www.cya.unam.mx/index.php/cya

Contaduría y Administración 71 (3), 2026, e572

Forecasting of the variation of the Banxico reference rate using a text classification model of the monetary policy committee minutes

Pronóstico de las variaciones de la tasa de referencia del Banxico usando un clasificador de textos para las minutas del comité de política monetaria

Gilberto Anzaldo San Vicente[1*], Guillermo Benavides Perales[1], Mario Graff Guerrero[2]

[1]Banco de México, México
[2]SECIHTI-INFOTEC Centro de Investigación e Innovación en TIC, México

Received April 11, 2024; accepted February 26, 2025
Available online May 1, 2026

Abstract

This study proposes a classification model that combines Natural Language Processing (NLP) and Machine Learning techniques to analyze 96 documents published by Banxico concerning Monetary Policy decisions related to adjusting the reference rate. The document is organized as follows. The introduction highlights the critical role of the reference rate set by Banxico in ensuring the stability and functioning of the Mexican economy. It also reviews the application of NLP in financial documents over last past decade, with a focus on research related to Monetary Policy decisions. The methodology section outlines the CRISP-DM framework used to develop the classification model. It details the steps taken, including the use of Python code to expand the training dataset and the model construction process. The result section evaluates the model's performance, which is deemed satisfactory based on the predefined evaluation metrics. Finally, the conclusion discusses the model's limitations, suggests potential future research, and emphasizes the importance of this research in enhancing financial decisions-making through the use of the Mexican reference rate.

[*]Corresponding author.
E-mail address: anzaldo.gilberto@hotmail.com (G. Anzaldo San Vicente).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.
1*/ The comments and points of view in this research document are those of the authors only and do not necessarily represent the point of view of Banco de México.

https://doi.org/10.22201/fca.24488410e.2026.5548

https://doi.org/10.22201/fca.24488410e.2026.5548

JEL Code: C82, C88, E52, E58
Keywords: reference rate; text classification; monetary policy; EvoMSA; financial forecasting

Resumen

El presente estudio propone un modelo de clasificación que combina técnicas de procesamiento de lenguaje natural (PNL), y algoritmos evolutivos con máquinas de soporte vectorial que se emplean en las aplicaciones de aprendizaje de máquina, para analizar 96 documentos publicados por el Banxico relacionados con las decisiones de Política Monetaria que inciden sobre la tasa de interés de referencia. El documento consta de una introducción en la que se explica el valor crucial que tiene esta tasa en la estabilidad financiera para la economía mexicana. Posteriormente, se expone el estado del arte sobre las aplicaciones del PLN en información financiera de la última década, haciendo hincapié en las decisiones de política monetaria. Para el desarrollo del clasificador se utilizó la metodología CRISP-DM, la cual es un marco de trabajo abierto para los proyectos relacionados con ciencia de datos. Así mismo, se incluyen fragmentos del código Python para incrementar el conjunto de entrenamiento del modelo. En la sección de resultados se valora la capacidad predictiva del modelo propuesto, los cuales se consideran satisfactorios de acuerdo con las métricas empleadas en los proyectos de aprendizaje de máquina. Finalmente, se valoran los alcances, y las potenciales líneas futuras de investigación que pueden desarrollarse sobre la toma de decisiones financieras relacionadas con la tasa de referencia de México.

Código JEL: C82, C88, E52, E58
Palabras clave: tasa de referencia; clasificación de textos; política monetaria; EvoMSA; pronóstico

---

Introduction

One of the constitutional mandates of the Bank of Mexico (Banxico) is to preserve the value of the national currency, the Mexican peso[1]. To achieve this, the central bank employs one of the financial instruments at its disposal: the interbank interest rate, commonly known as the reference rate. This monetary policy instrument serves two main functions. First, it helps control short-term inflation, ensuring it remains within the target range of 2% to 4%.[2] Second, it signals the central bank's expectations regarding future inflation trends to economic and financial agents. The reference rate represents the percentage at which Banxico lends money to financial intermediaries, who, in turn, incorporate it into the rates they offer to their private and corporate clients[3].

[1] Author's Note: The constitutional mandates of Banxico are 1) stability of the financial system, 2) issuance and circulation of paper currency, 3) preservation of the value of the currency, and 4) promotion of the stability of the payment system.
[2]Author's Note: Banxico uses the reference rate to control inflation. If inflation is high, the Bank of Mexico may increase the reference rate to reduce the money supply and lower the demand for goods and services in Mexico. Conversely, if inflation is low, the reference rate will be reduced to increase the money supply, stimulating the economy by encouraging an increase in aggregate consumption.
[3] Author's Note: The reference rate is also crucial because it affects the interest rates that banks charge their customers. If the reference rate is high, banks will charge higher rates. Conversely, if the reference rate is low, banks will charge lower rates.

https://doi.org/10.22201/fca.24488410e.2026.5548

Banxico determines the reference rate based on information provided by banks operating in Mexico. Its impact is monitored daily. During the Monetary Policy Committee Meetings, decisions are made to adjust the rate in increments of 25 basis points —either upward or downward — relative to its current value, or to leave it unchanged. For the new reference rate to become official, it must be published in the Official Gazette of the Federation.

Additionally, the reference rate directly influences the structure of interest rates in Mexico. Short-term rates (e.g., 28 and 180 days) tend to exhibit more pronounced movements, whereas long-term rates (e.g., 360 or 720 days) tend to experience moderate changes. Meanwhile, interest rates for 3, 9, 10, 20, or 30 years do not always exhibit significant fluctuations. This is because short- and medium-term rates are directly affected by monetary policy decisions, whereas long-term rates are primarily influenced by macroeconomic conditions and expectations regarding the national economy's future trajectory.

State of the art

To explore research and applications of Natural Language Processing (NLP) in forecasting of the reference rate, this study reviewed relevant studies published between 2013 and 2023. This review identified six key studies summarizing 220 sources on financial text processing applications. These studies include works by Allahayari et al. (2017), Chan et al. (2017), Kearney et al. (2014), Kumar et al (2016), Mishev et al. (2020), and Schumaker et al. (2012), as detailed in Annex Table A1.

The content of the reviewed sources was analyzed, identifying 117 NLP research studies on financial text processing with various applications. Among these, 42% were most relevant to this research, focusing on forecasting the foreign exchange market (5%), stock market indices (26%), or various combinations of financial indices (11%), as detailed in Annex Table A2.

The classification of these documents employed binary labels such as {good, bad}, {bullish, bearish}, {buy, sell} for predicting financial assets behavior. For audit documents related to corporate financial reports, labels such as {"evidence of fraud", "no evidence of fraud"}, {"reliable", "not reliable"}, and others were used (Allahyari et al., 2017; Chan et al., 2017; Kearney et al., 2014; Kumar et al., 2016; Mishev et al., 2020; Schumaker et al., 2012).

In portfolio management, NLP has been used to predict financial asset behavior. When optimizing portfolios, a classifier generates labels {buy, sell, hold} to indicate the best action for each asset (Chan et al., 2017; Kumar et al., 2016; Mishev et al., 2020).

https://doi.org/10.22201/fca.24488410e.2026.5548

Notably, no studies on processing the minutes of Mexico's Monetary Policy Committees were found in the analyzed sample.[4] Instead, only one study in this field was identified, analyzing the behavior of North American financial markets after U.S. Federal Reserve (FED) monetary policy announcements (Anzaldo & Benavides, 2020).

Regardless of the computational learning model supporting NLP applications in the financial information (see Annex Figure A1), all identified applications adhere to the weak assumption of the Efficient Market Hypothesis, as proposed by Fama in 1964 (Bachelier, 1900; Bouziane et al., 2015; Chan et al., 2017; Fama, 1965, 1970; Fu et al., 2020; Kumar et al., 2016; Mishev et al., 2020; Pandya et al., 2021; Schumaker et al., 2012).

Without this assumption —which posits that financial asset prices incorporate all available market information— the use of computational models to process this information for improving financial decision-making would be meaningless.

Regarding the information sources used by researchers, studies have shown that economic news providers are the most frequently used due to their perceived reliability. Alternatively, company-provided information is also utilized (Kearney & Liu, 2014; Kumar & Ravi, 2016; Schumaker et al., 2012). In both cases, the primary challenge is accurately classifying the sentiment of analyzed texts, which serve as inputs for the computational model (Chan & Chong, 2017; Kearney & Liu, 2014).

Techniques of NLP used to process financial information

Processing the vast amount of financial information in public documents has led to the development of agnostic techniques and methodologies designed to extract the most relevant information for financial decision-making in companies (Allahyari et al., 2017; Chan & Chong, 2017; Kearney & Liu, 2014; Schumaker et al., 2012).

Since no standardized methodology exists for processing documents containing financial information, such as COPOM minutes, researchers employ similar sequences of agnostic steps. The process begins with document cleaning, which involves removing words that do not contribute to information transmission, such as repetitive words or connectors (stopwords), while preserving those within a Bag of Words (BoW). This process results in terms referred as tokens (Mishev et al., 2020).

[4]Author's Note: A forthcoming research document has been identified. This document applies NLP to 199 minutes published by the Bank of England between 1997 and 2017, examining their impact on promoting stability in three key areas of interest to researchers: 1) inflation stability, 2) productivity, and 3) the financial sector in the United Kingdom. The study combines a vector autoregressive model that links the reference rate to a matrix of unigrams constructed using the one-hot encoding method (Benavides & Colla, 2023).

https://doi.org/10.22201/fca.24488410e.2026.5548

Subsequently, one of two types of transformations is applied to the tokens to reduce their quantity and consequently minimize the number of model training parameters. The first method transforms tokens into simpler expressions by extracting their roots (stemming) or transforming them according to the economic and financial context being addressed (lemmatization). The flow of steps is presented in Annex Figure A2 (Téllez et al., 2017).

Once the relevant tokens for the NLP model are identified, a term matrix is constructed using one of the following techniques. The simplest method, known as one-hot encoding, generates a binary matrix. Another method, the TF-IDF technique, assigns a relevance value to each token in the processed text (Mishev et al., 2020).

After constructing and training the computational learning model, it undergoes statistical evaluation. Depending on the type of result, regression techniques such as Mean Absolute Percentage Error (MAPE) or Mean Squared Error (MSE) are applied. For classification models, evaluation methods such as Receiver Operating Characteristic (ROC) curves and the area under the ROC curve (AUC), among others, are used, as presented in Annex Figure A3.

Opportunity areas of NLP in financial information

Based on the literature review, 42% of financial applications focus on the analysis of capital market documents while largely overlooking the minutes of Mexico's COPOM or those of any other central bank. Moreover, the development of these techniques suggests the use of diverse methodologies (Kumar & Ravi, 2016).

Multiple studies have employed different learning models, likely due to the absence of standardized ontologies and methodologies for processing financial documents. Depending on the type of document, these techniques yield varying degrees of accuracy, prompting researchers to select the model that provides the best performance, either individually or with ensemble models (Kumar & Ravi, 2016; Chan & Chong, 2017; Kearney & Liu, 2014). This variability may stem from the fact that the analyzed texts used are often short or consist of financial news headlines, which may not provide sufficient information for the models (Chan & Chong, 2017; Kearney & Liu, 2014; Kumar & Ravi, 2016; Mishev et al., 2020).

Regarding financial information sources, the provision of a financial vocabulary for document processing individually results in a limited, poorly standardized corpus that lacks content quality. A benchmark study is deemed relevant to assess its value based on the type of application being addressed (Allahyari, 2017; Kearney & Liu, 2014; Kumar & Ravi, 2016).

https://doi.org/10.22201/fca.24488410e.2026.5548

Objective

Given the importance of the reference rate for public and private financial decision-makers in Mexico, this study aims to develop a text classifier focused on Banxico's monetary policy decisions. The classifier will categorize the reference rate using the labels {up, down, same} based on the first 96 COPOM minutes published on Banxico's website (Banco de México, 2023a).

The classification model selected for this study is Evolutionary Multilingual Sentiment Analysis (EvoMSA) version 2.0, developed by researchers at INFOTEC (Graff et al., 2020; INFOTEC-INGEO, 2023a, 2023b, 2023c). EvoMSA stands out for its ability to process textual information in multiple languages. It includes a pre-trained model encompassing over four million words. Additionally, its classification algorithm integrates support vector machines with evolutionary optimization methods. This comprehensive, open-source library facilitates the processing of Banxico's minutes, although its individual components can also be utilized separately (see Annex Figure A1).

The methodology employed in this study is the Cross-Industry Standard Process for Data Mining (CRISP-DM). This framework has been widely used in data science projects due to its structured approach, which enables the seamless execution of activities from defining business problems to deploying solutions. A key strength of CRISP-DM is its technology-agnostic nature, making it well-suited for interdisciplinary research. Its six-steps process provides a systemic approach that ensures effective execution and control of the project (Han & Kamber, 2006).

The initial step of CRISP-DM involves understanding the business problems, in this case, developing a document classifier for COPOM minutes. The next step requires an exploratory analysis the minutes to identify regularities and extract relevant information, such as arguments that justify future interest rate movements. The data preparation phase then focuses on cleaning, transforming, and structuring the text for analysis.

Subsequently, the model is developed using EvoMSA library, with training data used to define its parameters The classifier is then evaluated using the test dataset to measure its performance.

Finally, if the model achieves an accuracy of at least 80%, it will be considered suitable for deployment. This threshold is widely accepted, as it helps prevent overfitting, ensuring that the model performs well on test data. The steps of this methodology are presented in Annex Figure A4.

Methodology

Considering the steps of CRISP-DM, presented in Annex Figure A4, the first step aims to understand the characteristics of the problem, which have already been addressed in the introduction and objectives sections of this document.

Understanding the data, as suggested in step 2, began with collecting the first 96 minutes of Banxico's COPOM, which are available as PDF files. Subsequently, the time series of the first 96 values of the reference rate (TII) was retrieved from Banxico's Economic Information System (SIE).

The third step, which aims to analyze and filter the data for the proposed classification model, involved manually estimating its variations (Change_TII). These calculations were manually linked to the text of each minute (arguments and decisions) and compiled into the COPOM.JSON file, as shown in Figure 1, with its results presented in Figure 2. Subsequently, the cleaning model described in Annex Figure A2 was applied. This process resulted in a dataset consisting of more than 5 million words. After applying the cleaning process, it was determined that the 96 minutes in the COPOM.JSON file can be represented with 4 666 tokens, as shown in Table 1.

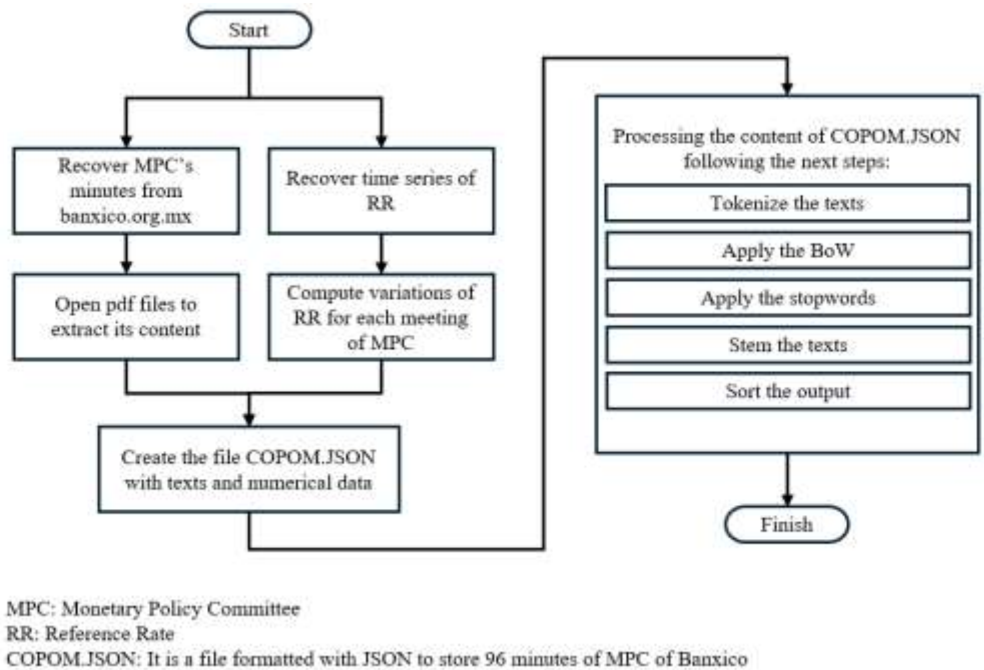


Figure 1. Text processing diagram to build the COPOM.JSON file.
Source: Author's own elaboration.

https://doi.org/10.22201/fca.24488410e.2026.5548

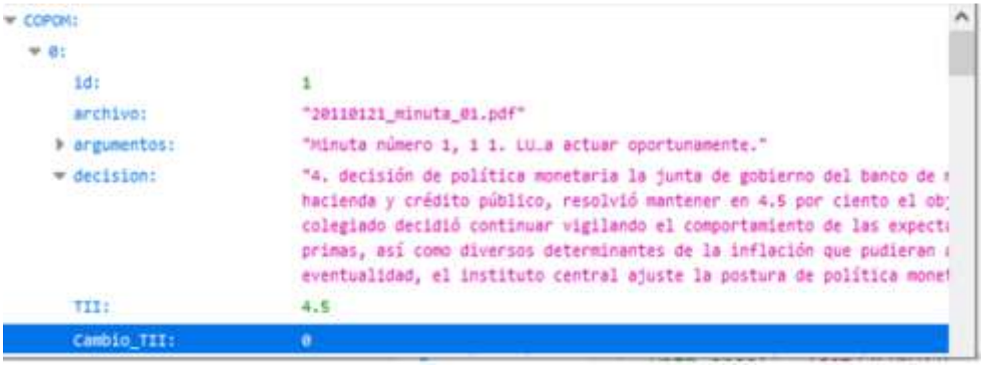


Figure 2. Contents of the COPOM.JSON file.
Source: Author's own elaboration.

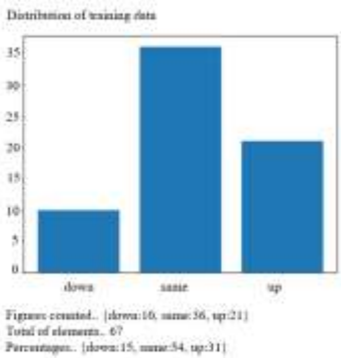
Table 1
Statistics of the screening process for the 96 COPOM minutes.

Concept of Computation	Results obtained
Number of PDF documents	96
Number of UTF-8 characters (original text)	5 446 470
Number of UTF-8 characters (clean text)	5 200 407
Number of unpurged tokens	831 961
Number of tokens purged	4 666

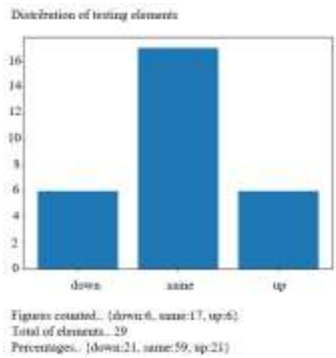
Source: Author's own estimations.

From the 96 processed documents, the training and test datasets were selected in a split of 70% and 30%, resulting in 67 and 29 sets of data, as presented in Figure 3, where it is shown that the decisions to decrease, maintain, or increase the reference rate are distributed as 17%, 55%, and 28%, respectively.

https://doi.org/10.22201/fca.24488410e.2026.5548



(A)



(B)

Figure 3. Distribution of the training (A) and test (B) sets of the original data.
Source: Author's own estimations.

Artificial augmentation of the training sample

Data processing in Data Science projects typically involves millions of observations. This amount of information allows learning models to better capture variations between them. However, in the context of monetary policy decisions, the generation of documents is limited. The available datasets contained 312 622 and 135 314 tokens for the training and test sets, respectively. This limitation motivated an increase in their size.

To artificially augment the data, the method proposed by Shorten et al. (2021) was employed. This method suggests taking an original text and randomly replacing various tokens with their synonyms. In this way, variation can be incorporated into the training sample. The new COPOM texts were generated using a dictionary of 143 synonyms, which was based on the content of Banxico's dictionary (Banco de México, 2023a) to increase the training set from 67 to 10 000, resulting in a set of 64.4 million tokens. To verify that the distribution of the new training set matched the original distribution (Figure 3-A), the texts were chosen with uniform probability. The result of this process generated a training set with the same distribution, as shown in Figure 4.

https://doi.org/10.22201/fca.24488410e.2026.5548

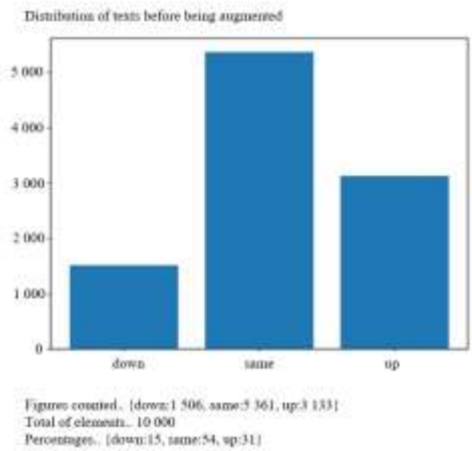


Figure 4. Distribution of the 10 000 elements, to train the classifier, artificially generated.
Source: Author's own estimations.

It is important to note that the training set was expanded solely to introduce more variations into the texts of the minutes. The testing set was not augmented, as it was specifically used to statistically validate the accuracy of the classifier[5].

The fourth step focuses on building the classification model. For this purpose, a program was developed in Python, where the training set was structured with 10 000 observations and stored in the variable str_dato, as presented in Figure 5. The following paragraphs provide a detailed explanation of the key elements in this figure, outlining the fundamental components necessary for constructing a forecasting model from a Machine Learning perspective.

A forecasting model establishes a relationship between input and output data; this is a common data structure used to organize a cause-effect relationship, where *cause* refers to relevant input data at a specific time, denoted as *T*.

The effect represents the corresponding output data. For the purpose of this research, the output specifies the response at time T+1, given that Banxico's newly announced reference interest rate takes effect on the next business day. Based on this framework, the outputs of the *evo* classifier can be interpreted as the result generated by a forecasting model from the perspective of data science.

Taking these ideas into account, the information related to cause is assigned to variable text (see line 5 in Figure 5), meaning that each of the 10 000 training data records is stored in one of the 10 000

[5] Author's Note: It should be noted that accuracies around 80% are commonly accepted in data science models. If the test set had been increased in the same proportion as the training set, accuracies exceeding 90% could have been achieved. However, this would indicate that the classifier had memorized the variations, making the trained model unreliable for new sets of texts.

https://doi.org/10.22201/fca.24488410e.2026.5548

text variables, describing the available information at time T. On the other hand, the output data represent the value of the forecast at time T+1, which is stored in 10 000 klass variables (see line 6 of Figure 5). The klass variables contain the variations of the interest rate, previously classified as {up, down, same} as described in Figure 4. Subsequently, the formatted data were transformed into a dictionary-type variable (data) and fed into the EvoMSA library using the TextRepresentations() function, with the final classifier stored in the variable evo, as shown in Figure 6.

```
1  %%time
2  s=''
3  for texto, clase in zip(X_train, y_train):
4      s = s + '{'
5      s = s + '"text" :' + '"' + texto + '"' + ","
6      s = s + '"klass":' + '"' + clase + '"'
7      s = s + '},'
8
9  s=s[0:len(s)-1]
10 str_dato='{"datos":[' + s + ']}'
Wall time: 1h 26min 25s
```

Figure 5. Construction of the inputs for the EvoMSA model using the 10 000 observations generated by separating the texts from the minutes (X_train) and the variation of the reference rate (y_train).
Source: Author's own elaboration.

```
from EvoMSA.evodag import TextRepresentations
```

```
%%time
text_repr =  TextRepresentations(lang='es')
Wall time: 28.8 s
```

```
%%time
evo=text_repr.fit(list(datos["datos"]))
Wall time: 1h 4min 59s
```

Figure 6. Construction of the classifier with the 10 000 observations generated. The values had been stored in the variable evo[6].
Source: Author's own elaboration.

Results

To assess the precision of the results of the *evo* classifier, the authors provided a file containing 29 documents that were not included in the training process; this file is named COPOM_Prueba.json. The content of this file is divided into a cause-effect structure (see Section 3.2) and coded in Python, as shown in Figure 7.

[6] Authors' Note: A simplified way to invoke the library is as follows: evo = TextRepresentations(lang= 'es').fit(X_train, y_train)

https://doi.org/10.22201/fca.24488410e.2026.5548

To separate the cause-effect variables, the testing data are stored in two objects. First, the texts of COPOM are stored in a list object named lst_pred; this object contains the information that was available at time T, prior to making a decision. On the other hand, the resulting variation of the reference interest rate is stored in the object lst_obs, representing the effect that will impact the Mexican Financial System at the time T+1. Both assignments can be seen in lines 10 and 11 of Figure 7.

Once the testing data have been prepared, the authors provided the evo classifier with the content of lst_pred, thereby supplying the classifier with the data from time T. Subsequently, the forecasting results of the variations in the reference interest rate are compared with lst_obs to assess the precision of this forecasting model.

Model results after one iteration

To determine the distribution of the results from the evo classifier, a confusion matrix was constructed to quantify the movements of the reference rate for the 29 documents provided. It was observed that the quality of the results reached 84.33% (precision_score). When quantifying the number of correct movements through completeness, a score of 83.33% (recall_score) was achieved, and when combining both metrics, the precision reached 83.02% (f1_score), averaging 83.56%. This indicates that evo achieved an acceptable level of accuracy for Machine Learning models, both in individual and combined metrics, as presented in Figure 7.

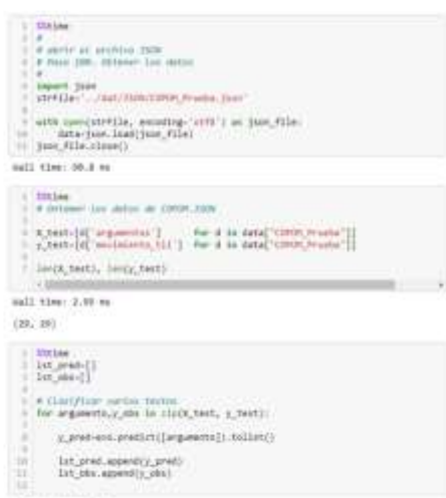


Figure 7. Assignment of test data and calculation of the evo classifier’s accuracy.
Source: Author’s own elaboration.

https://doi.org/10.22201/fca.24488410e.2026.5548

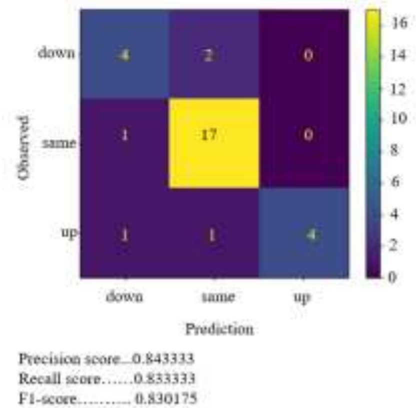


Figure 8. Confusion matrix and precision tests using the 29 test texts.
Source: Author's own estimations.

Calculation of confidence intervals

The results provided by the classifier indicate that it is reliable, as it meets the 80% accuracy criterion commonly accepted for Machine Learning models. However, to examine its behavior under variations of the test set, we resampled the set of 29 test documents 500 times (Graff, 2023). The interval was 83.02% to 84.33% precision.

Conclusions

The contribution of this research can be expressed in two main areas. From the data science perspective, the contribution lies in processing text data that summarizes the economic and financial condition in México on specific dates, denoted by time T, which influences the value of Banxico's reference interest rate on the following day, i.e., T+1. This is achieved thought the use of NLP techniques applied to the content of COPOM minutes and their processing with the evo classifier, all guided by the CRISP-DM methodology. To validate the precision of the proposed forecasting model, a subset of 29 documents was used as testing data, with the evo classifier provided only with the text of the minutes to predict the movement of the interest rate that could be applied on the following day. The resulting precision was between 83% and 84%, which is considered acceptable.

The second perspective concerns financial practice. The chosen research topic aligns with a promising trend in both academic and professional activities related to financial forecasting,

https://doi.org/10.22201/fca.24488410e.2026.5548

complementing the information considered in assessing financial decisions aimed at measuring the optimality levels of credit, market, and operational risks.

It is noteworthy that, even though 42% of the published applications are dedicated to predicting the value of various financial variables, no publications were found regarding the reference rate in Mexico. However, an upcoming document was discovered that addresses the topic with an econometric approach, specifically for the Bank of England reference rate (Benavides & Colla, 2023).

The absence of research on forecasting the reference rate published by Banxico may be due to the limited information available. In contrast to the capital markets, which generate and accumulate thousands of gigabytes (GB) or terabytes (TB) of information annually, central banks have a limited capacity to generate information. Monetary policy meetings are held between 9 and 15 times a year, and the aggregate volume of published documents is typically measured in kilobytes (KB) or megabytes (MB). As a result, it is necessary to artificially generate training data, which serves as the most effective solution[7].

Despite the lack of research on estimating the Interbank Interest Rate, this work can be considered innovative for exploring a financial topic of relevance to analysts and economic-financial decision-makers due to its significant impact on business performance within a modern economy.

Based on the experience gained during the development of this research, it is suggested that the following research avenues in sentiment analysis or the classification of financial variables would be of academic and professional interest.

- Apply sentiment analysis, or the classification of financial stability documents published by Banxico, to assess their similarity.
- Assess the immediacy with which prices assimilate new information from the environment to identify opportunities for financial arbitrage, assuming that economic agents may experience a lag in identifying this information.
- Evaluate the speed at which news is incorporated into the prices of financial assets.
- Build a glossary of economic and financial terms in Spanish to enhance the precision of models based on Natural Language Processing (NLP).

Considering the processing capabilities of EvoMSA, potential lines of research focused on processing economic and financial information could be framed within the following:

- Assess the impact of incorporating this reference rate movement classification model into sensitivity analysis methodologies of financial portfolios.

[7] Autor's Note: KB: Kilobyte (1 024 bytes), MB: Megabyte (1 024^2 bytes), GB: Gigabyte (1 024^3 bytes), and TB: Terabyte (1 024^4 bytes).

• Model the behavior of other economic and financial variables based on variations in the reference rate.

Further research can be conducted by comparing the methodology presented in this document with other standard procedures used to predict monetary policy decisions from central banks. For example, one could compare classical statistical analyses, such as the Taylor Rule, with our methodology and findings. Additionally, it may be beneficial to extend the methodology of this research to include other financial variables beyond those related to monetary policy. Examples include stock markets, commodities, and certain fiscal policy decisions.

References

Allahyari, M., Pouriyeh, S., Assefi, M., Safeai, S., Trippe, E. D., Gutierrez, J. B. & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv: 1707.02918. https://doi.org/10.48550/arXiv.1707.02919

Anzaldo, G. & Benavides, G. (2020), Expectativas en las tasas de interés y noticias de política monetaria de EEUU. Revista Mexicana de Economía y Finanzas (Nueva Época). 15(1), 17-35.

Bachelier, L. (1900). Théorie de la speculation. Annales Scientifiques de I'Ecole Normale Supérieure. 3(17), 21-86. https://doi.org/10.24033/asens.476

Banco de México. (2023a). Diccionario de Banxico Educa. 2022. Available online: https://anterior.banxico.org.mx/dyn/divulgacion/glosario/glossary.html (Accessed on 15/02/2023)

Banco de México. (2023b). Minutas del COPOM del Banxico. Available online: https://www.banxico.org.mx/publicaciones-y-prensa/minutas-de-las-decisiones-de-politica-monetaria/minutas-politica-monetaria-ta.html (Accessed on 19/02/2023)

Benavides. G., & Colla, E. (2023). El poder de las publicaciones de las minutas de política monetaria. El caso del Reino Unido. 2022. Próximo a publicarse.

Bouziane, A., Bouchiha, D., Doumi, N., & Malki, M. (2015). Question answering systems: Survey and trends. Procedia Computer Science. 73, 366-375. https://doi.org/10.1016/j.procs.2015.12.005

Chan, S. W. K., & Chong, M. W. C. (2017). Sentiment analysis in financial text. Elsevier. 94, 53-64. https://doi.org/10.1016/j.dss.2016.10.006

Fama, E. (1965). The behavior of stock prices. Journal of Business. 38(1), 34-105.Available online: https://www.jstor.org/stable/2350752 (Accessed on 13/12/2020)

Fama, E. (1970). Efficient capital markets: A view of theory and empirical work. Journal of Finance. 25(2), 383-417. https://doi.org/10.2307/2325486

Fu, B., Qiu, Y., Tang, C., Yu, H., & Sun, J. (2020). A survey on complex question answering over knowledge base: Recent advances and challenges. Available online: https://arxiv.org/abs/2502.13069 (Accessed on 05/11/2021)

Graff, M., Miranda-Jiménez, S., Téllez, E. S., Moctezuma, D. (2020). EvoMSA: A multilingual evolutionary approach for sentiment analysis. Computational Intelligence Magazine. 15(1) 76-88. https://doi.org/10.1109/mci.2019.2954668

Graff, M. (2023). Uso de la técnica Bootstrap para comparar el desempeño de modelos. Available online: https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/bootstrap.ipynb#scrollTo=0lmvZb_pzhyQ (Accessed on 15/01/2023)

Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques. M. K. Publishers, San Francisco, CA. ISBN 978-0-12-381479-1

INFOTEC-INGEO. (2023b). EvoMSA verion2.0. Documentación en Internet sobre EvoMSA versión 2.0. Available online: https://evomsa.readthedosc.io/en/docs (Accessed on 09/01/2023)

INFOTEC-INGEO. (2023c). DenseBoW. Documentación en Internet sobre la función TextRepresentation de EvoMSA versión 2.0. Available online: https://evomsa.readthedocs.io/en/docs/text_repr.html (Accessed on 09/01/2023)

INFOTEC-INGEOTEC. (2023a). EvoMSA version 2.0. Documentación en Internet sobre EvoMSA versión 2.0. Available online: https://evomsa.readthedocs.io/en/docs (Accessed on 09/09/2021)

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. Elsevier. 33, 171-184. https://doi.org/10.1016/j.irfa.2014.02.006

Kumar, S. B., & Ravi, V. (2016). A survey of the application of text mining in financial domain. Knowledge-Base Systems. 114, 128-147. https://doi.org/10.1016/j.knosys.2016.10.003

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: From lexicons to transformers. IEEE Access. 8, 131662 – 131682. https://doi.org/10.1109/ACCESS.2020.3009626

Pandya, H., & Bhatt, B. (2021). Question answering survey: Directions, challenges, datasets, evolution matrices. arXiv:2112.03572v1. https://doi.org/10.48550/arXiv.2112.03572

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. Elsevier. 53, 458-464. https://doi.org/10.1016/j.dss.2012.03.001

Shorten, C., Khosgoftaar, T.M., Furht, B. (2021). Text data augmentation for deep learning. Journal of Big Data, Springer Open, Open Access. 8(101), 1-34. https://doi.org/10.1186/s40537-021-00492-0

Tellez, E. S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O. S., & Villaseñor, E. A. (2017). A case study of spanish text transformations for twitter sentiment analysis. Elsevier. 81, 457-471. https://doi.org/10.1016/j.eswa.2017.03.071

Annex

Table A1
Types of information and data sources used (1984-2020).

Category	Type of information sources	Number of sources	Percentage	Cumulative percentage
Business news	Financial news and information	40	18%	
	Stock indices	8	4%	
	Corporate information	14	6%	
	Post servers	6	3%	31%
Social networks	Facebook	3	1%	
	Twitter	7	3%	
	News on social networks	1	<1%	5%
Financial corpus	DICTION	8	4%	
	G.I./Harvard	31	14%	
	Specific financial terms	18	8%	
	Others	4	2%	28%
Cybersecurity	Phishing	16	7%	
	Spam	11	5%	
	Malware	10	5%	
	Intrusion	7	3%	
	Fraud detection	36	16%	36%
		220	100%	100%

Source: Author's own estimations.

Table A2
Distribution of financial applications of text mining and NLP.

Category	Type of information sources	Number of sources	Percentage	Cumulative percentage
Cybersecurity	Phishing	12	10%	
	Spam	9	8%	
	Malware	10	9%	
	Intrusion	7	6%	
	Frauds	5	4%	37%
CRM	Mining customer reviews	5	4%	4%
Risk and portfolio management	Minimize portfolio variance	9	8%	
	Relationship between negative comments of companies versus their financial performance	11	9%	17%
Financial forecasts	FOREX	6	5%	
	Stock indices and share prices	30	26%	
	Combination of indicators	13	11%	42%
		117	100%	100%

Source: Author's own estimations.

https://doi.org/10.22201/fca.24488410e.2026.5548

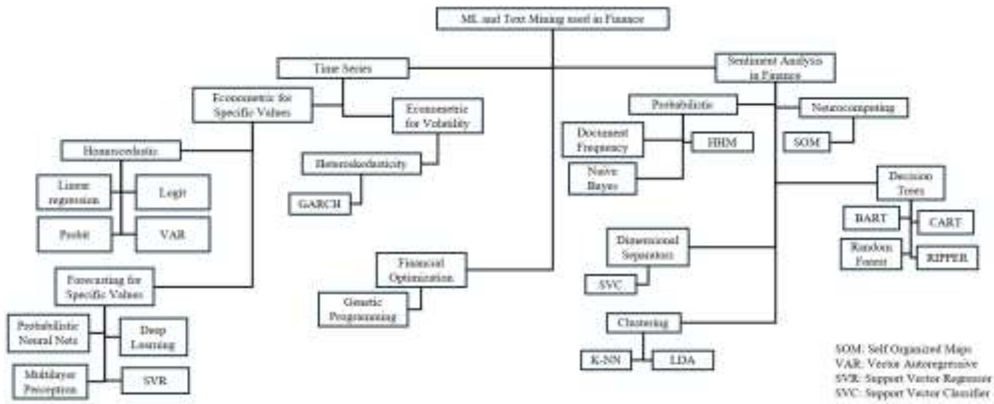


Figure A1. Classification of Computational Learning Models.
Source: Author's own estimations.

The applied methods are: Vectors Autoregressive (VAR), Hidden Markov Model (HMM), Latent Dirichlet Allocation (LDA), Support Vector Regression (SVR), Self-Organized Maps (SOM), Support Vector Classification (SVC), Bayesian Additive Regression Trees (BART), Classification and Regression Trees (CART), Repeated Incremental Pruning to Produce Error Reduction (RIPPER), K-Nearest Neighbors (K-NN), Knowledge Discovery in Databases (KDD), Decision Support Systems (DSS), Question/Answer Systems (QAS), and Knowledge-Base Question Answer Systems (KBQA) (Bouziane et al., 2015; Fu et al., 2020; Mishev, 2020; Pandya, 2021)

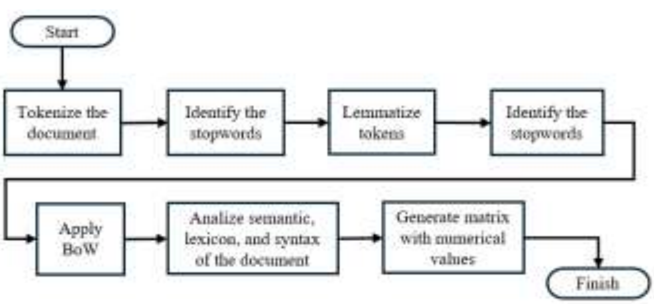


Figure A2. Linear flowchart of text screening.
Source: Author's own estimations.

https://doi.org/10.22201/fca.24488410e.2026.5548

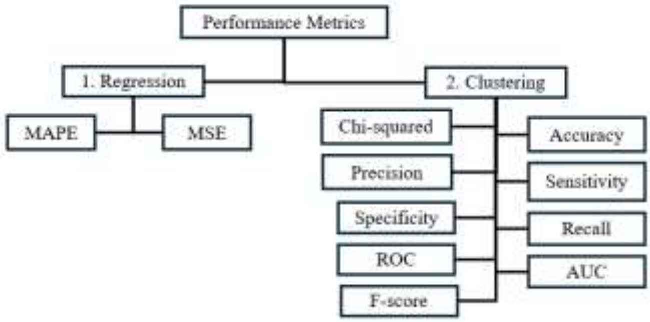


Figure A3. Performance metrics for computational learning models.
Source: Author's own estimations.

Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Receive Operating Characteristics (ROC), and Area Under Curve (AUC).

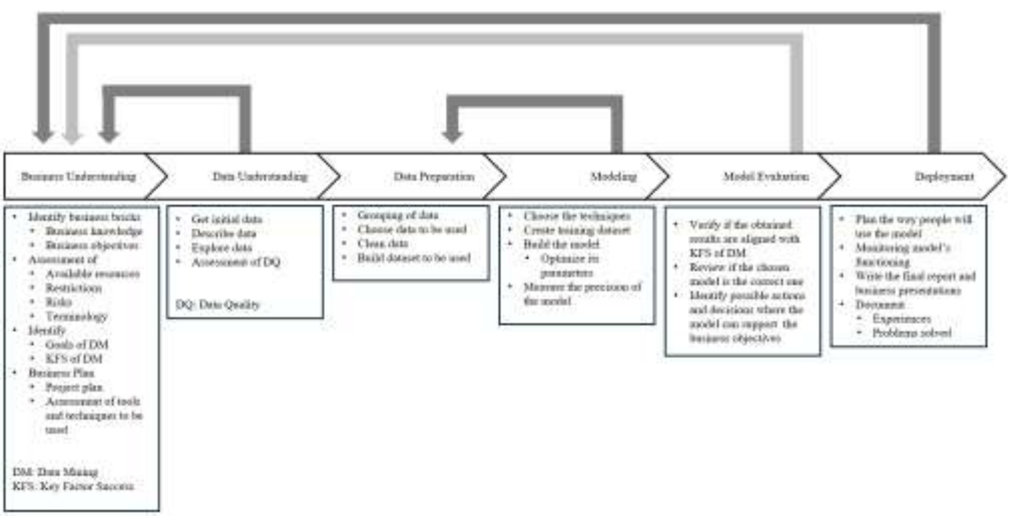


Figure A4. The six steps of the CRISP-DM methodology.
Source: Han & Kamber (2006).