

# Tracking consumer sentiment in the Dominican Republic: an approach based in granular data, a principal component analysis and a topology of neural networks

Lisette J. Santana Jiménez

Central Bank of the Dominican Republic



August 23th of 2024  
12<sup>th</sup> IFC Conference

# Outline



- I. Introduction
- II. Data and Methodology
- III. Results
- IV. Concluding Remarks



Given the impact of consumption on economic dynamics, the identification of the factors that act as guidelines in the trajectory of this variable (e.g. income level, trends, patterns, habits, prices, among others), as well as the evaluation of the balance of risks inherent to it, represent fundamental pivots for the decision-making processes by monetary policy makers...

- Variations in consumption constitute an essential component of the economic cycle, reflecting the incidence of elements such as the expectations of economic agents regarding the underlying outlook in a period of time. The period 2020-2023 was marked by various shocks (*i.e.* COVID-19 crisis, Russia-Ukraine geopolitical conflict, and general volatility in financial markets) that drastically impacted the global economy.
- From the perspective of the firms, there was a restructuring of the trade model, especially in micro, small and medium-sized companies, with the aim of ensuring their long-term sustainability (Jiménez and Santana, 2021). A modification of consumer habits and patterns was observed, mainly due to changes in the demand for certain goods and services that, prior to this period, were used to a lesser extent. This substitution effect observed during the pandemic is attributed mainly to factors such as the decrease in household income, which resulted in tighter budgetary constraints, and, on the other hand, to the awareness of a healthier lifestyle to minimize comorbid conditions that could exacerbate the likelihood of complicating COVID-19 clinical symptoms (Kirk and Rifkin, 2020).
- On the other hand, a modification of consumer habits and patterns was observed, mainly due to changes in the demand for certain goods and services that, prior to this period, were used to a lesser extent. This substitution effect observed during the pandemic is attributed mainly to factors such as the decrease in household income, which resulted in tighter budgetary constraints, and, on the other hand, to the awareness of a healthier lifestyle to minimize comorbid conditions that could exacerbate the likelihood of complicating COVID-19 clinical symptoms (Kirk and Rifkin, 2020).



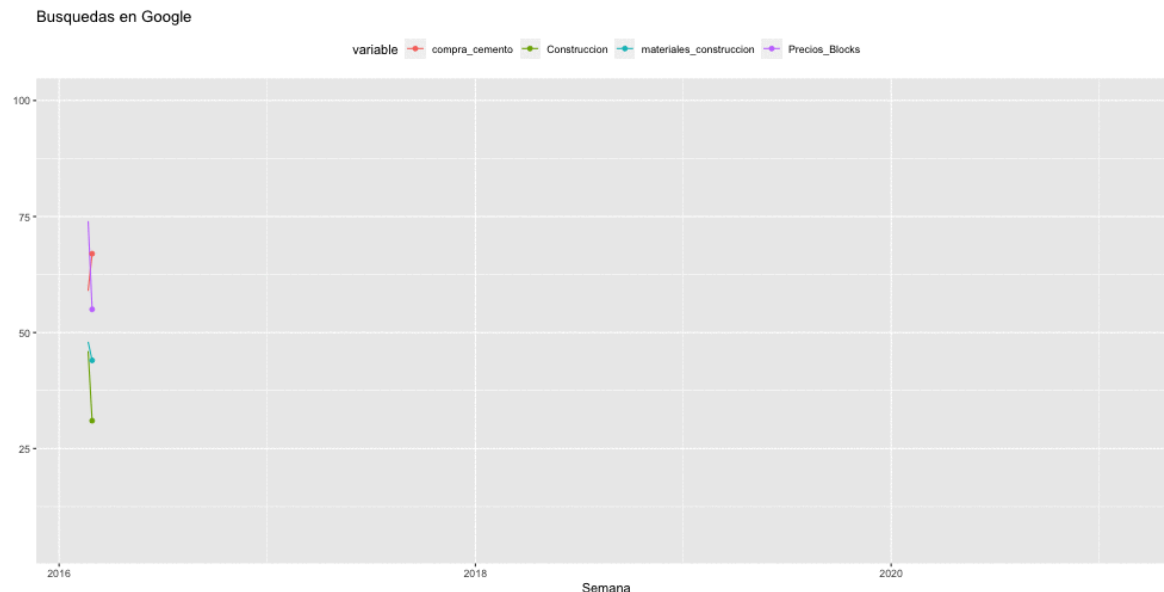
From the perspective of policy makers, one of the main reasons why consumer sentiment constitutes a fundamental input for decision-making processes is because it serves as a "thermometer" that allows capturing early signals about the health of the economy...

- Consumer sentiment indicators reliably capture, in advance, macroeconomic and financial conditions and, additionally, reflect how consumers' current financial situation affects their willingness to spend; in this case, consumer sentiment constitutes a causal force for the economy. The nano-economic data has played a preponderant role in monitoring the "footprint" of consumers or users of different goods and services, increasing the spectrum of information available so that: (i) at the macroeconomic level, central banks can manage an optimal monetary policy; (ii) in microeconomic terms, companies can adopt timely and efficient decisions for the management of their respective market niches.
- It is evident the effort that has been channeled to complement the results of surveys conducted with information from various platforms, which have led to the generation of leading indicators of a non-traditional nature, based both on qualitative and quantitative information, as well as structured and unstructured data. The objective of this research is to generate a consumer sentiment indicator for the case of the Dominican Republic concatenating quantitative and qualitative high-frequency information from different sources.

# Why are consumers' behavioral reactions relevant from the perspective of policy makers?

- The answer is simple: aggregate consumption represents the component with the greatest weighting in the behavior of economic activity and, precisely, the optimism of economic agents' conditions decisions to purchase goods and services, acquire financial assets, as well as incur debt. This set of "micro-decisions" has a direct impact on the trajectory of private consumption and, consequently, on the country's economic activity (Cote-Barón et al., 2023), to the point that consumer sentiment is considered a causal force in the economy (Kellstedt *et al.*, 2015).
- It is important to note that the scope of these decisions does not merely underlie the current economic situation of individuals, but is also subject to expectations about future income, employment conditions, prices and the evolution of interest rates. This analytical framework considered "non-traditional" has led to a significant enrichment of the empirical literature on synthetic indicators of consumer sentiment, making it feasible to identify the sources of uncertainty or the factors to which the oscillations recorded in this type of metric are attributed (Bholat, 2015; Quiñónez *et al.*, 2020). The methodologies used to quantify consumer expectations and sentiment are in a state of flux. This is due to the speed with which artificial intelligence (AI)-based techniques have progressed, creating new opportunities to develop metrics that have significantly changed the way expectations are quantified and conceptualized.

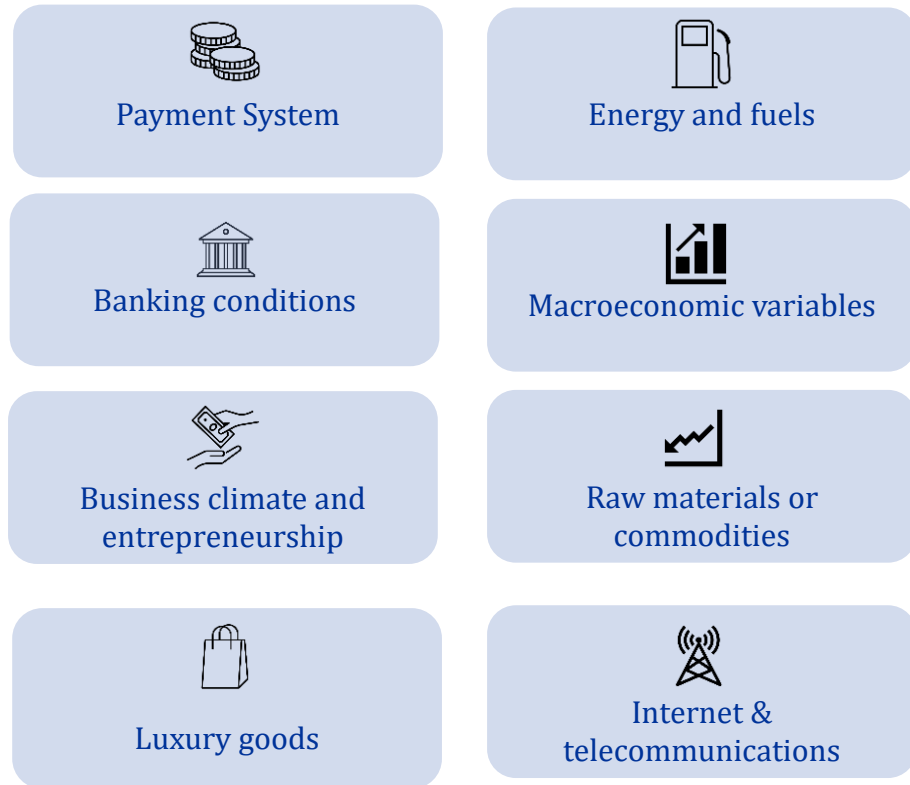
It is established that the key to the construction of synthetic indicators based on GTr data is the appropriate identification of search terms. The objective is to capture data vectors that reflect the perception of the economic agents about the underlying economic outlook...



- The Google Trends (GTr) platform is positioned as an important source of information, since it reflects the "footprint" of Google search engine users. Given that approximately 8.5 billion searches are performed daily globally, it can be inferred why the information from this platform has great potential to predict consumer behavior patterns and economic dynamics.

- A set of 133 variables is used, for the period January 2019-September 2023, with weekly periodicity, from different information sources. In the case of Google Trends, each vector is a normalized series in the interval  $[0,100]$ , whose value is subject to the relative volume or popularity, in intertemporal terms.
- For the selection process of GTr variables, the work of Santana (2019) and Della Penna and Huang (2009) is taken as a reference, using a set of unigrams and bigrams as starting point linked to the objective variable.

In order to consolidate the variables in data sets with similar characteristics, the information is grouped in eight blocks. A principal component analysis is made using these blocks of information with information from January 2019-September 2023...



- Given the massive amount of information produced on a daily basis, problems associated with the high dimensionality of data sets are recurrent. In this sense, there is a cost associated with the use of a larger amount of data, as a different treatment is required from traditional techniques, which have inferior performance when analyzing large datasets (e.g., curse of dimensionality).
- This is a point that supports the use of dimensionality reduction techniques, such as the PCA.
- A simple form to explain how the PCA works is to consider a line that captures the principal variations in a given data set, with direction and magnitude, considering the PCA as a set of orthogonal projections of data in a lower dimensional sub-space.



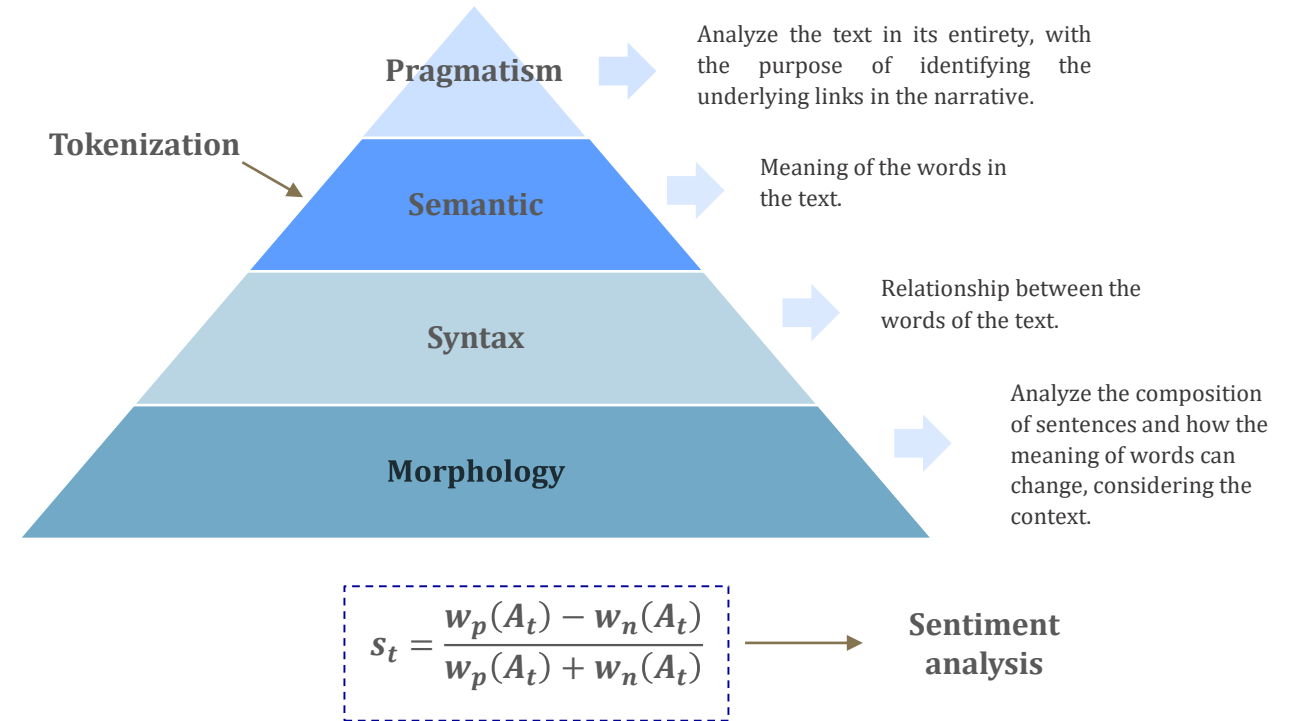
The visualization and interpretation of the existing relations among the variables in high dimensional spaces can be a complex exercise. For this reason, the techniques to reduce dimensionality, such as the PCA, allow preserving information about the explanatory variables...

- On the first stage of this exercise, a PCA is used to preserve the information about the explanatory variables and to obtain the weights of each block of data to generate the Consumer Sentiment Indicator (CSI). Given a set of information, composed by  $n$  observations and  $p$  variables, the mathematical representation of an analysis based in principal components is the following system:

$$\begin{aligned}f_1 &= a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p \\f_2 &= a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p \\&\vdots \\f_n &= a_{k1}x_1 + a_{k2}x_2 + \dots + a_{kp}x_p\end{aligned}$$

- $a_{ij}$  is the weight of the variable  $x_i$  on the component  $f_i$  and  $x_j$  is the  $j$ -th variable on the  $X$  matrix.

- On the second phase, a text mining exercise is carried out, for the analogous time period that the CSI is constructed. Briefly, the dynamics of the text mining algorithm can be described through the figure:





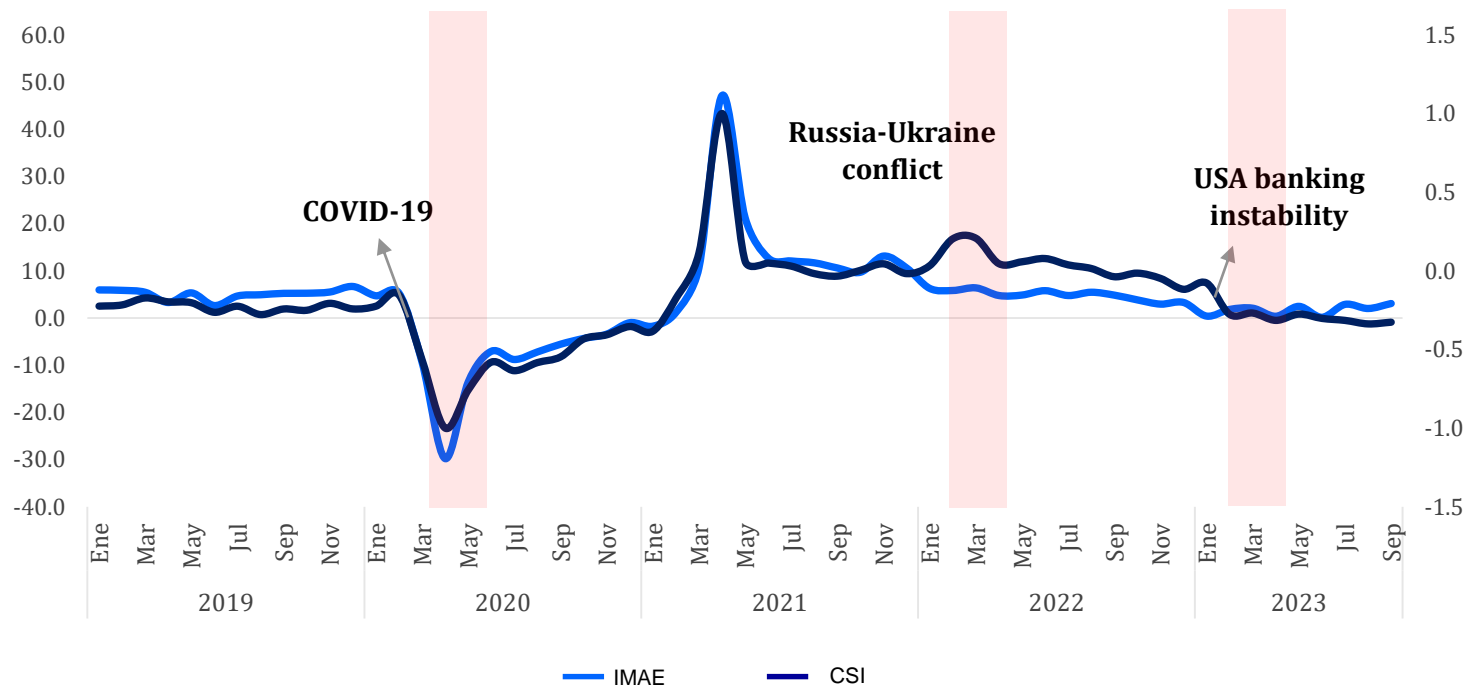
On the last phase of this research, the results obtained from the PCA concatenated with the metric that results from the text mining exercise are concatenated and mapped into the IMAE, using a neural network model...

- $Y_t$  is the metric related to economic growth (in this case IMAE),  $X_j$  is the vector of independent variables, and  $\alpha_j$  is an activation function. Once the number of hidden layers and neurons has been established, the aimed is trying to minimize the function:

$$\min_{\alpha, \theta} SSD = \sum_{t=1}^T \left[ Y_t - h \left( \sum_{k=1}^K \alpha_k f \left( \sum_{j=0}^J \theta_{jk} X_{jt} \right) \right) \right]^2$$

- The empirical literature does not present a definitive rule to carry out the optimal selection of hidden layers and neurons. One of the strategies to determine these hyperparameters is linked to the performance of the model in the test phase, observing that the network is no saturated, which would lead to overfitting (underfitting) problems. Regarding the number of hidden layers, the Cybenko theorem (1989) is used, which establishes that, with a hidden layer and a finite number of neurons, it is possible to approximate continuous functions with assumptions about the activation function.

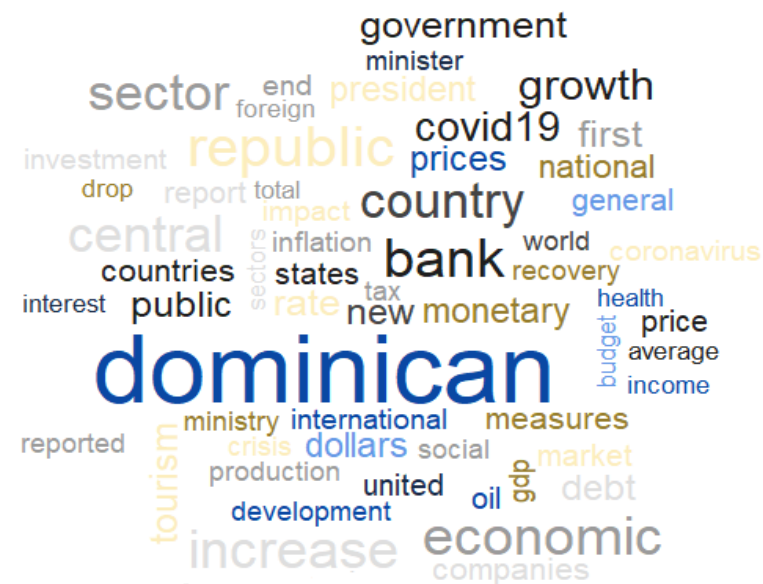
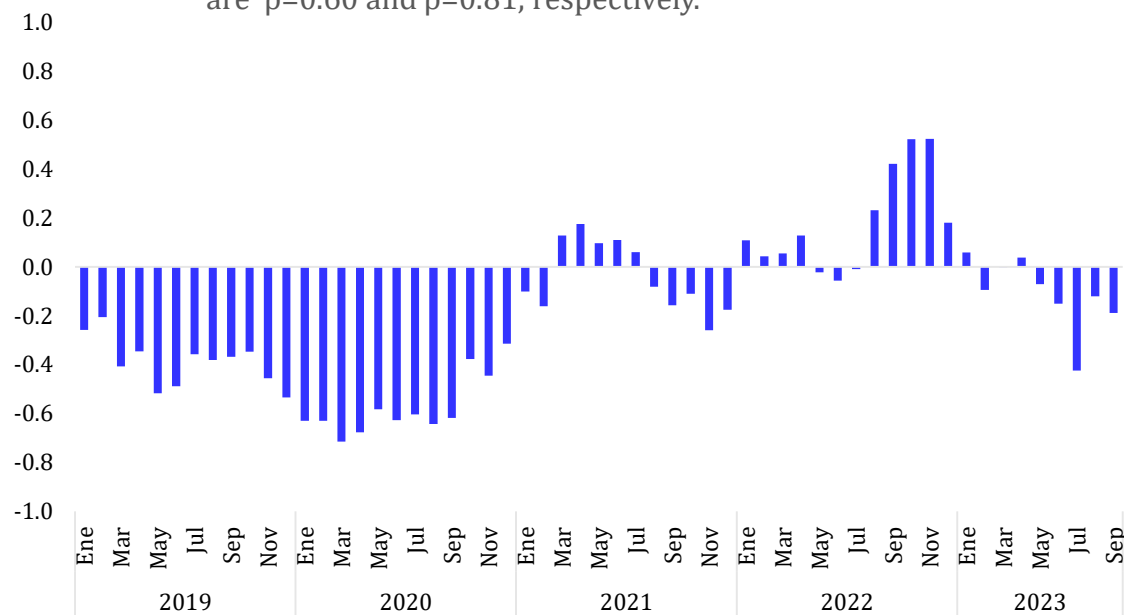
To generate the CSI, the first three loading factors (PC1, PC2 and PC3) are considered, which explain 86 % of the joint behavior of the variance of the blocks used as inputs in the framework of this analysis. A preprocessing of the inputs contemplated for the construction of the CSI is carried out, using the min-max criterion...



- The main episodes that occurred during the considered period are identified, noting that the turning points in the trajectory of the CSI coincide with the structural changes that occurred in the Dominican economy as a result of the COVID-19 crisis. The correlation coefficient between the CSI and the IMAE is  $\rho_1 = 0.77$ .
- The minimum value of the CSI is recorded in March 2020, remaining in negative territory until February 2021. It should also be noted that the period considered in this exercise was marked by the occurrence of a series of shocks (*e.g.* the Russia-Ukraine geopolitical conflict, volatility in the financial markets and instability in the U.S. banking system) that significantly affected the macroeconomic outlook.
- For the subset of the sample corresponding to the time interval 2020M01 – 2023M09, a correlation coefficient of  $\rho_2 = 0.84$  is obtained, which is higher than the value of this statistic for the complete sample. In addition to the correlation relationships, the results of the causality test are obtained; from which the robustness of the results is supported

A metric based on a text mining exercise is computed, in order to transform qualitative data from high-frequency news items in quantitative data, that contributes to the CSI. This is relevant particularly in periods of high uncertainty, anticipating changes in chaotic and non-linear scenarios...

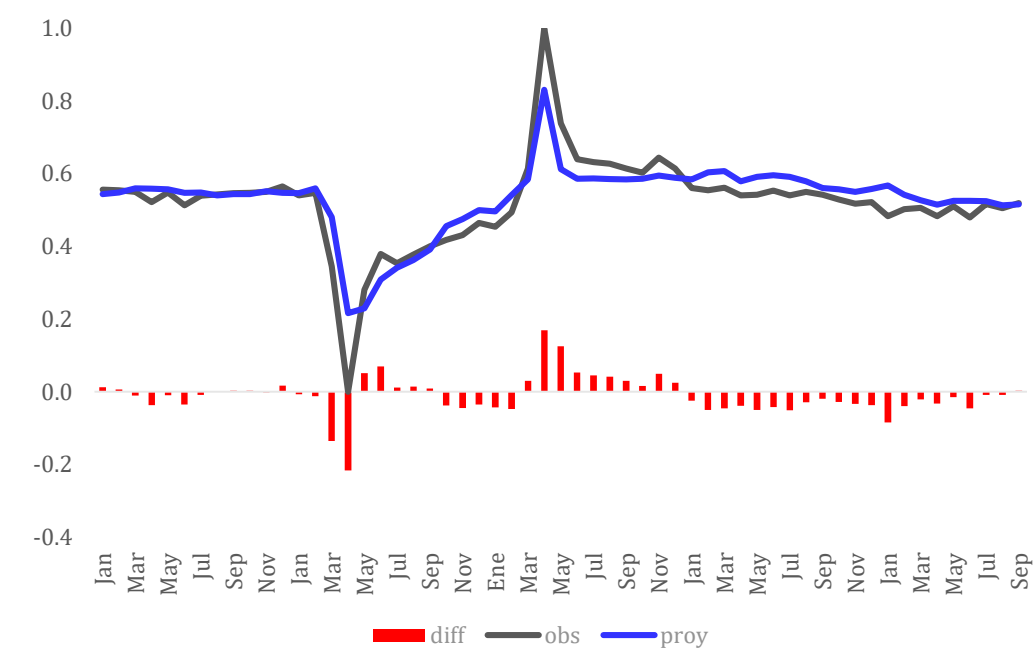
- The correlation coefficients between the IMAE and the text mining metric and between the IMAE and the CSI are  $\rho=0.60$  and  $\rho=0.81$ , respectively.



- The behavior of the text mining indicator shows consistency with the trajectory of economic activity for the period January 2019 - September 2023, both by the value of the correlation coefficient, and by the behavior of the inflection points at different time intervals.



In the MLP model, the overall RMSE is 0.1523. The functional specification depends on the CSI, CSI(-1) and the growth expectations. In this case, the CSI is already a composed metric that includes the text mining series. A series of simple OLS models is carried out, in order to evaluate the potential and contribution of each variable...



$Imae_t$	$Imae_{t-1}$	$CSI_t$	$CSI_{t-1}$	$E_{t+3}(PIB)$	$\alpha_0$	$R^2$
MICO 1	-	0.319	-0.078	0.237	0.365	0.749
		(0.000) *	(0.004) *	(0.000) *	(0.000) *	
MICO 2	0.3316	-	-	0.2895	0.1530	0.510
	(0.003) *			(0.000) *	(0.000) *	
MICO 3	-	0.406**	-	0.105	0.533	0.863
		(0.000) *		(0.032) *	(0.000) *	

\*\*Note: This CSI metric includes the text mining component.

# Concluding remarks

- It is estimated that an amount of 4.02 million exabytes of information are generated daily and approximately 90% of this data was created in the last 3 years. It is time that the central banks keep making efforts to optimize the use of this datasets considering the potential of the binominal big data-machine learning. The analysis, characterization and evaluation of the consumption patterns of economic agents has been the object of scrutiny in the areas of economics and finance, given that the magnitude of these decisions influences the management of monetary policy and business optimization processes. In this sense, the literature on this variable has expanded considerably, ranging from the construction of leading indicators, designed to assess consumer sentiment, perception or appreciation of prevailing macroeconomic conditions, to projection models or algorithms used to forecast the trajectory of this variable over a given time horizon, together with the inherent balance of risks.
- In the case of consumer sentiment indicators, it is pertinent to question why the construction of these metrics has gained so much relevance? In general terms, it can be established that, to the extent that it is feasible to quantify or approximate the uncertainty components associated with the choices of economic agents in a given period, both policy makers and companies can enrich the pool of information that supports their respective decisions. From a macroeconomic perspective, central banks can manage an optimal monetary policy, while, in microeconomic terms, firms can make timely decisions to manage their respective market niches.

# Concluding remarks

- Gathering information to identify signals about possible changes in consumer sentiment has become an increasingly meticulous and exhaustive task. This seeks to take advantage of the various data sets available, both traditional and non-traditional. In the context of the digital era, the footprints of economic agents are preserved in real time, allowing timely access to exabytes of granular and heterogeneous information from diverse sources. This facilitates the inference of patterns and trends of consumers of certain goods and services.
- The interval 2020-2022 highlights the importance of considering how structural changes, especially drastic breakdowns, underline the need to use different alternatives and exploit the potential of available information. In the midst of the Covid-19 crisis and the rapidity with which these changes occurred, timely access to diverse sources of information was essential for both central banks and companies to maintain their operations, make decisions and identify optimal solutions.
- It is also imminent to use projection models that accommodate the underlying dynamics of the crisis in the econometric modeling process. These models must be based on a strong assumption about the drastic structural change that this process implied in the time series and include the significant distortions produced after the pandemic shock (Shrebati, 2023). In the face of these problems, computational learning-based models are becoming increasingly attractive empirical strategies, demonstrating favorable performance in high-uncertainty and nonlinear scenarios, and allowing to complement the conclusions derived from traditional methods.



# BACK-UP

August 23th of 2024  
12<sup>th</sup> IFC Conference



CSI vs IMAE...In this case the exercise does not include the text mining component neither other variables that are not available for the time frame January 2019-September 2023...

