Irving Fisher Committee on
Central Bank Statistics

BIS

12th biennial IFC Conference: "Statistics and beyond: new data for decision making in central banks"

22-23 August 2024

Tracking consumer sentiment in real time in the Dominican Republic: an approach based on granular data, text mining techniques and a topology of neural networks[1]

Lisette Josefina Santana Jiménez,
Central Bank of the Dominican Republic

[1] This contribution was prepared for the conference. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Tracking consumer sentiment in real time in the Dominican Republic

An approach based on granular data, text mining techniques and a topology of neural networks

Lisette Josefina Santana Jiménez [1]

Abstract

The behavior of economic agents and their expectations regarding the future trajectory of the economy, at a multidimensional level and in a certain time horizon, has a primary role in their choices in t, serving as one of the main guidelines for the decisions adopted by policy makers. Given that the dynamics of private consumption constitutes the most important factor in the evolution of the Gross Domestic Product (GDP), the literature concerning the study of consumer sentiment is extensive, incorporating both traditional and non-traditional metrics, with the purpose of identifying turning points in consumption behavior, as well as analyzing, in a timely manner, changes in trends, patterns and preferences of economic agents. Consumer sentiment indicators are generally based on data from surveys of different organizations and economists. On the other hand, non-traditional indicators appeal to large sets of data (big data), both qualitative and quantitative. This document extends the approach of Santana (2019) and Della Penna and Huang (2009) to generate a consumer sentiment indicator (CSI), for the case of the Dominican Republic. High-frequency quantitative and qualitative information from different sources is concatenated within the framework of a principal component analysis and a text mining algorithm. The results show consistency between the behavior of the CSI and the economic activity for the period considered, identifying a correlation of $\rho = 0.77$, as well as a significant causal relationship, which provides robustness to the exercise. Finally, the effect of the CSI is mapped on the monthly indicator of economic activity, using a topology of multilayer neural networks; it is confirmed that the CSI, jointly with economic growth expectations, allows to obtain projections of the economic activity with minimal errors.

Keywords: consumer sentiment, economic activity, neural networks, text mining.

JEL classification: E21, E27, C4.

[1] Monetary Programming and Economic Studies Department, Central Bank of the Dominican Republic. For questions or suggestions write to: lj.santana@bancentral.gov.do. The opinions expressed by the author do not necessarily represent the views of the Central Bank of the Dominican Republic. Any errors are sole responsibility of the author.

I. Introduction

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). 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).

The concatenation of these factors caused significant and definitive changes in the market demand for certain goods and services. Under these considerations, it became clear that there was a need to generate metrics that would make it possible to track, in real time, consumer perception or sentiment with respect to the prevailing domestic and international panorama, as well as to track changes in the trends, habits and patterns that determine the choices made by certain segments of the population.

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. Specifically, 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 (Piger, 2003).

In this sense, the democratization of user data (nanoeconomic data) has played a preponderant role in monitoring the "footprint" of consumers or users of different goods and services, increasing the pool of information available so that, at the macroeconomic level, central banks can manage an optimal monetary policy and, in microeconomic terms, companies can adopt timely and efficient decisions for the management of their respective market niches.

Consumer sentiment indicators are generally based on data from surveys of different organizations, businessmen, economists or instruments that are applied directly to a representative sample of economic agents (e.g., University of Michigan consumer sentiment indicator, European Central Bank consumer expectations survey, Central Bank of the Dominican Republic consumer confidence indicator). Recent literature has appealed to a broad spectrum of non-traditional information (big data[2]), with the purpose of

[2] It refers to large data sets that are characterized by: (i) the speed with which they are produced, (ii) variety, (iii) veracity and (iv) volume.

constructing alternative metrics to track consumer sentiment in a timely manner and in real time. In this sense, it is evident the effort that has been channeled to complement the results of surveys conducted with information from various platforms (e.g., Google, Twitter, Instagram) (Curtin, 2019), which have led to the generation of leading indicators of a non-traditional nature, based on qualitative and quantitative information, as well as structured and unstructured.

This paper extends the approach of Santana (2019) and Della Penna and Huang (2009) to generate a consumer sentiment indicator (CSI) for the case of the Dominican Republic. Quantitative and qualitative, high-frequency information from different sources is concatenated. The treatment of this information is based on a principal component analysis, which allows us to reduce the dimension of the information matrix and, at the same time, to rescue the variables that explain, in greater proportion, the oscillations in the behavior of the indicator in question. Additionally, an exercise based on natural language processing (NLP) is carried out, specifically making use of text mining tools to quantify the underlying tone in the news coming from the Diario Libre newspaper's web portal.

The structure of this article is as follows: in the subsequent section a review of the literature associated with the construction of consumer sentiment indicators is carried out; in section III the data and the empirical approach used to achieve the objective are presented; in section IV the results obtained are discussed and, finally, the conclusions reached in the framework of this research are presented.

II. Literature Review

The concept of consumer sentiment or confidence refers to households' perception of the recent economic trajectory, together with their prospective expectations and assessment of the balance of risks of the main macroeconomic and financial variables (Kellstedt *et al.,* 2015). These types of metrics are designed with the objective of capturing signals about individuals' willingness to consume in the short term, considering the optimistic, pessimistic or neutral view on the underlying outlook (Curtin, 1983).

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.

The empirical literature concerning the construction of sentiment indicators has expanded considerably in recent years, emphasizing that the scope and information consigned in this type of metrics is consistent with the weighting given to them (Piger, 2023; Howrey, 2001). In this sense, a distinction is made between: (i) metrics elaborated from information of a traditional nature, which is compiled from various international organizations, as well as by conducting surveys of a representative sample of households and companies; (ii) metrics that have been computed by appealing to "non-traditional" data, which encompass a broad spectrum of information of a qualitative and quantitative nature, as well as structured and unstructured, which is available in real time, without lags and with the potential to minimize compilation costs (Curtin, 2022; Santana, 2019).

Under the traditional scheme, one of the most prominent and widely used indicators is the consumer sentiment index developed by the University of Michigan in the U.S. (Curtin, 2007). This index is a construct derived from a dynamic panel (with monthly rotation) of approximately 500 surveys conducted in the United States (U.S.A.); the surveys are reapplied to the same economic agents in t+6 (t refers to the month in which the first survey was conducted; $t \in \mathbb{N}$). The design of this survey makes it possible to track changes in the attitudes and behavior patterns of economic agents (both in individual and aggregate terms). Another of the advantages of this indicator is that it is computed using monthly information, which makes it available in a timelier manner than economic growth data, making it a fundamental input for making projections on U.S.A. economic dynamics.

However, in a recent publication Curtin (2022), Richard Curtin, who is currently overseeing the development of the University of Michigan's consumer sentiment index, states that 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 to predict consumer trends, habits and potential decisions. In this sense, surveys of economic agents are being complemented and, in some cases, replaced by the collection of information from different internet platforms, generated based on various approaches such as Boolean classification, keyword search, clustering algorithms, among others (Baker *et al.,* 2016).

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). In this vein, Carbonero *et al*. (2021) study the origin of the intertemporal anxiety of economic agents in financial markets, based on a synthetic indicator of sentiment (using qualitative data), with the purpose of evaluating the impact of such anxiety on investment in riskier assets, in scenarios in which the expected return in $t + n, (n \in \mathbb{R})$ interacts with multiple narratives. It is concluded that the anxiety of economic agents has a negative impact on the volatility of financial markets.

It is important to note that these "non-traditional" data sets are not without constraints for users. In this regard, while it has been documented that sentiment indicators constructed from information posted on social networks (e.g., Instagram, Twitter and Facebook) reliably capture trends, habits and consumption patterns (Becerra and Sagner, 2021; Curtin, 2022; Qi and Shabrina, 2023), there are different challenges in obtaining this data, considering that, recently, more restrictive information privacy policies have been established[3].

However, it is logical to think that, just as different authors have generated consumer sentiment metrics based on the results of opinion surveys (Brown and Cliff, 2005; Lemmon and Portniaguina, 2006; Qiu and Welch, 2006), it is also possible to construct indicators of consumer sentiment or confidence based on Internet searches. The results of different researches reflect that the attitudes and behavior patterns of economic agents, captured through internet searches, have the power to predict, in the short term, the volatility of financial markets, the return of certain assets, and the dynamics of different productive sectors (Brochado, 2020; Scott and Varian, 2016; Daas *et al.,* 2014; Vosen and Schmidt, 2011; Choi and Varian, 2012).

3 One of the most recent measures (July 2023) was the limitation on the display of information on the social network Twitter (renamed "X"), precisely to prevent data scraping.

Thus, the Google Trends (GTr) platform is positioned as an important source of information, since it captures nanoeconomic metadata, reflecting 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 (Choi and Varian, 2012 Della Penna and Huang, 2009). 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 being to capture data vectors that reflect the positive or negative perception about the underlying economic outlook in a given period.

For the case of the Dominican Republic (DR), Quiñonez and Santana (2021) study the predictive power of variables coming from the GTr platform, specifically to perform real-time forecasting exercises, in the framework of a structural Bayesian time series model (Scott and Varian, 2015) and a multilayer neural network model, for the monthly economic activity indicator (IMAE). The authors obtain out-of-sample results that evidence the favorable performance of these models, along with the inputs used, even in periods of high uncertainty. Santana (2019) constructs a consumer sentiment indicator (CSI) for the Dominican Republic, using GTr data and following the approach proposed by Della Penna and Huang (2009). A synthetic indicator is obtained, with monthly frequency, based on different blocks of information, which presents a correlation coefficient of 0.60 with respect to the IMAE and which allows us to anticipate the turning points of economic activity.

III. Data and methodology

3.1. Data

A set of 133 variables is used, for the period January 2019-September 2023, with weekly periodicity, from different information sources: (i) GTr; (ii) Central Bank of the Dominican Republic (BCRD); and (iii) Bloomberg. The GTr platform provides vectors for each of the keywords considered. Each vector is a normalized series in the interval [0,100], whose value is subject to the relative volume or popularity, in intertemporal terms; the maximum value is one hundred (100) (highest volume of searches) and zero (0) is the minimum value. 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[4] linked to the objective variable.

In order to consolidate the variables in data sets with similar characteristics, the information is grouped in eight blocks:

1. Payment system;
2. Banking conditions (bankruptcy);
3. Business climate and entrepreneurship;
4. Luxury goods;
5. Energy and fuels;
6. Macroeconomic variables;
7. Raw materials or commodities;

4 In the context of machine learning, the n-grams are sequences of n words, where $n \in \mathbb{N}$. In this case, unigrams and bigrams are considered, where n=1 and n=2, respectively.

8. Internet and telecommunications[5].

A transformation of the values of the variables corresponding to each block is carried out. In the first instance, the inter-annual variation is computed, with the aim of capturing the seasonal component of each series; subsequently, a normalization based on the min-max criterion is carried out, in order to minimize noise and homogenize the values of the series considered.

In a second stage, an indicator based on text mining algorithms is constructed, using qualitative information from the local news portal diariolibre.com. A Boolean search of terms is carried out, with the purpose of compiling a collection of economic and financial information with the potential to impact the expectations of economic agents and, consequently, their decision-making processes. A total of 1,353 news items are available, with daily frequency, for the period January 2019-September 2023.

3.2. Empirical approach

3.2.1. Principal component analysis

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 principal component analysis (PCA)[6].

In the machine learning scheme, the principal component analysis is one of the unsupervised algorithms for dimensionality reduction by first identifying the hyperplane closest to the data set and making a projection of the data set onto the hyperplane. It is important to note that, before performing a projection of the training set onto a lower dimensional hyperplane, an appropriate hyperplane must be selected, so that the variance of the original set can be preserved, in higher proportion (Géron, 2017).

In the left panel of Figure 1, a 2D dataset is visualized and in the right panel the result of the projection of each of these datasets on different axes is presented. It is evident that the projection in the first panel (solid line) preserves the maximum proportion of the variance, while the projection in the third panel maintains a minimum level of variance. It is reasonable to select the axis that preserves the largest proportion of the variance of the original set; another way to carry out this selection is to choose the axis that minimizes the average of the square of the distance between the initial data set and its projection on a given axis. This last premise is the underlying idea of the PCA. The original data set is transformed into a new set of factors $(f_n, n \in \mathbb{N})$, named principal components, that represent linear combinations of the original variables.

Another form to explain how this algorithm 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. In this case, the dimension of a data set is the number of characteristics that are used in the analysis. 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.

5 See disaggregation by blocks in appendix A1.

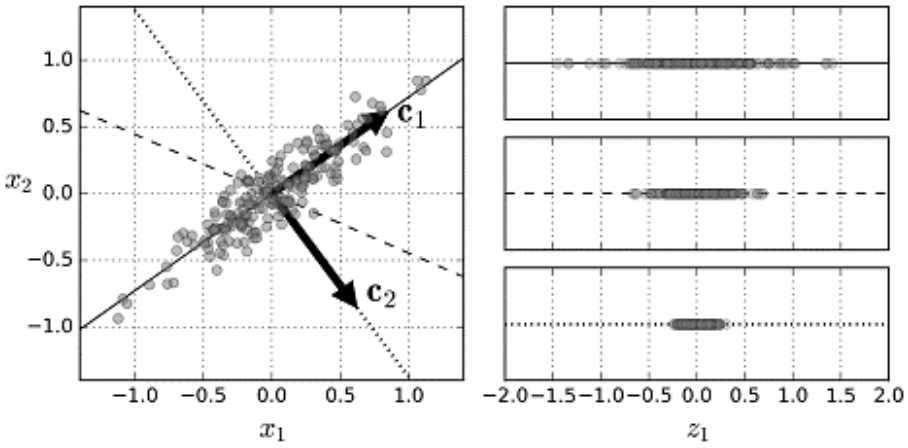
6 Other approaches used to reduce the dimensionality of a data set are manifold learning and projection (Géron, 2017).

Given a set of information, composed by n observations and p variables, represented by the matrix $X_{(n\,x\,m)}$, the mathematical representation of an analysis based in principal components is the following system:

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

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

Figure 1. Projection of a training data set in different axes.



*Source: Géron (2017).

The principal components are ordered in such a way that the first component f_1, collects the linear combinations of variables that explain, to a greater extent, the oscillations of the dependent variable[7]. Under this approach, the problem of visualizing the behavior of certain variables in intricate or complex environments is considerably simplified, making the process of identifying possible patterns or relationships between factors simpler and minimizing the noise of the data used as inputs (Di and Biswal, 2022).

3.2.2. Text mining

The text mining approach is a natural language processing technique that makes it possible to derive or generate quantitative information from document banks, using statistical, learning and computational linguistic tools (Weiss *et al.,* 2005). More explicitly, this concept encompasses a broad spectrum of computational and statistical techniques to quantify the underlying tone of relevant texts. The set of tools under this "umbrella" of text mining has become increasingly important in decision-making processes, given the information it provides for monetary and financial analysis that could not be inferred by conventional approaches.

The economic literature based on text mining algorithms has shown a significant increase, considering that, firstly, the increased availability of large sets of documents demands increasingly

[7] The significance of each component diminishes from f_1 a f_n, $n \in \mathbb{N}$.

sophisticated processing techniques and, secondly, the fact that it is feasible to extract and process qualitative and unstructured information opens up a "sea of opportunities" to address or enrich various research topics. The process of transforming qualitative to quantitative information is extensive and ranges from the pre-processing phase of the data sets to the generation of tokens and the contextualization of the corpus[8].

The process of transforming qualitative to quantitative information is extensive and ranges from the pre-processing phase of the datasets to the generation of tokens and the contextualization of the corpus depending on the type of algorithm selected. In this case, contextualization of text partitions is carried out using specialized[9] dictionaries; this approach is ideal for inferring sentiment metrics or indicators that reveal the underlying tone of a text or a document bank. It is important to note that, in the scheme of a text mining exercise, different stages are exhausted to analyze a token or partition of the corpus (Figure 3.2), in order to have a better understanding of the structure of each unit into which the text is divided.

This approach is ideal for inferring sentiment metrics or indicators that reveal the underlying tone of a text or a document bank. It is important to note that, in the scheme of a text mining exercise, different stages are exhausted to analyze a token or partition of the corpus (Figure 3.2), in order to have a better understanding of the structure of each unit into which the text is divided. The first step is to define the corpus or set of documents to be analyzed. Once the corpus has been defined, the following steps concerning the preprocessing of the text must be completed (Bholat et al, 2015):

i. Removing empty words (stopwords): words that are considered irrelevant to reveal information about the underlying tone of the documents[10] (*e.g.* articles, prepositions) are extracted from the corpus.

ii. Remove punctuation marks, numbers and blanks.

iii. Convert all words to lowercase (case folding); in this way, the same tonic is extracted from the same words, isolating the fact that they are presented in upper or lowercase letters.

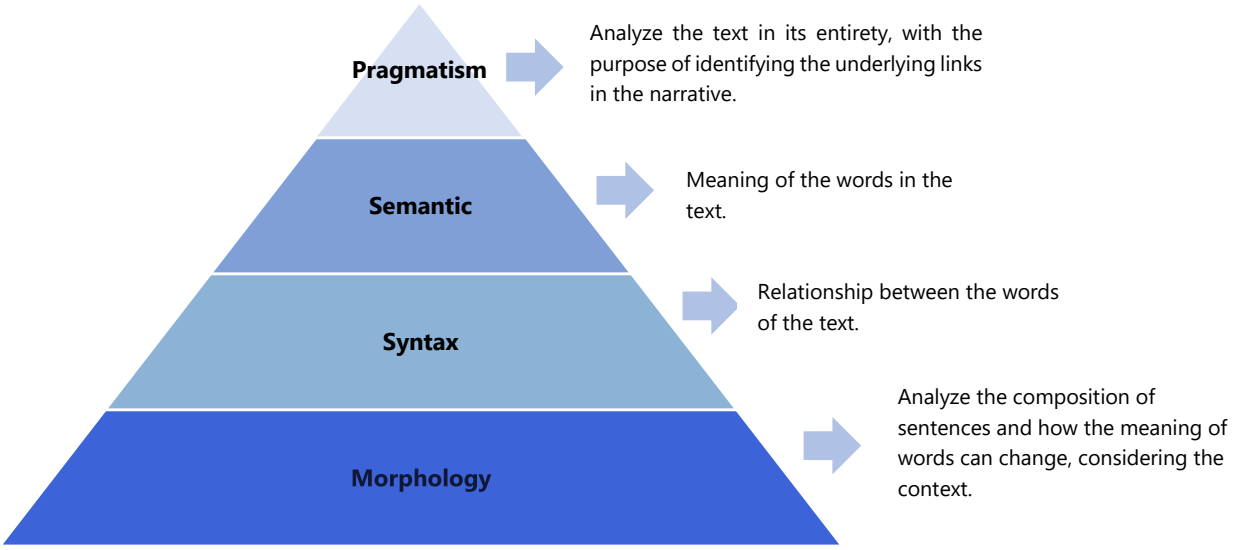
iv. To reduce the size of the corpus by stemming the words to their root.

8 In this context, the term corpus alludes a set big data set composed by unstructured/qualitative information.

9 Other alternatives for the text mining exercises are based both on supervised and non-supervised learning (e.g. Boolean classification, latent semantic allocation, latent Dirichlet allocation).

10 In this phase of the process, a gradual debugging of empty words was carried out by evaluating word histograms and using a list of empty words available at http://snowball.tartarus.org/algorithms/english/stop.txt.

Figure 2. Pyramid of the stages of analysis in text mining.



*Source: author's elaboration.

In a second stage, a tokenization or separation of the words in the text is carried out, in order to establish the frequency of each word and group them into different clusters, according to a specific metric (in this case, a Euclidean metric is used). The clustering phase is the most important problem in supervised or unsupervised learning, where an attempt is made to compose a structure from data with or without a "label". In this case, the K-Nearest Neighbor (KNN) algorithm (supervised learning) makes it possible to group the data with the greatest possible coherence between the elements of its group. In a second stage, a tokenization or separation of the words in the text is performed, in order to establish the frequency of each word and group them into different groups (clusters), according to a specific metric (in this case, a Euclidean metric is used). The clustering phase is the most important problem in supervised or unsupervised learning, where an attempt is made to compose a structure from data with or without a "label". In this case, the K-Nearest Neighbor (KNN) algorithm (supervised learning) makes it possible to group the data with the greatest possible coherence between the elements of its group.

The KNN algorithm allows to delimit neighborhoods of points, according to the established metric. In the framework of this model, each point has its largest neighborhood, which encompasses the maximum number of points with the same classification or label, called "local neighborhood". Once the points have been classified in their respective neighborhoods, the words are contextualized using Henry's Finance Dictionary and the Quantitative Discourse Analysis Dictionary (QDAP). These dictionaries contain a list of words and classify the tokens in the corpus according to a positive or negative tone.

Sentiment is computed by the following equation:

$$s_t = \frac{w_p(A_t) - w_n(A_t)}{w_p(A_t) + w_n(A_t)} \quad (2)$$

where A_t corresponds to the number of news items available at time *t*; $w_p(A_t)$ is the total number of positive news items in A_t; $w_n(A_t)$ represents the total number of negative words and s_t is the corresponding sentiment indicator; the value of s_t is normalized under the min-max criterion, with values in the interval [-1, 1], and represents the balance between the count of positive and negative words divided by the sum of positive and negative words present in A_t news items.

3.2.3. Neural Networks

Within the framework of this research, the effect of the sentiment indicator, derived in the previous stages, on economic activity (measured in terms of the IMAE and GDP) is mapped. The objective is to verify whether the indicator constructed allows us to anticipate turning points and generate accurate projections on economic dynamics. A functional specification is used as follows:

$$Y_t = \sum_{j=1}^{j} X_j \alpha_j, \tag{2}$$

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

$$mín_{\alpha k \theta(j,k)} SSD = \sum_{t=1}^{T} \left[Y_t - h\left(\sum_{k=1}^{k} \alpha_k f\left(\sum_{j=0}^{j} \theta_{ik} X_{jt} \right) \right) \right]^2 \tag{3}$$

IV. Results

In order to consolidate the variables into sets that share similar characteristics, the information is grouped into eight blocks[11]. A principal component analysis (PCA) is used to establish the weighting of each block in the generation of the consumer sentiment indicator (CSI). Under the PCA scheme, the eigenvalues λ (eigenvalues) indicate the proportion of the variance that is explained by each block, in the direction of its corresponding eigenvector x (eigenvector). Thus, the eigenvector linked to the largest eigenvalue is the factor of greatest relevance or weighting in the analysis.

In the context of this exercise, it is identified that for the block "Payment Systems" $\lambda_1 = 1.95$, being this the eigenvalue of greater magnitude, followed by "probability of bankruptcy" and "business and investment climate" with $\lambda_2 = 1.15$ and $\lambda_3 = 0.74$, respectively; in each case we have eigenvectors with positive direction (Figure 1). The cumulative ratio (Table A3.1) suggests that, as a function of these three loading factors, 86.2 % of the total joint volatility of the blocks and the IMAE is explained. Additionally, the correlation coefficients of each block with respect to the IMAE (Table A3.3) allow us to obtain more intuition about the exercise in question. In this sense, it is logical to expect that POS card transactions, as well as ATM withdrawals, capture the oscillations of economic dynamics to a greater extent, obtaining a correlation coefficient (γ=0.765) between both variables.

[11] See disagregation by blocks in the table A1.

It is important to note that the CSI is one of the input variables in the multilayer neural network model, for the purpose of projecting economic activity, measured in terms of the WEAI. In this way, it is possible to deal with the problems linked to the curse of dimensionality, reducing the number of input variables in the topology of the MLP model and avoiding a saturation of the neural network. To build the CSI, the first three loading factors (i.e. 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, through which the information blocks are normalized in the interval [-1, 1], using the min-max criterion, where -1 represents the minimum value of the indicator and 1 is the maximum value, using the equation:

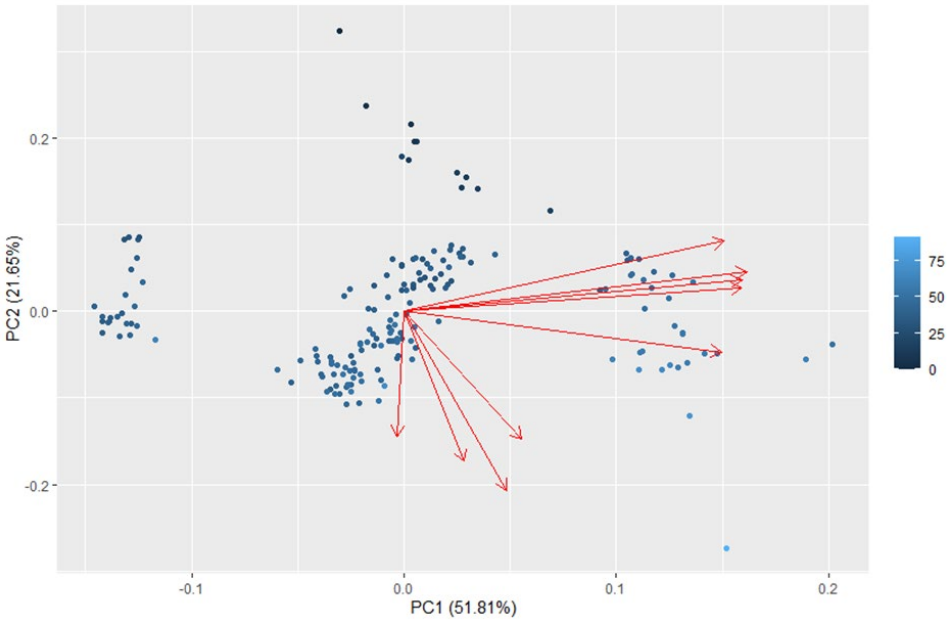
(4) $$x_i = 2 * \left\{ \frac{(x_n - x_{min})}{(x_{max} - x_{min})} \right\} - 1,$$

In equation (4), x_n is the *n-th* observation, while x_{min} and x_{max} represent the minimum and maximum value, respectively, of the series. Thus, to the extent that the CSI tends to the lower limit of the interval, there is a greater deterioration in the sentiment of economic agents. On the other hand, the periods in which the values have been located in the epsilon-neighborhood (V_ε) of the upper limit are linked to episodes of stability or a rebound in economic growth.

(4.1) $$\lim_{n \to \infty} \left| \left[2 * \left\{ \frac{(x_n - x_{min})}{(x_{max} - x_{min})} \right\} - 1 \right] - (-1) \right| < \varepsilon,$$

(4.2) $$\lim_{n \to \infty} \left| \left[2 * \left\{ \frac{(x_n - x_{min})}{(x_{max} - x_{min})} \right\} - 1 \right] - 1 \right| < \varepsilon.$$

Graph 1. Loading factors (PCA).



*Source: author's elaboration.

The results for the CSI are presented in Figure 2. The main episodes that occurred during the period January 2019 - September 2023 that generated changes in the behavior of the index 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 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, both nationally and internationally, and which resulted in a deterioration of the CSI.

In the framework of this exercise, it is useful to segment the sample to establish correlation relationships in different time intervals. For the period January 2019 - September 2023, the correlation coefficient between the CSI and the IMAE is $\rho_1 = 0.77$, visualizing a convergence between these variables. On the other hand, for the subset of the sample corresponding to the time interval January 2020 - September 2023, a correlation coefficient of $\rho_2 = 0.84$ is obtained, which is higher than the value of this statistic for the full sample. In addition to the correlation relationships, the results of the causality test are obtained; from which the robustness of the results is supported (see Appendix, Table A2).

In this sense, it is appropriate to point out that this convergence between the trajectory of the IMAE and the CSI, even in periods of high uncertainty, can be attributed to a higher level of credibility[12] in the CBRD's management by economic agents (Quiñonez *et al*., 2021). The alignment between the monetary authority's discourse and actions has a direct impact on the expectations channel, which is one of the main transmission mechanisms of the monetary policy. To the extent that the entropy of expectations is minimized, it is expected that consumer sentiment will be more consistent with the behavior of economic activity.

In addition to the blocks of variables specified for the construction of the CSI (section 2.3), a metric based on a text mining exercise is calculated (see figure 3), in a time interval analogous to the quantitative information described above. This process facilitates access to a set of documents composed of high-frequency news items, which are subsequently processed and transformed to generate an indicator that contributes to the CSI. In this context, correlation coefficients of ρ=0.60 and ρ=0.81 are observed between the IMAE and the text mining metric and between the IMAE and the CSI, respectively.

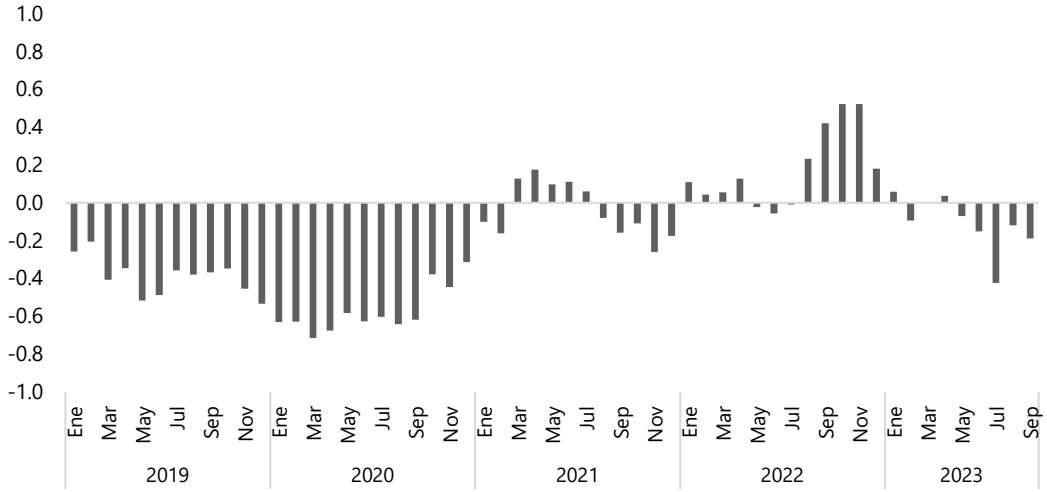
[12] The CBDR's communication strategy, after the COVID-19 crisis, became increasingly explicit and consistent with the actions adopted, which, despite the underlying circumstances, led to an increase in the credibility of economic agents in relation to monetary policy management.

Graph 2. Consumer sentiment index vs IMAE.

*Source: author's elaboration.

**Notes: 1/ Values normalized with the min-max criteria in the interval [-1, 1];
2/ Values expressed as a moving average.

Graph 3. Results of text mining exercise.

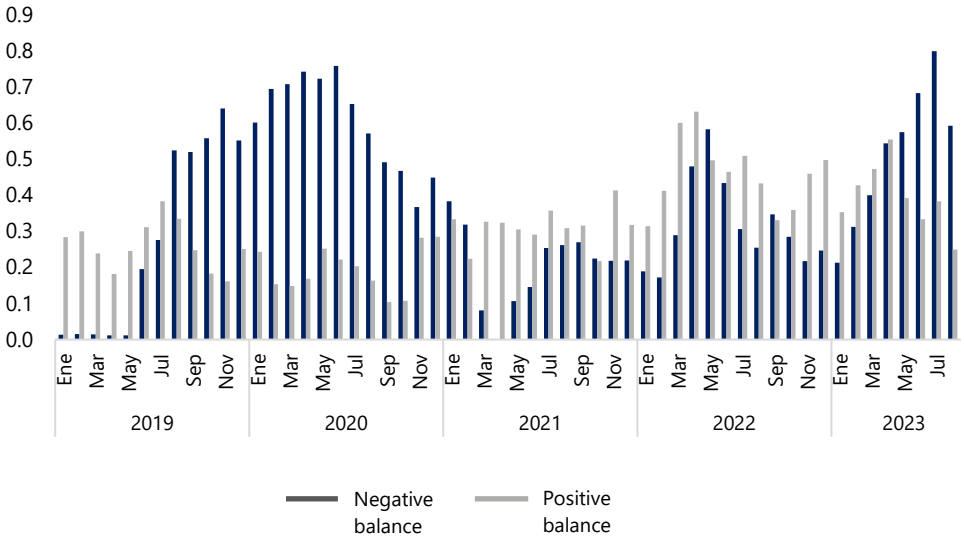


*Source: author's elaboration.

The behavior of the text mining indicator shows coherence with the trajectory of the economic activity for the period January 2019 - September 2023, both by the value of the correlation coefficient, and by the consistency visualized in the inflection points at different time segments. The information derived from these turning points is especially relevant in periods of high uncertainty, as it reflects the indicator's

ability to anticipate structural changes in chaotic and non-linear scenarios, where it is more likely to observe increases in forecast errors.

Graph 4. Balance of positive and negative words.



*Source: author's elaboration.

It is logical that, when segmenting the sample, the negative word balance was significantly higher than the positive balance (Figure 4) in the second half of 2019, when there was a deterioration in the country's tourism sector, and then in 2020 with the onset of the COVID-19 crisis, followed by other shocks (e.g. commodity price volatility, Russia-Ukraine geopolitical conflict) that significantly affected the performance of economic activity, as well as the behavior of prices.

Based on the CSI, it is feasible to carry out projections of economic dynamics in the framework of a multilayer neural network model (MLP), under the following functional specification:

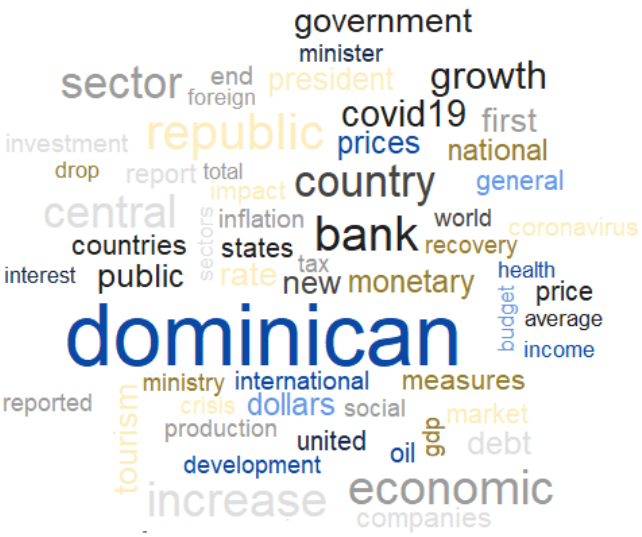
$$IMAE_t = f(CSI_t, CSI_{t-1}, E_{t+12}(GDP)), \quad (5)$$

where:

CSI_t: consumer sentiment indicator in t;
ISC_{t-1}: lag in $t-1$ of the CSI;
$E_{t-12}(GDP)$: expectations of economic growth in $t+12$.

The selected hyperparameters are a logistic activation function. Three hidden neurons are used, considering the number of input variables of the model. Additionally, a training sample covering 70 % of the total data is used. To homogenize the data, an IMAE transformation is applied under the min-max normalization criterion. The use of these metrics minimizes the noise in the information incorporated in the neural network, given the sensitivity of these structures to the variables considered as inputs in the model.

Figure 3. Word cloud of economic and financial news from the Diario Libre website (January 2019 - September 2023)



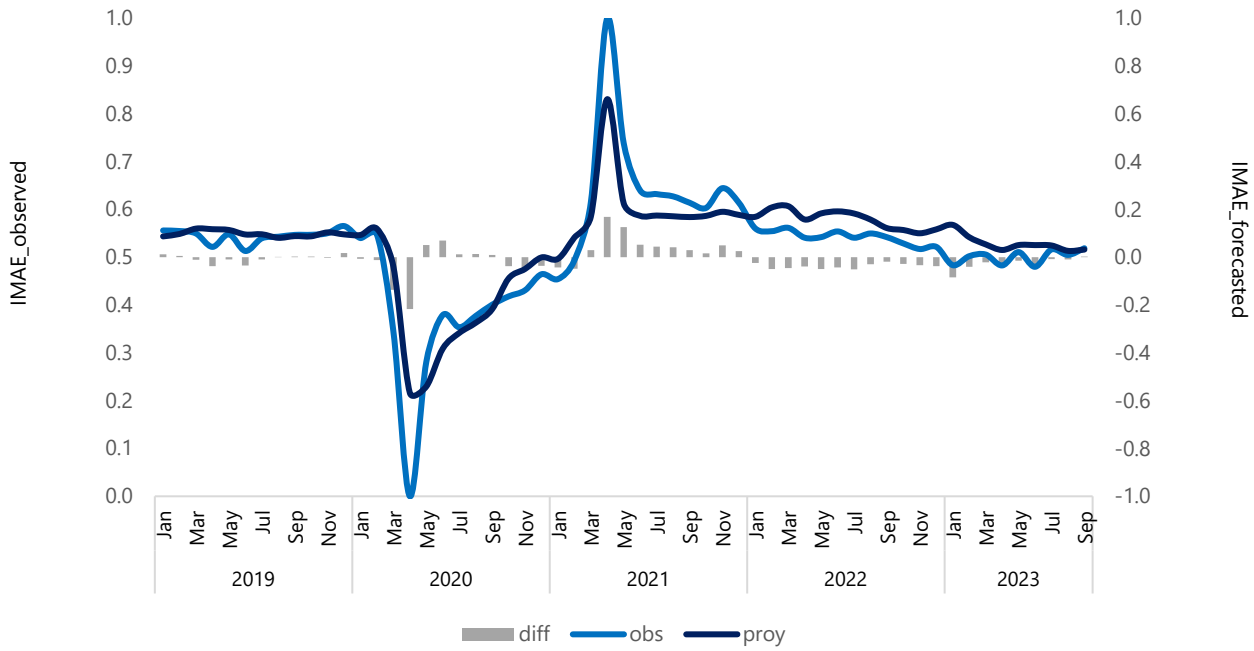
*Source: author's elaboration.

On the other hand, a simple average is computed for the CSI and for the metric based on news from the diariolibre.com website, in order to reduce the number of variables present in the architecture of this algorithm and to avoid network saturation. Additionally, one is incorporated for the CSI_{t-1}, as well as the series of economic growth expectations $E_{t+12}(GDP)$. Under this specification, an overall mean square error of 0.1523 is observed (Figure A1), showing a coherence between the observed values for the IMAE series and the projected figures in the framework of the multilayer neural network topology (Figure 5), as well as potential to predict inflection points in the trajectory of the series.

The convergence of the forecasted series towards the effectively observed values reflects the learning ability of the algorithm, considering the information in the training dataset, even during the Covid-19 crisis. This period was characterized by a chaotic and high uncertainty both in the macroeconomic and financial scenario. Under the MLP model, the projections generated presented a relatively low forecast error, evaluated in terms of the mean square error (MSE). For the years 2020 and 2021, the MSE was -0.027 and 0.041, respectively, while, for the post-pandemic period, the average MSE was -0.033.

The out-of-sample projections, carried out over a period of a time horizon h=1, allow us to corroborate the robustness of the results obtained in relation to the goodness-of-fit of the MLP model. Table 1 presents the results of these projections for the period January 2023 - October 2023, with their respective forecast errors, evaluated in terms of the ECM, obtaining a general mean square error (GMS) equivalent to 0.2158. This figure is consistent with the forecast errors of other economic dynamics projection models (Quiñonez and Santana, 2022) in which the discrepancy between observed and projected values is minimized.

Graph 5. Multi-layer neural network model results fit for IMAE (Normalized values min-max criteria 1/).



*Source: author's elaboration.
Note: 1/The values in this graph are normalized using the min-max criteria in the interval [0,1] and expressed as a moving average.

An additional validation exercise is a simple estimation of three linear regression models, using ordinary least squares, (OLS) with different functional specifications. The aim of this exercise is to capture the added value of the CSI and the text mining-based metric (Table A3.4) to project the modal path of the IMAE in the Dominican Republic. For the OLS_1 model a basic functional specification with an autoregressive AR (1) component and economic growth expectations is used. For both the OLS_2 and OLS_3 models, the CSI is incorporated, as well as a lag of this variable and the economic growth expectations; the difference between these specifications is the metric based on the text mining algorithm, which is considered in the OLS_3 model. Although the variables are significant in all the models, the best goodness of fit, measured in terms of the R^2, is found in the OLS_3 model, with an $R^2 = 0.86$; thus, the joint potential of the CSI and the text mining metric to improve the projections of the IMAE and, in general terms, to optimize the results of the short-term economic dynamics forecasting system can be observed.

These results are satisfactory, considering that one of the main challenges in the generation of forecasts (in the post-pandemic period) is to mitigate the effects of anomalies or the existence of outliers in the training data. In a scenario in which, the appropriate modeling strategies are not adopted and that an exhaustive information pre-processing is not carried out, suboptimal projections would be obtained.

In this order of ideas, it is imminent to use projection models that allow the underlying dynamics of the crisis to be accommodated in the econometric modeling process. These models must integrate a solid hypothesis on the drastic structural change that this process implied in the time series and consider the significant distortions produced after the pandemic shock (Shrebati, 2023).

Table 1. MSE evaluation of the MLP model, IMAE forecast for $h = 1$.

Period	Variation y-o-y (%)		MSE
	IMAE_Obs	IMAE_proy	
Jan-23	0.388	0.802	0.4140
Feb-23	1.798	1.182	0.6160
Mar-23	2.059	2.130	0.0710
Apr-23	0.349	0.323	0.0264
May-23	2.445	2.692	0.2470
Jun-23	0.138	0.378	0.2402
Jul-23	2.871	2.532	0.3390
Aug-23	2.001	1.967	0.0339
Sep-23	3.059	3.058	0.0012
Oct-23	3.646	3.815	0.1688

*Source: author's elaboration.

Finally, it is important to note that, under this scheme, by minimizing forecast errors, it is feasible to perform projections, over a short-term horizon, using a Multiple-Input Multiple-Output (MIMO) approach. This methodology employs as inputs the model projections in a period $t - k$, where $k \in \mathbb{R}$, as well as assumptions for the independent variables (see, for example, Mateo and Santana, 2018).

V. Concluding remarks

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.

In this paper, a CSI is constructed for the Dominican Republic, considering the period January 2019-September 2023. High-frequency quantitative and qualitative information from different sources is

combined within the framework of a principal component analysis and a text mining algorithm. In addition, the effect of the CSI on the IMAE is mapped using a multilayer neural network topology. From the results, it is found that, for the time period considered, the correlation coefficient between the CSI and the IMAE is $\rho_1 = 0.77$; it is shown convergence between these variables even in periods of high uncertainty and volatility.

For the time interval January 2020-September 2023, a correlation coefficient of ρ=0.84 is obtained. In addition to the correlation relationships, the results of the causality test are obtained; from which the robustness of the results is supported. In this sense, it is appropriate to point out that this consistency between the trajectory of the IMAE and the CSI can be attributed to greater credibility in the management of the monetary policy by economic agents (Quiñonez et al, 2021).

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. In this sense, the democratization of user data (nanoeconomic data) has been crucial to monitor and track the behavior of users of different goods and services.

The juncture of the 2020-2023 period 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. From the business perspective, there was a restructuring of the business model, particularly in micro, small and medium-sized enterprises, and a modification of consumption habits and patterns, motivated by changes in the demand for certain goods and services that previously had less use. 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 employ 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.

In this sense, the model based on a multilayer neural network topology, used to map the effect of the CSI on the behavior of the IMAE, yields satisfactory results, both in terms of the neural network fit and in terms of the magnitude of the out-of-sample forecast errors. The overall mean square error obtained is 0.2158, which is a relatively low value and it is also similar to the errors of other machine learning models, employed to project economic activity (Santana, 2017; Quiñonez and Santana, 2021), and also considering the challenges in generating forecasts in the post-COVID-19 crisis period. These challenges include anomalies or outliers in the training data. In a scenario where, appropriate empirical strategies and a thorough pre-cleaning process of the information are not adopted, the projections would be suboptimal.

References

Baker, Scott R. y Bloom, Nicholas and Davis, Steven J. (2016). "Measuring Economic Policy Uncertainty". The Quaterly Journal of Economics, vol. 131 (4), pp.1593-1636. https://doi.org/10.1093/qje/qjw024.

Becerra, J. y Sagner, A. (2020). "Twitter-Based Economic Policy Uncertainty Index for Chile," Working Papers Central Bank of Chile 883, Central Bank of Chile. https://www.bcentral.cl/documents/33528/133326/DTBC_883.pdf.

Bholat, D. (2015). "Big data and central banks," Bank of England Quarterly Bulletin, Bank of England, vol. 55(1), pp. 86-93. https://www.bankofengland.co.uk/-/media/boe/files/quarterly-bulletin/2015/big-data-and-central banks.pdf?la=en&hash=91BA29740CFE8B0334F7E874EEB57D3053DFC1C0.

Brochado, A. (2020). "Google search based sentiment indexes". IIMB Management Review, 32 (3), pp. 325–335. https://www.researchgate.net/publication/336743381_Google_Search-Based_Sentiment_Indexes.

Brown, G. y Cliff, M. (2005). "Investor sentiment and asset valuation". The Journal of Business, 2005, vol. 78, issue 2, pp. 405-440. http://dx.doi.org/10.1086/427633.

Carbonero, F., Davies, J., Ernst,E. Ghosal, S. y McSorley, L. (2021). "Anxiety, Expectations Stabilization and Intertemporal Markets: Theory, Evidence and Policy," Working Papers 2021_12, Business School - Economics, University of Glasgow. https://www.gla.ac.uk/media/Media_799873_smxx.pdf.

Choi, H. y Varian, H. (2012). "Predicting the present with Google Trends". The Economic Record, The Economic Society of Australia, vol. 88(s1), pp. 2-9. http://hdl.handle.net/10.1111/j.1475-4932.2012.00809.x.

Cote-Barón, J., Pulido-Mahecha, K., Rodríguez-Rodríguez, N., Rojas-Martínez, C. (2023). "El ISAE: Un indicador para monitorear la actividad económica colombiana en alta frecuencia". Borradores de Economía, No. 1225. https://www.banrep.gov.co/es/isae-indicador-monitorear-actividad-economica-colombiana-alta-frecuencia.

Curtin, R. (1983). "Curtin on Katona. Contemporary Economists in Perspective," Henry W. Spiegel and Warren J. Samuels, eds. New York: Jai Press, pp. 495-522.

Curtin, R. (2007). "Consumer Sentiment Surveys: Worldwide Review and Assessment". Journal of Business Cycle Measurement and Analysis, 2007, vol. 2007, issue 1, pp.7-42. https://doi.org/10.1787/jbcma-v2007-art2-en.

Curtin, R. (2019). "Consumer expectations: micro foundations and macro impact". Cambridge University Press; 1era. edición.

Curtin, R. (2022). "Measuring Economic Expectations by Machine Learning". 36th CIRET Conference (Istanbul, Turkey). https://ciret2022.conference.marmara.edu.tr/dosya/kongre/ciret.

Daas, P. y Puts, M. (2014). "Social media sentiment and consumer confidence". European Central Bank; Statistics Paper Series No.5. https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.en.pdf.

Della Penna, N. y Huang, H., (2009). "Index for U.S. Using Internet", Working Paper No. 2009-26, University of Alberta. https://sites.ualberta.ca/~econwps/2009/wp2009-26.pdf.

Di, X. y Biswal, B. (2022). "Principal components analysis reveals multiple consistent reponses to naturalistic stimuli in children and adults". Humm Brain Map. 2022 Aug 1;43(11):3332-3345. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9248318/.

Géron, A. (2017) "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow". Publisher(s): O'Reilly Media, Inc.

Howrey, P. (2001). "The Predictive Power of the Index of Consumer Sentiment," Brookings Papers on Economic Activity, Economic Studies Program, The Brookings Institution, vol. 32(1), pp. 175-216. https://www.brookings.edu/wp-content/uploads/2001/01/2001a_bpea_howrey.pdf.

Jiménez, G. y Santana, L. (2021). "Tendencias de pago bajo la crisis del Covid-19". Informe de Política Monetaria (IPOM), junio-2021, Banco Central de la República Dominicana. https://www.bancentral.gov.do/Publicaciones/Consulta?categoryId=93.

Kellstedt, P., Linn, S., y Hannah, A. (2015). "The Usefulness of Consumer Sentiment: Assessing Construct and Measurement". Public Opinion Quarterly, 79 (1), 181-203. https://corescholar.libraries.wright.edu/political_science/34.

Kirk, C. y Rifkin, L. (2020). "I'll trade you diamonds for toilet paper: consumer reacting, coping and adapting behaviors in the Covid-19 pandemic". Journal of Business Research, Vol. 117, pp. 124-131. https://www.sciencedirect.com/science/article/pii/S0148296320303271.

Lemmon, M. y Portniaguina, E. (2006). "Consumer confidence and asset prices: some empirical evidence". The Review of financial studies, vol.19, issue 4, pp.1499-1529. https://doi.org/10.1093/rfs/hhj038.

Mateo, P. y Santana, L. (2017). "Evolución de los factores determinantes del crédito bancario: evidencia para la República Dominicana basada en probabilidades de Google Trends y un Modelo Bayesiano Estructural de Series de Tiempo. Galardonado con el Tercer Lugar del Concurso de Economía de la Biblioteca Juan Pablo Duarte, Banco Central de la República Dominicana http://bcshareportalp:8085/biblioteca/documentos/pdf/ganadores/2017/tercer.pdf.

Piger, J. (2003). "Consumer Confidence Surveys Do They Boost Forecasters' Confidence? The Regional Economist, April 2003, pp. 10-11. https://www.stlouisfed.org/publications/regional-economist/april-2003/consumer-confidence-surveys-do-they-boost-forecasters-confidence.

Qi Y. y Shabrina Z. (2023) "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach". Soc Netw Anal Min. 2023;13(1):31. doi: 10.1007/s13278-023-01030-x. Epub 2023 Feb 9. PMID: 36789379; PMCID: PMC9910766.

Qiu, L. y Welch, I. (2006). "Investor Sentiment Measures". Available at SSRN: https://ssrn.com/abstract=589641 or http://dx.doi.org/10.2139/ssrn.589641.

Quiñonez, J. y Santana, L. (2023). "*Nowcasting* de la dinámica económica en la República Dominicana bajo la crisis del COVID-19: un enfoque basado en *big data* y técnicas de *machine learning*". Serie de Estudios Económicos, Banco Central de la República Dominicana. https://www.bancentral.gov.do/a/d/2582-serie-de-estudios-economicos.

Quiñónez, J., Rosa, J., Santana (2020). "Incertidumbre, gestión de la política monetaria y entropía de las expectativas en la República Dominicana: un análisis basado en algoritmos de minería de textos y redes neuronales". Documento de trabajo 2020-02, Banco Central de la República Dominicana. https://cdn.bancentral.gov.do/documents/trabajos-de-investigacion/documents/2020-02_Incertidumbre_Internacional.pdf?v=20200917?v=1722540039997.

Santana, L. (2019). "Construcción de un indicador de confianza del consumidor para la República Dominicana: una aproximación basada en probabilidades de *Google Trends*". Revista Oeconomia, Banco

Central de la República Dominicana; Vol.9 No.2. https://cdn.bancentral.gov.do/documents/trabajos-de investigacion/documents/coleccion_ensayos_vol_xiii_no2.pdf?v=1722539326691.

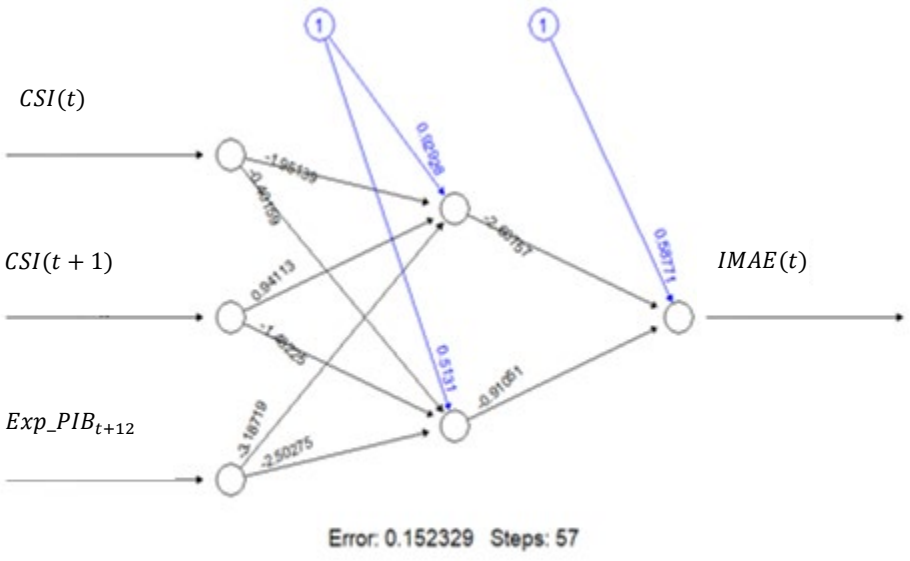
Schmidt, T. y Vosen, S. (2011). "Forecasting private consumption: survey-based indicators vs. Google trends," Journal of Forecasting, John Wiley & Sons, Ltd., vol. 30(6), pages 565-578, September.https://www.econstor.eu/bitstream/10419/29900/1/614061253.pdf.

Scott, S. y Varian, L. (2015). "Bayesian Variable Selection for Nowcasting Economic Time Series," NBER Chapters, in: Economic Analysis of the Digital Economy, pages 119-135, National Bureau of Economic Research, Inc. https://www.nber.org/system/files/chapters/c12995/c12995.pdf.

Shrebati, L. (2023). "Forecasting post COVID-19: How to improve forecasting models' performance when training data has been affected by exceptional events like COVID-19 pandemic?". KTH, School of Electrical Engineering and Computer Science (EECS). https://kth.diva-portal.org/mwg-internal/de5fs23hu73ds/progress?id=jyBFuzvbBg6aeYo5NDWIvQvl82pIISjjiMnnL6q7Xxw.

Appendix

Figure A1. Multilayer neural network model based on the consumer sentiment indicator.



*Source: MLP model.

Table A1. List of independent variables

No.	Variables	Bloque
1	POS_credit	Payment System
2	POS_debit	
3	Credit cards	
4	ATM	
5	Bankruptcy	Bankruptcy
6	Credit & Loans	
7	Finance	
8	Business Operations	
9	Bankruptcy Probability	
10	Unemployment	Business climate and entrepreneurship
11	Jobs	
12	Curriculums & Portfolios	
13	Commercial and real estate investments	
14	E-Commerce Services	
15	Business Operations	
16	Grants and financial assistance	
17	Credit facilities	
18	Household Financing	
19	Employment Agencies	
20	Automovile Financing	
21	Financial Planning	
22	Insurances	
23	Investment in commodities, futures and trading	
24	Residential and apartment leasing	
25	Investment	
26	Onapi	
27	Financial Business	
28	Financial Operations	
29	Office Services	
30	Office Supplies	
31	Risk Manegement	
32	Computers and electric equipments	
33	Luxury Goods	Luxury Goods
34	Wholesale and Department Stores	
35	Shopping Portals and Search Engines	
36	Apparel	
37	Electronics	Energy and fuel
39	Energy and utilities	
40	Electricity	

No.	Variables	Bloque
41	Oil and Gas	
42	Alternative and Renewable Energy	Energy and Fuel
43	WTI Oil Price	
44	EMBI	Macroeconomic Variables
45	Commercial lending rate	
46	Consumer Lending Rate	
47	Development Mortgage Lending Interest Rate	
48	Weighted Average Asset Interest Rate	
49	Nominal Exchange Rate Purchase Extrabank	
50	Nominal Foreign Exchange Rate Nominal Sale Extrabank	
51	Agricultural Private Sector Loans	
52	Private Sector Loans Construction	
53	Private Sector Loans Trade	
54	Private Sector Consumer Loans	
55	Private Sector Loans Electricity	
56	Private Sector Loans Hotels	
57	Private Sector Loans Industries	
58	Private sector real estate loans	
59	Private sector loans mines	
60	Other private sector loans	
61	Private sector social loans	
62	Private sector transport loans	
63	Private sector housing loans	
64	Regular gasoline	
65	Premium gasoline	
67	Coal, South African	Commodities
68	Natural gas, US	
69	Natural gas, Europe	
70	Liquefied natural gas, Japan	
71	Natural gas index	

No.	Variables	Bloque
72	Cocoa	
73	Coffee, Arabica	
74	Coffee, Robusta	
75	Tea, avg 3 auctions	
76	Tea, Colombo	Commodities
77	Tea, Kolkata	
78	Tea, Mombasa	
79	Coconut oil	
80	Groundnuts	
81	Fish meal	
82	Groundnut oil **	
83	Palm oil	
84	Palm kernel oil	
85	Soybeans	
86	Soybean oil	
87	Soybean meal	
88	Rapeseed oil	
89	Sunflower oil	
90	Barley	
91	Maize	
92	Sorghum	
93	Rice, Thai 5%	
94	Rice, Thai 25%	
95	Rice, Thai A.1	
96	Rice, Viet Namese 5%	
97	Wheat, US SRW	
98	Wheat, US HRW	
99	Banana, Europe	
100	Banana, US	
101	Orange	
102	Beef	

No.	Variables	Bloque
103	Chicken	
104	Lamb **	
105	Shrimps, Mexican	
106	Sugar, EU	
107	Sugar, US	
108	Sugar, world	
109	Tobacco, US import u.v.	Commodities
110	Logs, Cameroon	
111	Logs, Malaysian	
112	Sawnwood, Cameroon	
113	Sawnwood, Malaysian	
114	Plywood	
115	Cotton, A Index	
116	Rubber, TSR20 **	
117	Rubber, RSS3	
118	Phosphate rock	
119	DAP	
120	TSP	
121	Urea	
122	Potassium chloride **	
123	Aluminum	
124	Iron ore, cfr spot	
125	Copper	
126	Lead	
127	Tin	
128	Nickel	
129	Zinc	
130	Gold	
131	Platinum	

No.	Variables	Bloque
132	Silver	
133	Internet & Telecommunications	Internet & Telecommunications

Table A2. Granger Causality Test between the consumer sentiment indicator and the monthly economic activity index.

Nule Hypothesis:	Observations	Statistic F	Prob.
CSI does not Granger-cause IMAE	178	7.73836	0.0007

*Note: sample from january 2019 – september 2023; two lags.

Table A3. Eigenvalues, eigenvectors, correlation coeficientes.

Table A3.1. Eigenvectors (loading factors)

Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
IMAE	0.08	0.48	-0.57	0.22	0.60	-0.08	0.08	0.08	-0.03
Block 1	0.14	0.58	-0.31	-0.09	-0.73	0.09	-0.01	0.04	0.03
Block 2	0.45	-0.13	-0.01	0.05	-0.03	-0.13	-0.36	0.36	-0.71
Block 3	0.44	-0.10	-0.01	0.23	0.01	-0.09	-0.53	0.11	0.66
Block 4	0.44	-0.08	-0.05	0.14	0.05	0.50	0.03	-0.71	-0.16
Block 5	0.42	0.13	0.26	0.00	-0.03	-0.71	0.41	-0.26	0.02
Block 6	0.42	-0.23	-0.09	-0.18	0.02	0.34	0.59	0.48	0.19
Block 7	0.16	0.41	0.36	-0.71	0.31	0.15	-0.23	0.00	0.04
Block 8	-0.01	0.40	0.61	0.58	0.03	0.27	0.10	0.22	-0.03

Source: PCA results.

Table A3.2. Eigenvalues

No.	Value	Differential	Proportion	Cumulative values	Cumulative proportion
1	4.66	2.71	0.52	4.66	0.52
2	1.95	0.80	0.22	6.61	0.73
3	1.15	0.40	0.13	7.76	0.86
4	0.75	0.47	0.08	8.50	0.94
5	0.28	0.15	0.03	8.78	0.98
6	0.13	0.08	0.01	8.91	0.99
7	0.05	0.01	0.01	8.96	1.00
8	0.03	0.02	0.00	8.99	1.00
9	0.01	---	0.00	9.00	1.00

Source: PCA results.

Table A3.3. Correlation coefficients between different blocks of variables and the IMAE.

Blocks	Correlation Coefficients
Block 1	0.765
Block 2	0.186
Block 3	0.318
Block 4	0.453
Block 5	0.152
Block 6	0.369
Block 7	0.125
Block 8	0.080

Source: author's elaboration.

Table A3.4. Linear regression models for the IMAE.

Models	$Imae_t$	CSI_t	CSI_{t-1}	$E_{t+3}(PIB)$	$text$	α_0	R^2
OLS 1	**-**	**0.319** (0.000) *	**-0.078** (0.004) *	**0.237** (0.000) *	**-**	**0.365** (0.000) *	**0.749**
OLS 2	**0.3316** (0.003) *	**-**	**-**	**0.2895** (0.000) *	**-**	**0.1530** (0.000) *	**0.510**
OLS 3	**-**	**0.4064** (0.000) *	**-**	**0.105** (0.032) *	**-0.1861** (0.001) *	**0.533** (0.000) *	**0.863**

Source: author's elaboration.

Notes: (*): statistical significance $p \leq 0.05$.

2/Functional specifications:

Model OLS_1: $Imae = f\left(ISC_t, ISC_{t-1}, E(PIB)\right)$.

Model OLS_2: $Imae = f\left(Imae_{t-1}, ISC_{t-1}, E(PIB)\right)$.

Model OLS_3: $Imae = f(ISC_t, ISC_{t-1}, E(PIB), text)$.

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

BANCO CENTRAL
REPÚBLICA DOMINICANA

August 23th of 2024

12th IFC Conference

Outline

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

BANCO CENTRAL
REPÚBLICA DOMINICANA

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.

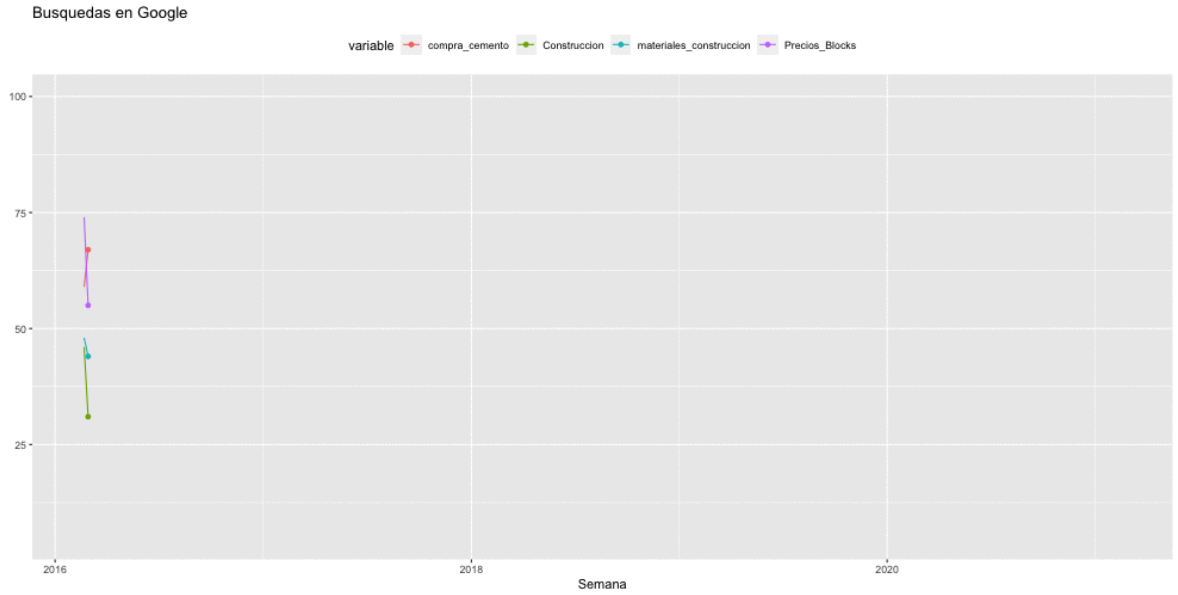
- It is evident the effort that has been channeled to complement the results of surveys conducted with information from various platforms, which have led to the generation of leading indicators of a non-traditional nature, based both on qualitative and quantitative information, as well as structured 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…

- Payment System
- Energy and fuels
- Banking conditions
- Macroeconomic variables
- Business climate and entrepreneurship
- Raw materials or commodities
- Luxury goods
- Internet & telecommunications

- 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.

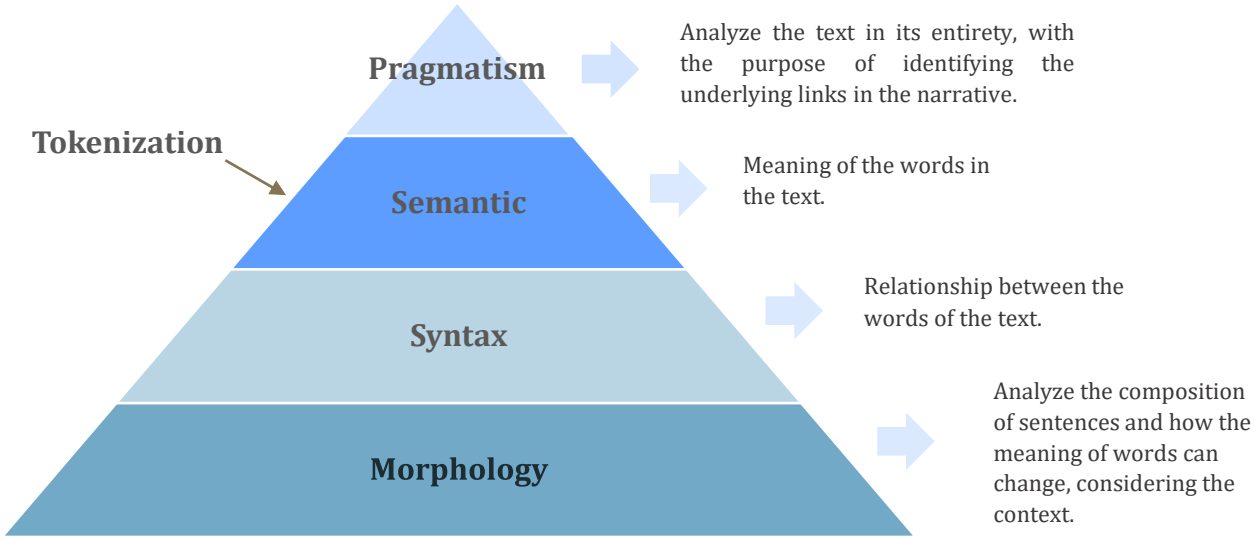
The visualization and interpretation of the existing relations among the variables in high dimensional spaces can be a complex exercise. For this reason, the techniques to reduce dimensionality, such as the PCA, allow preserving information about the explanatory variables...

- On the first stage of this exercise, a PCA is used to preserves the information about the explanatory variables and to obtained 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:

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

- 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 in constructed. Briefly, the dynamics of the text mining algorithm can be described through the figure:



$$s_t = \frac{w_p(A_t) - w_n(A_t)}{w_p(A_t) + w_n(A_t)} \longrightarrow \text{Sentiment analysis}$$

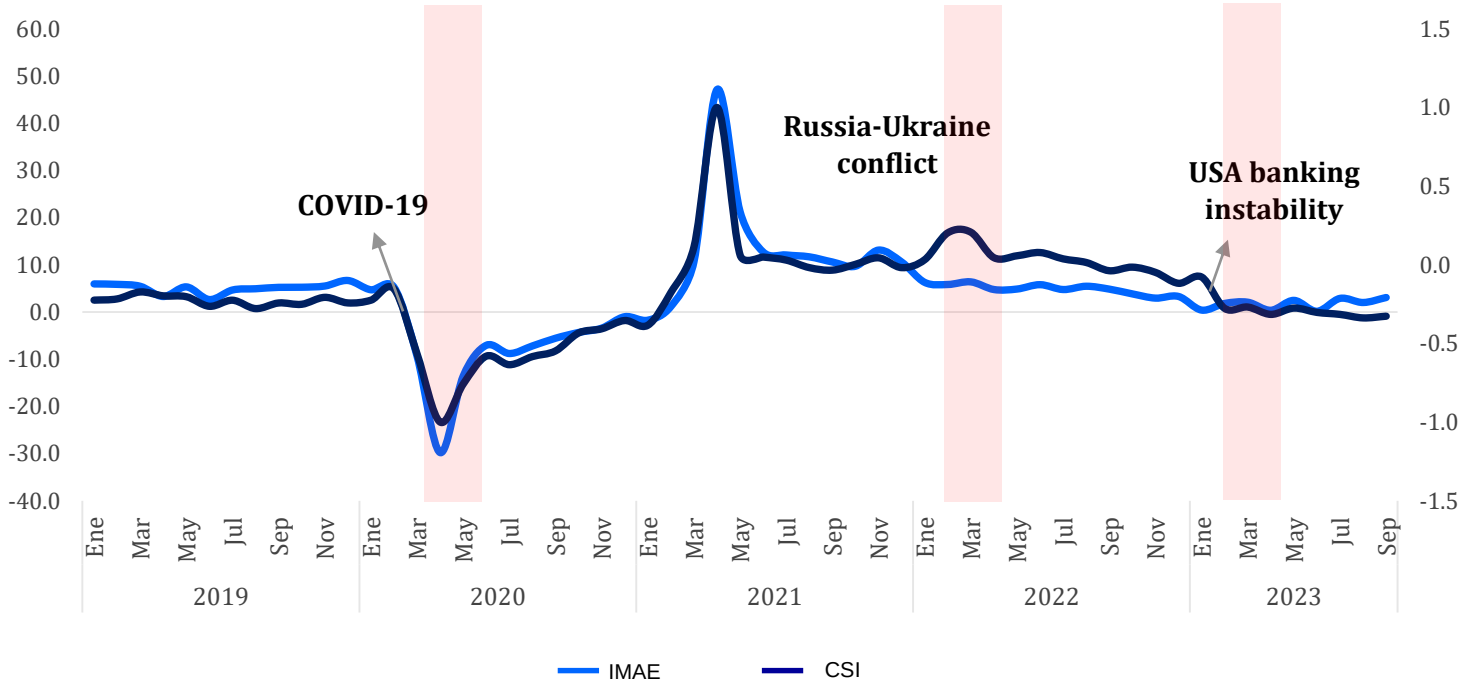
On the last phase of this research, the results obtained from the PCA concatenated with the metric that results from the text mining exercise are concatenated and mapped into the IMAE, using a neural network model…

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

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

- 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…



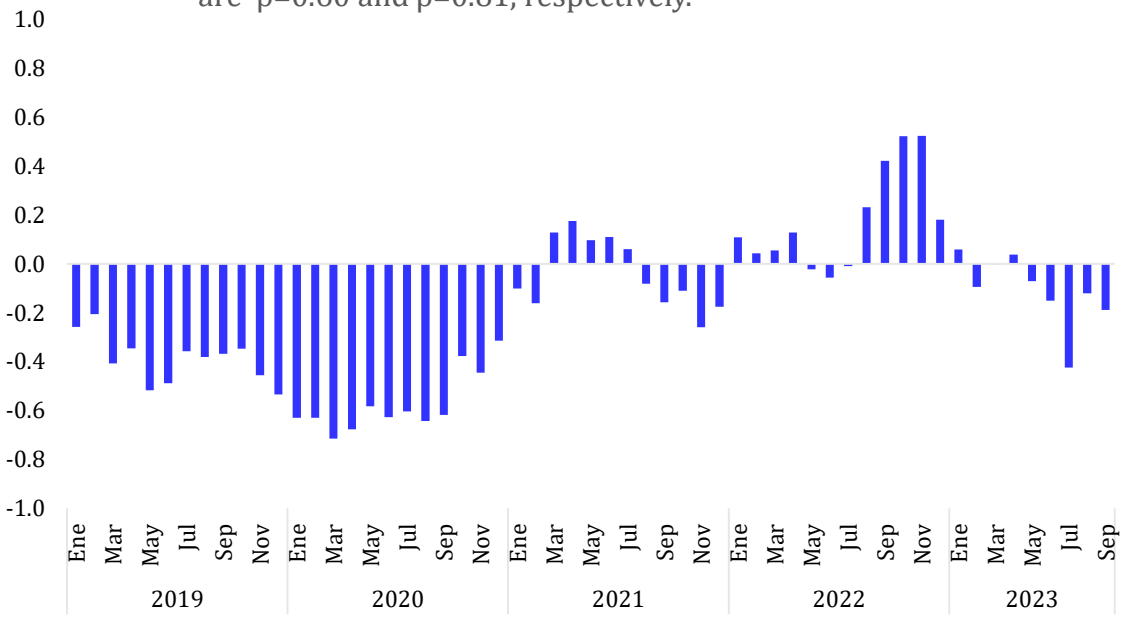
- The main episodes that occurred during the considered period are identified, noting that the turning points in the trajectory of the CSI coincide with the structural changes that occurred in the Dominican economy as a result of the COVID-19 crisis. The correlation coefficient between the CSI and the IMAE is $\rho_1 = 0.77$.

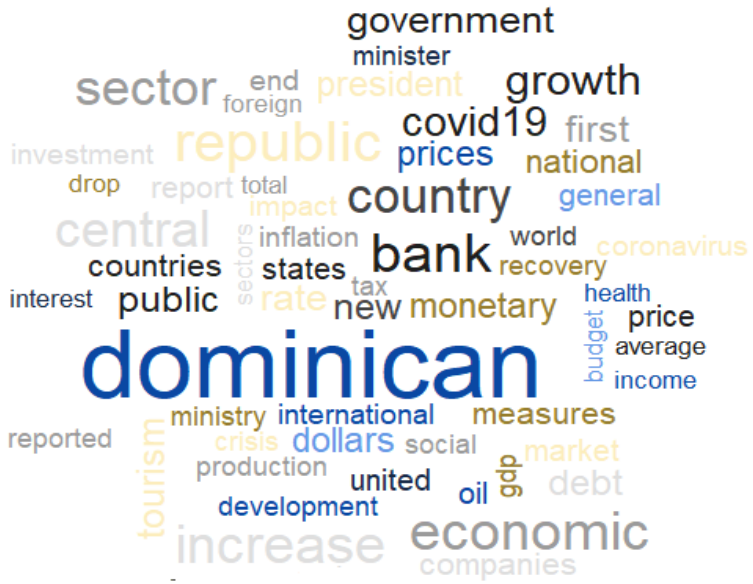
- The minimum value of the CSI is recorded in March 2020, remaining in negative territory until February 2021. It should also be noted that the period considered in this exercise was marked by the occurrence of a series of shocks (*e.g.* the Russia-Ukraine geopolitical conflict, volatility in the financial 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 ρ=0.60 and ρ=0.81, respectively.

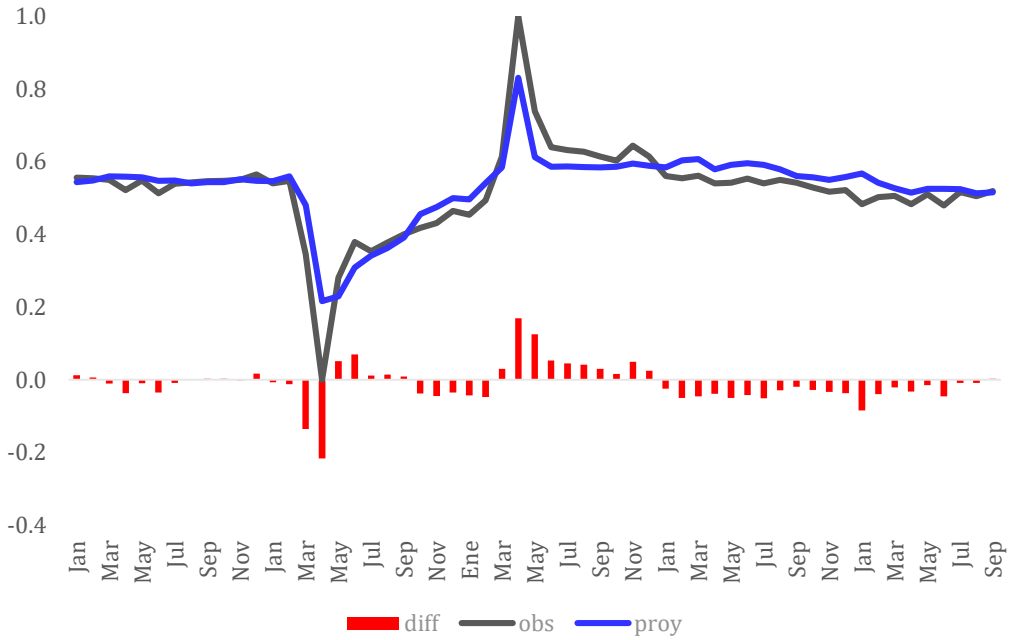




- The behavior of the text mining indicator shows consistency with the trajectory of economic activity for the period January 2019 - September 2023, both by the value of the correlation coefficient, and by the behavior of the inflection points at different time intervals.

BANCO CENTRAL
REPÚBLICA DOMINICANA

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}	$E_{t+3}(PIB)$	α_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.

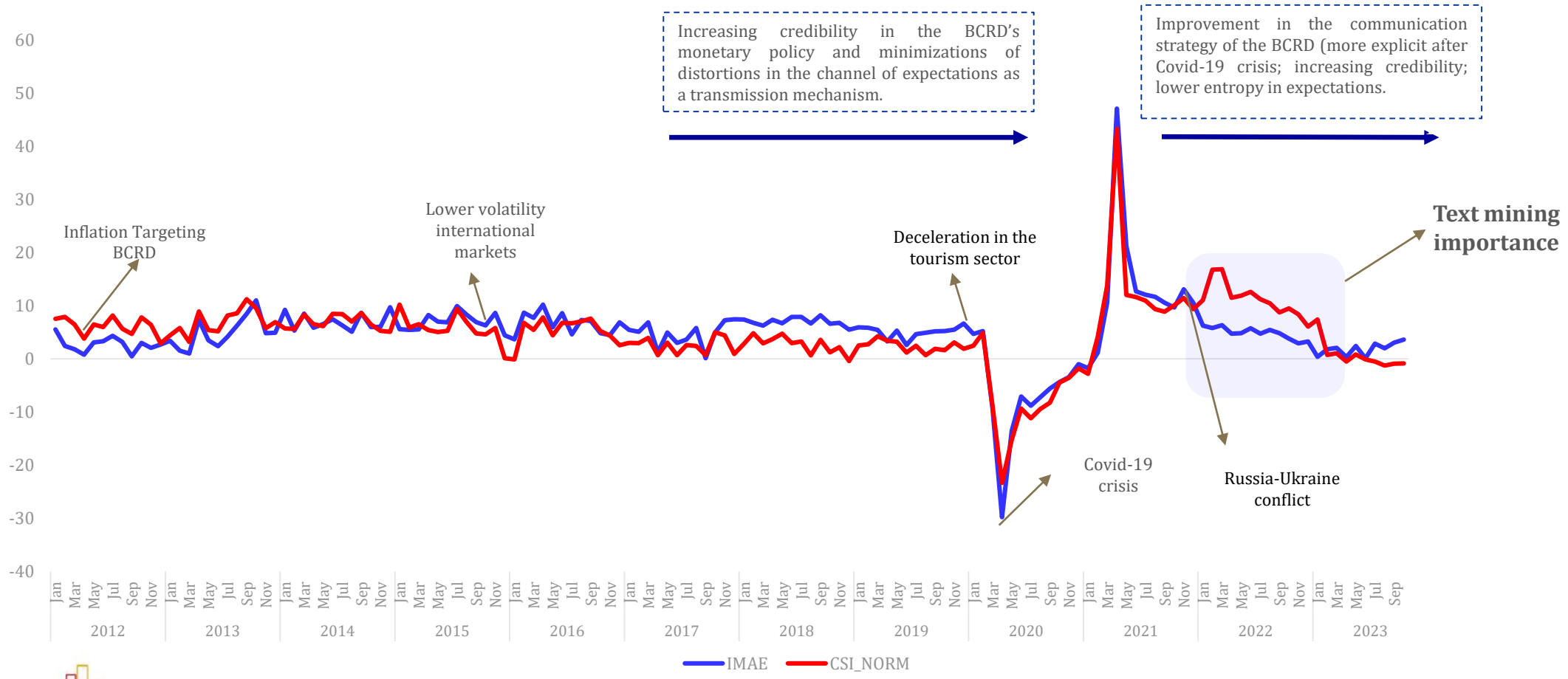
BACK-UP

August 23th of 2024

12th IFC Conference

BANCO CENTRAL
REPÚBLICA DOMINICANA

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…



BANCO CENTRAL
REPÚBLICA DOMINICANA