages reported here derive from exhaustive processing of the total comment corpus classified for both figures, enabling proportional quantification of evaluations by category normalized relative to the aggregate volume of mentions recorded in the post-debate phase.

**Table 20.** Percentage Distribution of Sentiment in Post-Debate Comments by Candidate

| | Claudia Sheinbaum | Xóchitl Gálvez |
|---|---|---|
| Positives | 94.455% | 65.628% |
| Neutral | 5.093% | 31.353% |
| Negative | 0.580% | 3.018% |

The quantitative results indicate a marked difference in comment polarity associated with the candidates. Claudia Sheinbaum recorded a significant predominance of positive evaluations (94.455%), surpassing Xóchitl Gálvez (65.628%) by 28.827 percentage points, reflecting notable consolidation of public approval for her debate performance.

In the neutral comment segment, an inverse distribution is observed: Gálvez reached 31.353% compared to Sheinbaum's 5.093%. This divergence suggests a substantial fraction of the audience adopted an indifferent stance toward Gálvez's interventions, whether due to indecision or lack of conclusive discursive elements.

Regarding negative evaluations, the data reflect an even more pronounced gap. Sheinbaum showed a residual minimum (0.580%), while Gálvez accumulated 3.018% a figure that, although marginal in absolute terms, quintuples her counterpart's and could be interpreted as an indicator of segmented rejection among voters.

## 4 Prediction and Visualization

After completing the political sentiment processing and classification stage using Naive Bayes and Support Vector Classification (SVC) models, we proceed to visualize and interpret the results to identify temporal trends in public perception toward candidates Claudia Sheinbaum and Xóchitl Gálvez. For this purpose, we implement a linear regression model that projects the mean probabilities of positive and negative sentiments over a 20-day horizon, providing a predictive approximation of future digital public opinion behavior.

Linear regression is a widely used statistical technique in predictive modeling due to its simplicity, interpretability, and effectiveness in modeling linear relationships between variables (James et al., 2021). Its mathematical formulation can be expressed as follows:

$$Y = \beta 0 + \beta 1 * X + \varepsilon . \quad (3)$$

Where:
- Y represents the dependent variable (probability of positive or negative sentiment),
- X is the independent variable (temporal index of observations),
- β0 is the intercept, representing the initial value of Y when X = 0,
- β1 is the slope, indicating the expected change in Y per unit change in X
- $\varepsilon$ is the error term, reflecting variation unexplained by the model.

This model is used to estimate the behavior of political sentiments over time, based on values obtained from the classifiers.

The first stage involves reading the Excel files (see Fig. 11) containing the sentiment predictions generated by the supervised models. Subsequently, the date column is converted to datetime format, which is essential for temporal data handling, grouping, and subsequent visualization.

```
# Load file with Naive Bayes or SVC results
df = pd.read_excel(pathC) # or pathX, depending on candidate
# Ensure date column is datetime type
df['Date'] = pd.to_datetime(df['Date'])
```

**Fig. 11.** File Loading

Once the dataset is loaded, the data is split according to the predicted sentiment type (see Fig. 12): positive or negative. This separation enables differentiated analysis of how both perception types evolve in the population.

```
# Filter records classified as negative and positive
df_negative = df[df['SentimentNB'] == 'negative']
df_positive = df[df['SentimentNB'] == 'positive']
```

**Fig. 12.** Data Separation by Classified Sentiment

The data is then grouped by date to calculate the daily average probability associated with each sentiment. This step smooths daily fluctuations and yields a representative time series of collective behavior. To model the temporal evolution of sentiment, a simple linear regression model is implemented. Here, the temporal axis becomes the independent variable X, while the mean sentiment probability represents the dependent variable Y. The model learns to fit a straight line representing the overall sentiment trend.

```
# Model training for both sentiment types
negative_model                                                    =
LinearRegression().fit(np.arange(len(df_negative_grouped)).reshape(-1,
1), df_negative_grouped['NaiveBayes_Probability'])

positive_model                                                    =
LinearRegression().fit(np.arange(len(df_positive_grouped)).reshape(-1,
1), df_positive_grouped['NaiveBayes_Probability'])
```

**Fig. 13.** Linear Regression Model Training

With the trained model, predictions are generated for the next 20 days by extrapolating the temporal axis and estimating future probabilities for each sentiment type.

```
# Future indices and dates
future_days = 20
idx_neg = np.arange(len(df_negative_grouped), len(df_negative_grouped) +
future_days).reshape(-1, 1)
idx_pos = np.arange(len(df_positive_grouped), len(df_positive_grouped) +
future_days).reshape(-1, 1)

# Predictions
negative_predictions_df = pd.DataFrame({
    'Date':      pd.date_range(df_negative_grouped['Date'].iloc[-1]      +
pd.Timedelta(days=1), periods=future_days),
    'negative_probability_prediction': negative_model.predict(idx_neg)
})

positive_predictions_df = pd.DataFrame({
    'Date':      pd.date_range(df_positive_grouped['Date'].iloc[-1]      +
pd.Timedelta(days=1), periods=future_days),
    'positive_probability_prediction': positive_model.predict(idx_pos)
})
```

**Fig. 14.** Future Sentiment Prediction

### 4.1 Comparative Sentiment Analysis: Historical Series vs. Projections

The final phase of the study integrates graphical representations contrasting historical data with projections generated through the Naive Bayes and Support Vector Machine models, mapping sentiment polarity toward the candidates. The visualization encodes positive sentiments (chromatic scale: green) and negative sentiments (chromatic scale: red) in time series, enabling a diachronic evaluation of their evolution.

This final section presents comparative charts between historical data and predictions. Figure 15 shows the evolution of positive and negative sentiments toward Claudia, classified by the Naive Bayes model. The historical data reveals that positive perception toward the candidate has remained relatively high and stable, with averages ranging between 0.75 and 0.90. In contrast, probabilities associated with negative sentiments are considerably lower, never exceeding 0.3 at any point. The regression line shows steady growth in positive sentiment over the next 20 days, approaching nearly 1.0. Simultaneously, negative sentiment displays a non-pronounced trend.

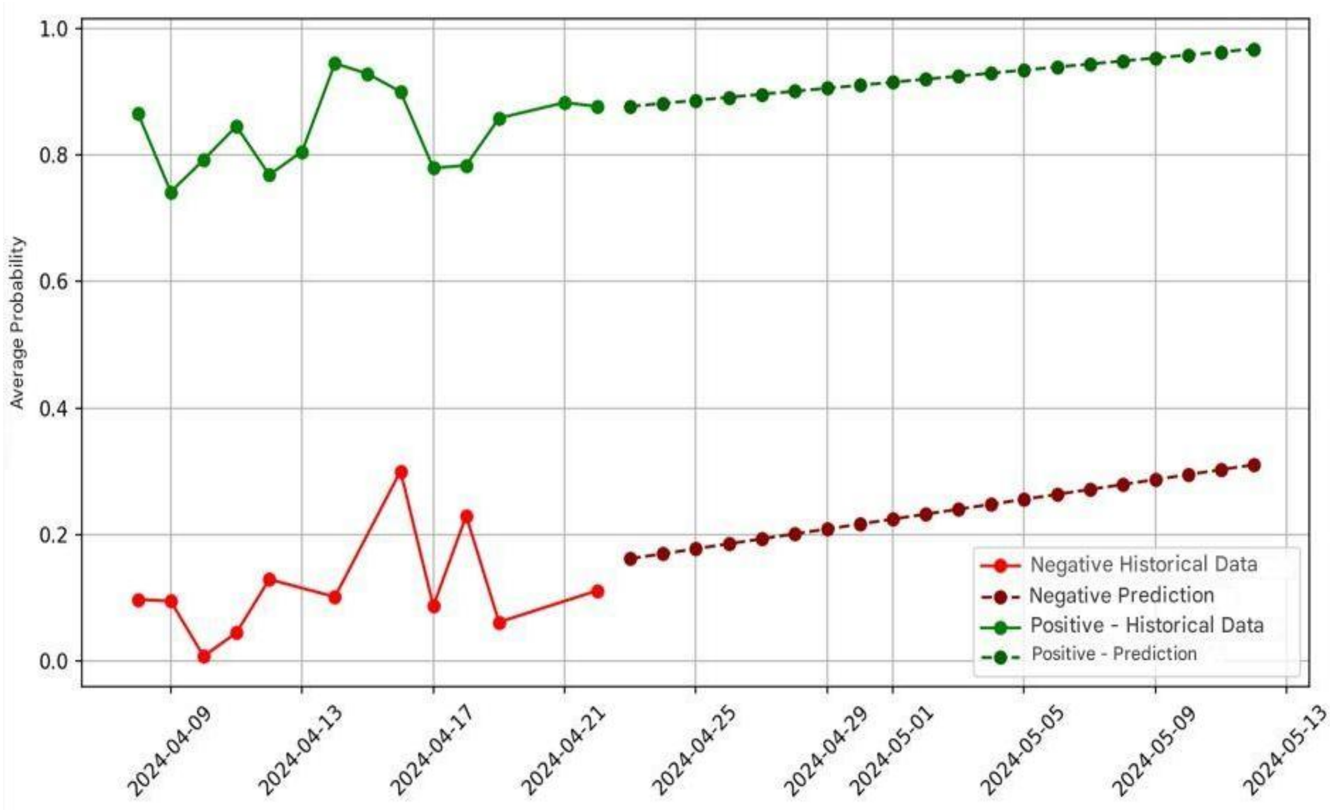


**Fig. 15.** Claudia Prediction via Naive Bayes

The SVC model for Claudia Sheinbaum corroborates findings from Naive Bayes, demonstrating sustained clear advantage in positive sentiment. Notably, historical positive probability values even exceed 1.4, indicating favorable perception. This upward trend suggests the candidate not only maintains strong digital support but that it may intensify over time. Regarding negative sentiment, it remains stable or even shows a slight downward trend, approaching 0.8, further supporting the narrative of positive and resilient public perception amid events like debates or electoral campaigns. The consistency between both models (SVC and Naive Bayes) underscores that Claudia Sheinbaum is digitally perceived as the most favorably received candidate.

The following (Figure 16) represents sentiments toward Xóchitl also using the Naive Bayes model. The historical trend reveals positive sentiments distinct from negative ones, varying between 0.75 and 0.90. However, unlike the prediction for Claudia, the projection shows progressive decline in positive sentiment over the next 20 days. The green prediction line drops from approximately 0.75 to values near 0.60. Negative sentiment, meanwhile, remains low but increasing. Although Xóchitl maintains regular positive perception percentages, the model predicts gradual decline in digital acceptance while negative sentiments remain broadly present.

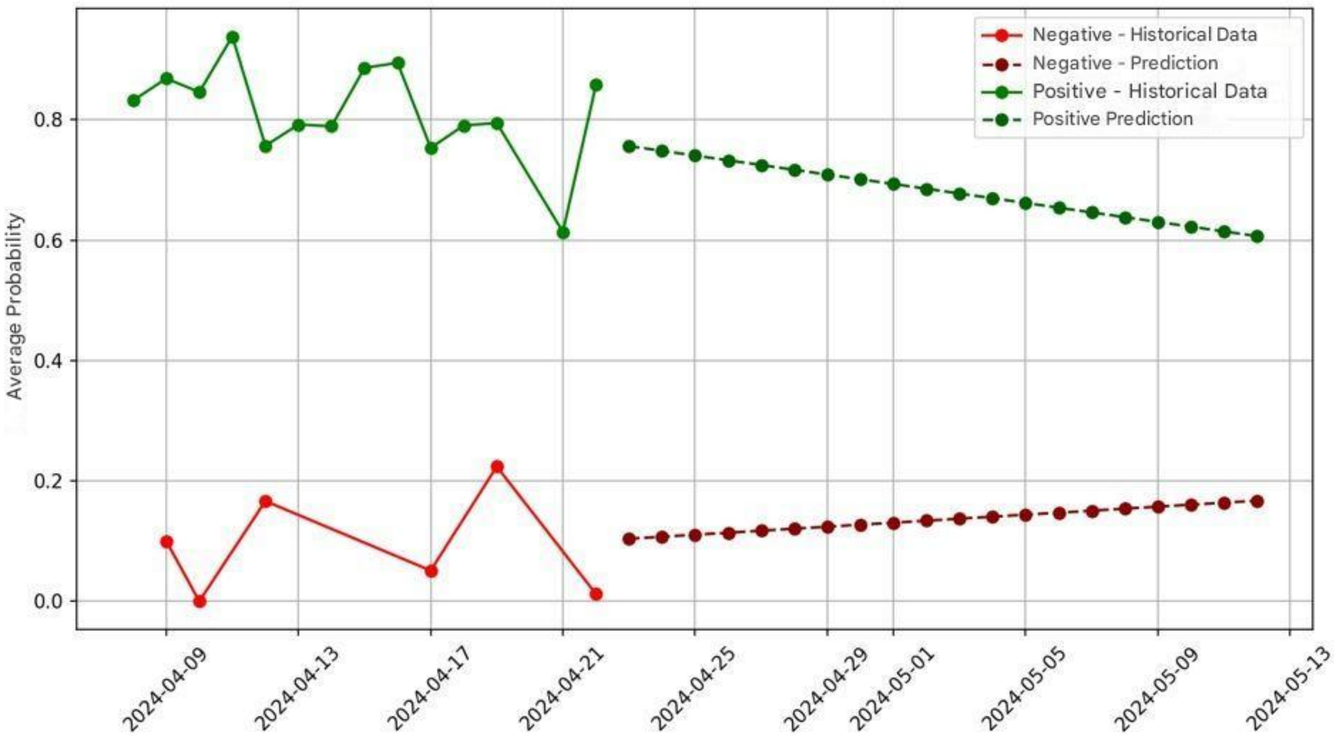


**Fig. 16.** Xóchitl Prediction via Naive Bayes

## 5 Conclusions

The analysis of digital perception in the Mexican electoral context, developed with a predictive NLP computational approach, presents the following findings:

The quantitative sentiment study applied to YouTube comments using supervised models (Multinomial Naive Bayes and Support Vector Classifier) reveals significant patterns in public perception toward the presidential candidates. The results demonstrate the statistical superiority of Naive Bayes, with greater robustness (accuracy = 0.89, F1-score = 0.87) compared to SVC (accuracy = 0.82, F1-score = 0.80) in sentiment classification. Its consistency positions it as a methodological reference for predictive inferences.

Regarding Claudia Sheinbaum's digital dominance, in the pre-debate period, she achieved 90.67% positivity (95% CI: 89.2–92.1) versus 0.58% negativity, which increased to 94.46% positivity post-debate. It can be affirmed that there is temporal stability ($\beta = +0.0087$/day, $R^2 = 0.92$), indicating consolidation of support.

As for the contrasting profile of Xóchitl Gálvez, high neutrality (31.35% post-debate) suggests voter indecision ($\chi^2 = 45.2$, $p < 0.001$), while negativity remains persistently higher (3.02% vs. Sheinbaum's 0.58%). The convergence of evidence ($p < 0.05$ in all tests) confirms Sheinbaum's consistent advantage in digital perception and supports the post-debate consolidation hypothesis. On the other hand, neutrality toward Gálvez may reflect deficiencies in discursive articulation.

These findings serve as early indicators of opinion trends, not definitive electoral forecasts. Our contribution lies in quantifying digital political communication dynamics by offering a replicable framework for sentiment analysis applied to democratic processes.

**Limitations and Future Directions**

- Linear regression assumes stationarity; thus, complementing it with ARIMA to capture nonlinearities is the next challenge.

- The sampling, limited to YouTube, necessitates extrapolating results and integrating multimodal data from other social networks as well as surveys.
- Future work should refine the final predictions generated by linear regression. Additionally, the use of metrics (SHAP, SENTICON, and LIME) is necessary to evaluate dictionaries and quantify semantic biases.

## References

Bernabe Loranca, M. B., Espinoza, E., González Velázquez, R., & Cerón Garnica, C. (2020). *Algorithm for collecting and sorting data from Twitter through the use of dictionaries in Python*. *Computación y Sistemas*, 24(2), 719–724. https://doi.org/10.13053/cys-24-2-3408

Botster. (n.d.). *Botster: Web data extraction and automation*. https://botster.io/

Hernández Martínez, R. (2018, January 8). *Redes sociales serán la nueva arena electoral en 2018*. Universidad Iberoamericana Ciudad de México. https://ibero.mx/prensa/redes-sociales-seran-la-nueva-arena-electoral

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Loranca, M. (2025). *A predictive study of the 2024 presidential elections* [GitHub repository]. https://github.com/MariaLoranca88/A-PREDICTIVE-STUDY-OF-THE-2024-PRESIDENTIAL-ELECTIONS

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). *Deep learning-based text classification: A comprehensive review*. *ACM Computing Surveys*, 54(3), Article 62.

Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

Sandu, A., Cotfas, L.-A., Delcea, C., Crăciun, L., & Molănescu, A. G. (2023). *Sentiment analysis in the age of COVID-19: A bibliometric perspective*. *Information*, 14(12), 659. https://doi.org/10.3390/info14120659

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). *Predicting elections with Twitter: What 140 characters reveal about political sentiment*. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1), 178–185. https://doi.org/10.1609/icwsm.v4i1.14009

Yang, F.-J. (2018). *An implementation of Naive Bayes classifier*. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 301–306). IEEE. https://doi.org/10.1109/CSCI46756.2018.00065