CIDE

LATIN AMERICAN ECONOMIC REVIEW

Volume 33 | Article 10 | 2024

Tweeting Inflation: Real-Time Measures of Inflation Perception in Colombia *

Jonathan Alexander Muñoz-Martínez† David Orozco‡
Mario A. Ramos-Veloza§

Submission received: October 15, 2023
Final version received: October 2, 2024
Accepted: October 2, 2024
Published: February 6, 2025

Abstract

This study follows recent developments in the literature using Twitter, now known as X, to measure inflation perceptions. By applying machine learning techniques, we implement two real-time indicators of inflation perception for Colombia and show that both exhibit a dynamic similar to that of inflation and inflation expectations between January 2015 and March 2023. This suggests that our indicators are closely related to the underlying factors that drive inflation. Additionally, we find that our indicators improve the forecast accuracy of inflation and inflation expectations up to six months. Overall, this approach provides a valuable instrument for gauging public sentiment towards inflation, and complements the traditional indicators used in the inflation-targeting framework.

Keywords: Inflation perceptions, Twitter, Real-time data, Central banks.
JEL codes: E31, E37, E52.

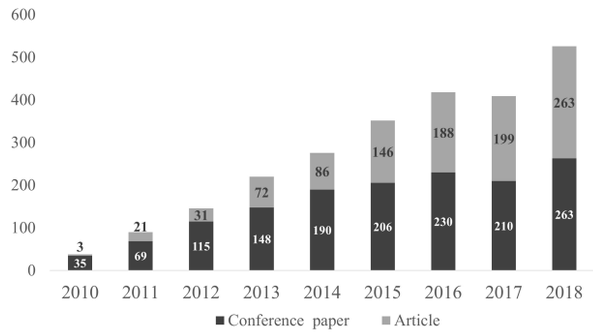
*The opinions contained in this document are the sole responsibility of the authors, and do not commit Banco de la República or its Board of Directors.

†Banco de la República, Colombia. Mail: jmunozma@banrep.gov.co
‡Universidad Icesi. Mail: dmorozco@icesi.edu.co
§Banco de la República, Colombia. Mail: mramosve@banrep.gov.co.

1. Introduction

Inflation expectations are a critical component of the inflation targeting framework adopted by many central banks. Policymakers regularly monitor multiple measures of inflation expectations derived from economic surveys and financial market instruments to understand short-run inflationary pressures. Angelico et al. (2022); Bricongne et al. (2022); Born et al. (2023) proposed that information on Twitter can be obtained in real time to gauge collective perceptions of inflation. Moreover, this information can be carefully analyzed, selected, and used to construct new indicators, thereby enhancing the measures available to policymakers. Twitter, which was rebranded as X since July 2023, may provide valuable real-time and detailed information in the policymaking process, as discussed by Vydra and Kantorowicz (2021).[1] Furthermore, text analytics and big data have revolutionized the construction of indicators using different social media. For instance, Antenucci et al. (2014) use Twitter to construct labor market indices for job loss, job search, and job posting. Bailliu et al. (2019) analyze labor market conditions using text analytics applied to Mainland Chinese-language newspaper. Glaeser et al. (2019) uses Yelp to nowcast the local economy. Taking advantage of these characteristics, Becerra and Sagner (2020) constructed an Economic Policy Uncertainty (EPU) index using tweets from media and Indaco (2020) show that the volume of tweets is a valid proxy for estimating GDP. More evidence of the increasing usefulness of Twitter data in academic research is provided by Salvatore et al. (2021), as depicted in Figure 1.

Figure 1: *Academic relevance of Twitter*



This figure shows the conference papers and articles based on Twitter, produced between 2010 and 2018. Source: Salvatore et al. (2021).

Twitter users primarily share their daily activities or seek information, making this social network a natural source of insights regarding beliefs about politics, economics, and current events. Active users enjoy being part of the community, regularly tweeting, and connecting with others. This provides an opportunity to access real-time and detailed information on how individuals perceive the inflation rate and how it affects their daily lives. These perceptions may be influenced by several factors including personal experience, media coverage, and economic education. As Bricongne et al. (2022) shows, using Twitter to construct an inflation indicator offers several advantages. First, it provides timely information in real time and delivers insights more quickly than other indicators. Second, it is cost-effective as there is no need to design, develop, or conduct extensive surveys. Additionally, Twitter includes a wide range of opinions from users with diverse backgrounds, offering a more comprehensive view of inflation perceptions than traditional surveys do. As a

[1]Twitter has been rebranded as X following its acquisition by Elon Musk in 2022. The platform underwent significant transformations, including modifications to its name, logo, and strategic vision. X aspires to evolve into a comprehensive "everything app", expanding beyond its current social media functionality to encompass a diverse range of services, including financial transactions, e-commerce, and communication platforms.

non-traditional indicator, Twitter data complements the information collected from traditional survey-based indicators and financial markets. Moreover, Twitter is more reactive to short-term events and may be useful in providing early signals of changes in inflation. By leveraging these new data sources, policymakers can better understand inflation perceptions and expectations and make informed decisions. The use of Twitter as an indicator of inflation expectations has drawbacks. First, Twitter audiences are neither conventional nor permanent in nature. Therefore, the measures obtained may be subject to significant volatility. Second, it is difficult to ascertain whether Twitter users refer to future or present expectations or whether they are discussing price levels or inflation itself. Nonetheless, as Angelico et al. (2022) suggest, current judgments of economic conditions and inflation expectations are closely related, and similarities may be due to various factors including historical inflation experience, inflation perceptions and anchoring to actual inflation.

In this study, we construct two inflation perception indicators for Colombia covering the period from January 2015 to March 2023. The first indicator builds on insights from previous works by Angelico et al. (2022); Bricongne et al. (2022); Born et al. (2023), whereas the second indicator introduces a semi-supervised Word2Vec approach useful for summarizing tweet signals, thereby contributing directly to the literature. Moreover, this is the first study that track consumers' real-time inflation perceptions in developing countries, specifically in Colombia, monitoring how Twitter users perceive inflation and how these perceptions may affect their decisions and showing their usefulness in forecasting other measures. Our results suggest that our perception indicators complement the existing measures of inflation and expectations. We believe that common factors underlie the evolution of inflation perception, inflation rate, and expectations and that our perception indicators exhibit inherent behavior that could be useful in monetary policy. First, we find that our indicators show a dynamic similar to both expectations and observed inflation, confirming the connection between expectations and inflation and demonstrating that our indicators provide a reliable measure of inflation perceptions. Second, our indicators offer additional information to policymakers. Including them in forecasting models improves the predictive accuracy of the AR(p) models for alternative measures of inflation and inflation expectations up to six months ahead.

This document comprises seven sections, beginning with this introduction. The second section details our approach to extracting tweets and constructing our data. Given that Spanish is the official language of more than 20 countries, we explain how to determine whether a tweet is Colombian. In the third section, we outline our methodology for constructing our indicators. First, we combine a dictionary-based and Latent Dirichlet Allocation techniques to eliminate tweets referring to promotions, popular sayings, and cryptocurrencies. Subsequently, we construct two inflation perception indicators using the dictionary-based and Word2Vec approaches. In the fourth section, we present a real-time analysis of the indicators. In the fifth section, we compare the dynamics of our indicators with observed inflation and expectations and discuss the fact that all these measures share some underlying structural factors that may reflect similar dynamics. In the sixth section, we perform an empirical exercise to show that our indicators provide additional information in forecasting six measures of annual inflation and inflation expectations. In the concluding section, we summarize our findings and highlight the potential implications of this research for future studies in the field.

2. Data Collection and Pre-processing

Twitter is a popular social media platform in Colombia. According to Datareportal (2022), as of January 2022, there were approximately four million Twitter users in Colombia, representing approximately 10% of the population. Additionally, internet users in Colombia spend an average of 2 hours and 41 minutes per day on social media platforms, with Twitter being one of the most popular. Therefore, as Twitter is an important social media platform in Colombia, with a significant number of users, businesses, and organizations, it can

Figure 2: *Terms searched.*

This figure presents the terms included in the search query in the Twitter API.

leverage Twitter's reach and popularity to connect with potential customers and engage with their audience. To access Twitter's vast collection of tweets and to search for specific keywords within a given time frame, we used the Twitter API v2 standard search endpoint. The query included tweets in Spanish that included at least one of the words related to prices, inflation, and cost of living, as presented in Figure 2. We did not constrain the tweet location and searched for tweets between January 2015 and March 2023.

As an application of Natural Language Processing (NLP), our study employs a corpus, i.e. a collection of documents, of 83.4 million documents, each of which is a tweet. The first challenge we faced was identifying tweets from Colombian users. Given that only about 1% of tweets provide information on geographic location, and Spanish is the official language of 21 countries, we used a combination of tweet and user information to classify tweets as Colombian.[2] We established four criteria for identifying Colombian tweets: (i) tweets geolocated in Colombia, (ii) tweets containing words that refer to Colombia, (iii) users geolocated in Colombia, and (iv) users with descriptions that indicate Colombian origin.

To filter the database of Colombian tweets, we included tweets that contained information on demonyms, main cities, words related to the Bureau of Statistics, the Central Bank, and Colombian universities using the list of demonyms provided in Table A.1 in Appendix A. This procedure is repeated using the user's information. Finally, we removed tweets that exclusively discussed inflation in other countries, using a list of demonyms to identify the main Spanish-speaking countries and cities outside of Colombia, as detailed in Table B.1 in Appendix B. Thus, for instance, we will include a tweet from a Mexican user about Colombian inflation, if it is geolocated in Colombia or if it includes any of the denonyms related to Colombia in the text. In our approach, a tourist who describes prices in a Colombian place would be captured, as well as an analyst who explicitly refers to our country. On the other hand, a user identified as Colombian who only tweets about Mexican prices will not be included in our sample.

We identify 3.5 million tweets as Colombian. The specific outcomes of our classification approach are presented in Table 1, where each cell shows the number of tweets that meet the conditions of the corresponding

[2]Spanish is the official language in the following countries: Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, Equatorial Guinea, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Spain, Uruguay, and Venezuela.

row and column. Diagonal elements denote the number of tweets identified solely by one criterion. The total count for each row signifies the number of tweets identified by the criteria, and may not necessarily sum up the numbers within the row. Among 176,619 tweets, the criterion of tweet location identified the smallest number of Colombian tweets. Hence, employing additional criteria for user information, encompassing geolocation and description, yields a greater number of identified Colombian tweets and, consequently, a more extensive database for our analysis. Notably, 29,336 tweets fulfilled both the tweet descriptions and user location criteria, confirming their origin in Colombia. It is evident that user location emerges as the most pivotal criterion for our classification, with over 2.6 million entries in our database corresponding to users located in Colombian cities, towns, or within the country.

Table 1: Classification of Colombian Tweets.

	Tweet		User		
	Location	Description	Location	Description	Total
Tweet location	33,682	27,284	128,531	29,336	172,916
Tweet description		284,309	327,675	126,951	642,912
User location			1,895,308	524,717	2,742,944
User description				235,761	798,251

This Table shows the number of tweets that satisfied our four classification criteria based on the user and tweet information. The values on the diagonal show the number of tweets that were identified only with specific criteria. The numbers off the diagonal show tweets that satisfy the criteria in the row and column. Finally, the column Total indicates the number of tweets that satisfy the criteria described in the row.

To prepare our Colombian data for machine learning analysis, we cleaned it by removing any irrelevant or duplicated information, lowercasing all text, and removing stop words, hashtags, mentions, and punctuation marks. This step helped to standardize the data and reduce the number of features.

3. Constructing the inflation perception indicators

In this section, we discuss the methodology used to construct two alternative perception inflation indices. First, we removed tweets related to topics that did not contain relevant price signals. Next, we describe how our indicators are constructed, an N-gram indicator based on a dictionary that identifies whether each tweet indicates a price increase, decrease, or stability. Additionally, we constructed a Word2Vec (W2V) indicator based on the proximity of our tweets to keywords signaling price movements.

3.1. Removal of non price related tweets

To clean our data and obtain a signal related to inflation, we extract tweets that contain price content relevant to the Colombian CPI's basket of goods and services. Thus, we eliminated three categories of tweets that do not contain any useful price information: tweets containing popular sayings in Colombia, tweets related to cryptocurrency price dynamics, and tweets related to sales and promotions. We employed two classification approaches to remove non-inflation-related categories from our database. We search for N-grams that contain words related to sayings in Spanish, and we also use Latent Dirichlet Allocation (LDA) to identify tweets related to sales or promotions and eliminate them from our database.[3] First, we eliminate tweets containing

[3]Other possible methods of topic classification include dictionary-based approaches using pre-trained ML algorithms like Robertuito Pérez et al. (2021), Xml-RoBerta Conneau et al. (2020), or developing and training our own model.

Table 2: *Popular sayings removed.*

lo barato sale caro	no tiene precio
a cualquier precio	al precio que sea
precio de la fama	a precio de gallina robada
precio de la libertad	precio de la rectitud
no importa el precio	por ningun precio

This Table shows popular sayings in Spanish frequently used in Colombia, which contain words related to prices, but their message is not considered price signals.

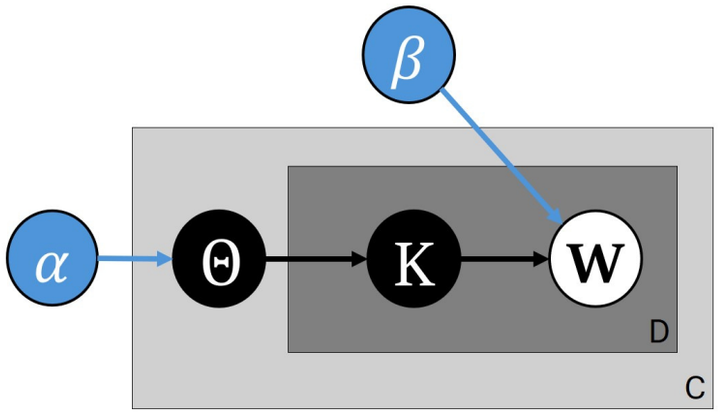
a list of n-grams including the terms "precio", "caro", or related words that belongs to phrases or popular expressions without price signal, Table 2.

Following the removal of popular sayings, we also eliminate promotions, as done in Angelico et al. (2022); Born et al. (2023), and remove cryptocurrencies, a topic not taken into consideration in previous works. This step is crucial for filtering out non-price signals from our database. In this step, we use a probabilistic topic modeling approach (LDA) proposed by Blei et al. (2003), which is also one of the most popular topic models in the natural language processing (NLP) field, a description of the algorithm is presented in appendix D.1. This is reinforced by its ease of use and the fact that it has been demonstrated to categorize text in the same way as people do Chang et al. (2009). Other studies that employ this strategy to summarize the corpus in subjects include Larsen et al. (2021), who created 80 time-series measures of the news topics the media reported, that is, the various types of news reporting. Guío-Martínez (2020) examine the underlying subjects in the Central Bank of Colombia's minutes and monetary policy reports and Gabrielyan et al. (2020) propose an index of inflation news that takes into account the intensity of a given topic using UK news.

LDA is used to identify the main topics discussed in a collection of documents (d) comprising a corpus (C). The algorithm assumes that each document is a mixture of a small number of topics (K) and that each word (W) in the document is drawn from one of these topics. The LDA algorithm follows several steps to identify topics: first, it creates a dictionary of words; second, it randomly assigns topics in each document (α) and words in each topic (β), with α and β being the parameters of the Dirichlet distribution. third, it iteratively updates the topic assignments based on the probabilities of the words belonging to each topic (Θ) and identifies the topics and the distribution of words in each topic. A simple scheme of the algorithm is shown in Figure 3. The output of the LDA algorithm is a set of topics, each represented by a distribution of words and a probability distribution of topics for each document in the corpus. Overall, the LDA algorithm provides a useful tool to discover underlying topics in a large corpus of text data. Finally, because the LDA estimation procedure does not provide the names of the topics, the assignment of labels to each topic is based on the most important terms within the topic.

To determine the optimal number of topics in our data, we applied the *Coherence Score* proposed by Röder et al. (2015). This metric involves grouping words into subsets, estimating the relative distance between words within each topic, and producing an aggregate coherence score. Thus, we can compare the consistency index calculated for 10, 20, or 50 topics in our database, and we found the highest score with 50 topics. Table 3 presents the terms with a high probability associated with each of the topics removed from our database to exclude information about promotions or cryptocurrencies. Each line in the table corresponds to one topic; thus, we removed ten topics related to promotions and eight related to cryptocurrencies. Finally, as a robustness check we performed a similar exercise of removing duplicates tweets as proposed by Born et al. (2023), to remove promotional bots. We filter out an additional 0.18% of tweets from our database.

Figure 3: *Graphical Model of LDA*



This figure shows a graphical representation of LDA model where alpha and beta parameters are chosen randomly to compute the probability (θ) that a word (W) is in a topic (K). Source: Li et al. (2017).

Table 3: *Words from topics related to promotions and cryptocurrencies.*

Promotions
venta, vende, info, informacion, excelente
bogota, ciudad, medellin, vendo
calidad, justo, pedidos, excelente
envio, pago, disponible, tienda, calidad, whatsapp, info
super, descuento, pagina, informacion, whatsapp, disponible, bogota
servicio, mes, calidad, whatsapp, excelente
productos, mejores, calidad, enda, info, producto, servicio, dia
domicilio, pide, contactanos, metro, whatsapp, servicio, envio, calidad, ciudad
compra, cafe, final, via, venta, especial, colombiano, descuento
envio, gratis, cali, whatsapp, bogota, medellin, info
Cryptocurrencies
usd, eth, ethereum, mar, final
usd, ltc, litecoin, mar, via, cara
usd, xmr, monero, mar, via
bitcoin, btc, tasa, fuente, usd
millones, dolares, colombianos, vale, pasa, empresa, mes
pesos, mil, colombianos, millones, vale
ana, final, bogota, despues, economia, mercado, gasolina, alza, informacion
xrp, usd, via, mas, bogota, mar, dolar, pais, alto, servicio, pagar, productos

This Table shows the words associated with the topics classified as promotions and cryptocurrencies according to the LDA. These topics were removed in our analysis.

3.2. N-gram Indicator

To construct this indicator, we use a dictionary-based approach by classifying our tweet signal as inflation up, down, or constant according to the selected n-grams. Thus, a tweet is classified into these categories only when an n-gram appears in its contents. The list of terms is a modified version of that proposed by Angelico et al. (2022), in which we include terms related to monetary policy, housing, and electricity important topics according to Born et al. (2023).[4] Finally, we add other relevant expressions for the Colombian case. We employ 333 expressions to denote inflation up and 204 to denote inflation down, as presented in Tables C.1 and C.2 in Appendix B. To illustrate our findings, Figure 4 presents the most frequent n-grams. Once we classify our tweets according to their signal, we add all information to construct the daily indicators of increasing inflation U_t and decreasing inflation D_t. We compute the balances using the daily aggregates of each signal and the formula of **Indicator #4** proposed by Angelico et al. (2022), that is, the difference between the logarithms of the upward and downward tweets per day $B_t = ln(U_t + 1) - ln(D_t + 1)$.[5].

In order to remove the volatility we find in the balance B_t, we employ two techniques. First, we apply winsorization, which involves capping any value that exceeds three standard deviations from the mean. This helps mitigate the influence of extreme values on the indicator. Second, we implement a smoothing procedure employing a backward-looking moving average over a period of 30 days. This moving average helps reduce short-term fluctuations, provides a more stable representation of the indicator's trend, and ease comparison with monthly series. Finally, our indicator is constructed as balanced divided by its standard deviation; that is, $\pi_t^T = \frac{B_t}{\sigma(B_t)}$. This strategy facilitates comparisons over time and across the variables. Moreover, this scaling approach not only maintains the sign of the initial balance between the upward and downward signals but also provides a scale based on the historical behavior of the indicator.

3.3. Word2Vec Indicator

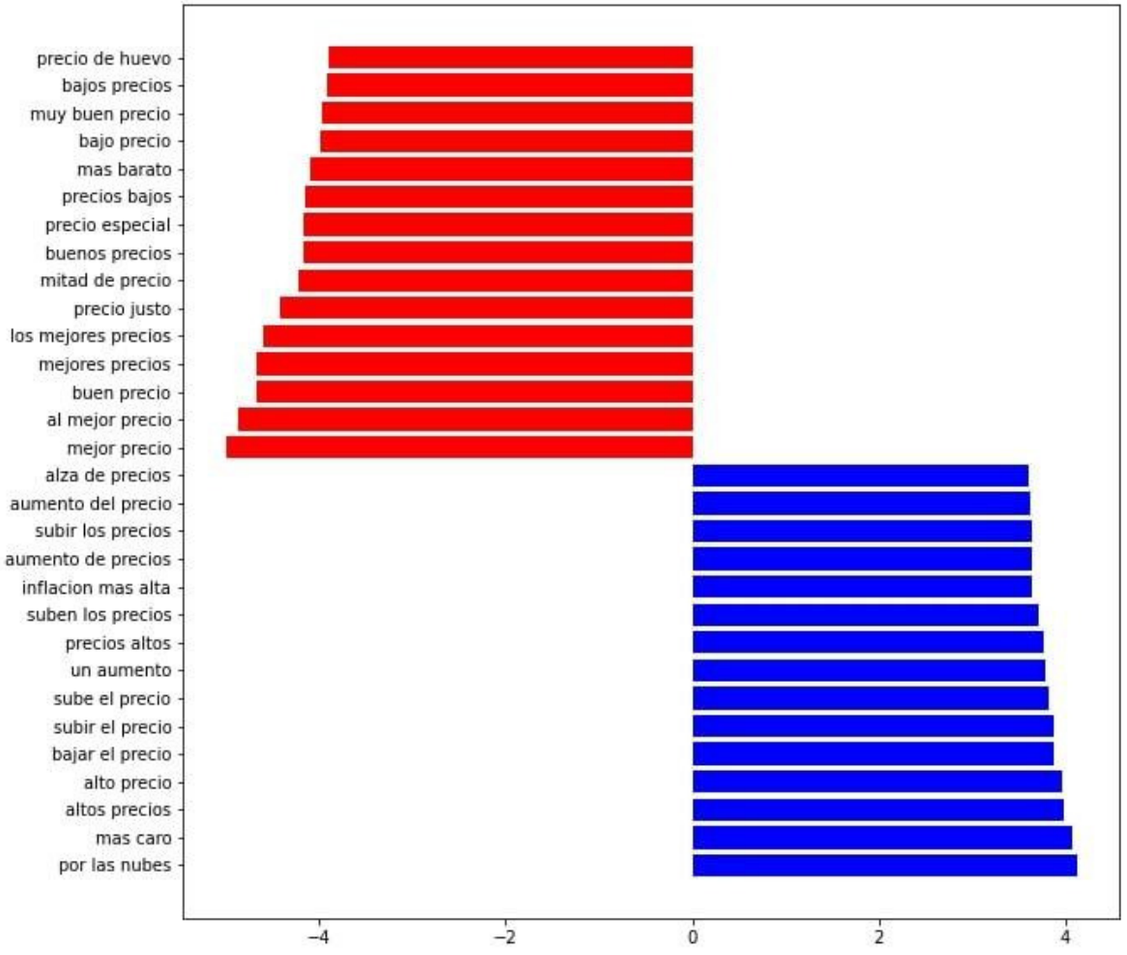
An alternative methodology for extracting a price signal from our Twitter corpus is to use the Word2Vec algorithm, which is a powerful tool for capturing semantic and syntactic relationships between words in a text corpus Mikolov et al. (2013). This technology extracts useful information from large amounts of unstructured text data and can be helpful in identifying patterns and trends that are difficult to notice using conventional approaches. This method is based on the idea that the meaning of a word can be inferred from its context. This method creates a vocabulary from text data and then learns the distributed representations or word embeddings of each word in the vocabulary. These embeddings encode the meaning of a word based on its co-occurrence patterns with other words in the text.

W2V addresses two main tasks: Continuous Bag-of-Words (CBOW) and Skip-Gram models. The first predicts the target words from the surrounding context words and the second predicts the surrounding context words from the target words. A more detailed description of the algorithm is provided in appendix D.2. We focus on Skip-Gram method under which to extract a rising price signal, we determine a set of terms linked to price rises, such as "expensive" or "price hike" and then compute the cosine similarity of the words in each tweet with the terms. This is a measure of vector similarity that ranges from 0 to 1, with 1 denoting a complete similarity. Tweets containing an increasing price signal are more likely to have a high similarity to price-up

[4]It is important to highlight a difference among Colombia and industrialized economies, using our approach we did not find an important contribution of electricity in our final indicator. This may be due to the fact that electricity is a regulated good in Colombia and its price evolution depends on technical factors in contrast to industrialized countries where electricity is a key determinant of headline inflation and its price fluctuations affect more strongly daily lives.

[5]In their study Angelico et al. (2022) present three other balances which we also implemented and have a close dynamic and results in the empirical analysis to our indicator

Figure 4: *Most frequent n-grams*



This figure shows the 15 most frequent N-grams indicating upward signals and the 15 most frequent N-grams indicating downward signals in our database.

terms. As an example, consider two vectors, A and B, and the cosine similarity is defined as:

$$similarity_{A,B} = \cos\theta = \frac{A \cdot B}{AB}$$

where A and B denote the norm of each vector.

To capture the tweets most closely related to each signal, we selected vectors with θ close to 0. In our application, we define three terms or centers to capture increasing price signals. “alto”, “alta”, and “alza”. The algorithm computes as nearest neighbors: ‘alto’, ‘alta’, ‘alza’, ‘aumento’, ‘subida’, ‘incremento’, ‘nubes’, ‘crecimiento’, ‘subiendo’, ‘caro’, ‘alimento’, ‘subieron’, ‘subido’, ‘encima’, ‘crisis’, ‘aumenta’, and ‘devaluacion’. With respect to the decreasing signals we select as centers: “bajo”, “bajo”, “caida”, which the algorithm associate as nearest neighbors: ‘bajo’, ‘baja’, ‘caida’, ‘debajo’, ‘bajan’, and ‘menor’. Next, we use words inferred from the corpus to classify the signal prices in our tweets. Finally, to construct our W2V inflation perception index, we summed the number of tweets of the upward price signal U_t and the downward tweets D_t, as well as their logarithmic balance, using the same procedure used in the N-gram indicator to remove volatility and provide a scale.

4. Real-time analysis

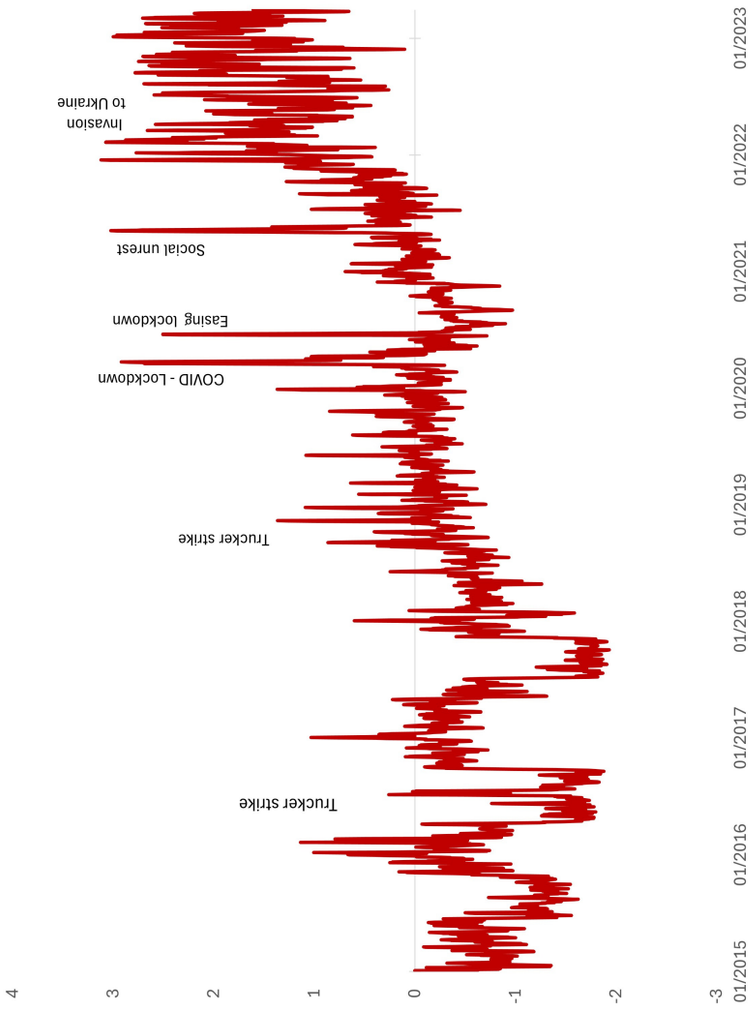
In this section, we present our estimated indices and discuss their dynamics and real-time properties. In Figure 5, we present our N-Gram index, which shows significant volatility, similar to that in previous studies. We describe the main events in Colombia during this period and focus on prominent peaks associated with particular events. At the beginning of our sample, Colombia experienced the "El Niño" weather phenomenon up to May 2016, with its effects intensifying during the first semester of 2016. We observe two peaks in our indicator on July 16th and 18th, related to the increase in prices due to the trucker strike, which reduced the supply of food and other goods in the main cities. Until 2019, the index exhibited stable behavior, with some peaks, most notable in November 2018, coinciding with the beginning of the longest trucker strike in our sample period. In 2020, our index showed two significant peaks: the first on March 23rd, coinciding with citizens' uncertainty about the beginning of the lockdown, and the resulting panic buying to provision for quarantine. The second peak occurred on June 17th when restrictions on some economic activities were lifted, leading to concerns about supply, contagion, and price increases.

After the lockdown measures ended in August 2020, we observed an increase in our indicator, primarily because of rising transport costs. On May 8, 2021, there was another peak associated with social unrest in Colombia, including several marches against the government and a trucker strike. Subsequently, there was a substantial increase in expectations, largely attributable to supply shocks resulting from cost hikes in transportation and the Russian invasion of Ukraine. This also marks a peak in Colombian tweets in February due to the uncertainty around oil prices and their effects on Colombia's inflation. The index became more volatile in 2022, with uncertainty in the economic environment up to May due to elections, and afterward, uncertainty arose from the effects on employment, economic growth, and prices due to the fiscal and social reforms proposed by the president.

As highlighted in the Introduction, one of the challenges encountered in classifying unstructured data is determining whether Twitter users express expectations regarding inflation or price levels. To address this issue, we exploit the known structure of the dictionary approach, classify our original n-grams, analyze whether each n-gram refers to prices or inflation, and compute the percentage of tweets that refer to price levels and inflation. Our analysis reveals that approximately 55% of tweets in Colombia, which indicates an increase in our inflation indicator, are related to inflation itself, while the remaining 45% are directly related to price levels. In contrast, when the indicator signals a decrease, approximately 42% of the tweets specifically refer to inflation, while 58% refer to general price levels or specific products. This suggests that Twitter users do not distinguish between these two concepts.

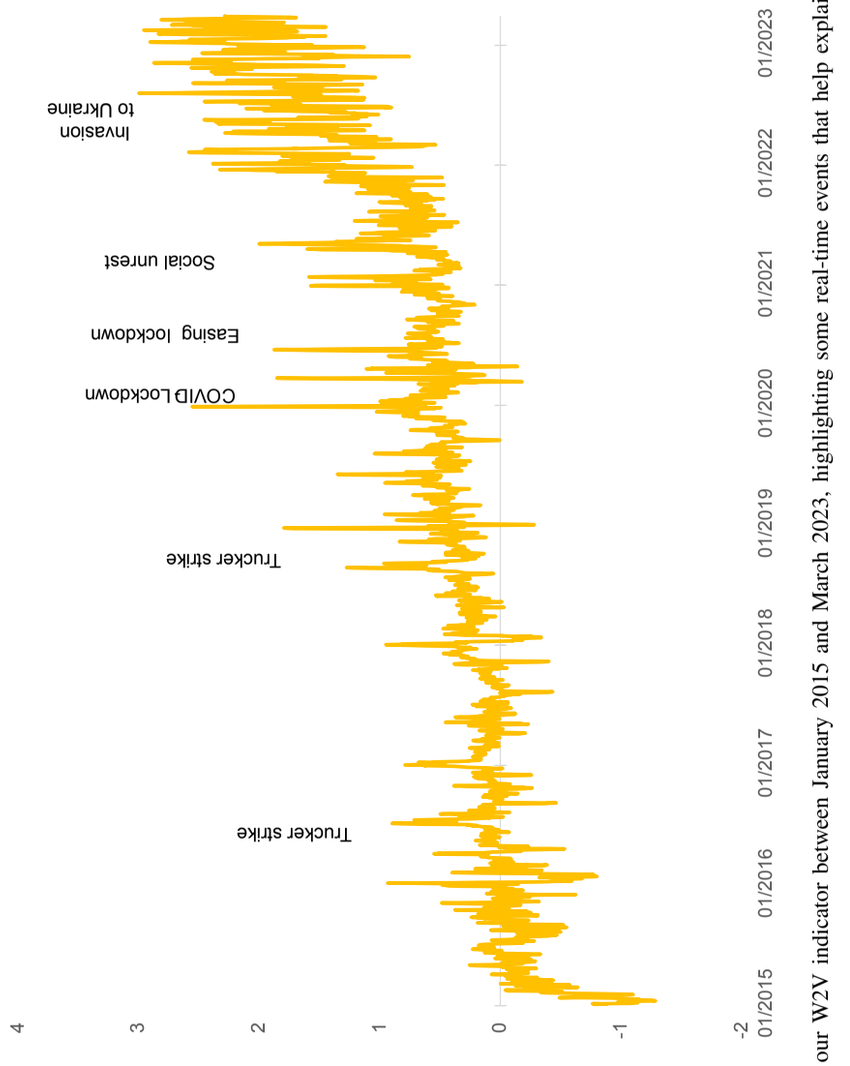
Figure 6 shows the W2V perception index. For reference, we also plotted the same events as those discussed for the N-gram indicator. We observe peaks on these dates, indicating that this index also captures these events. Notably, there is an increase in volatility following the Russian invasion of Ukraine. Additionally, we include a new event: the social unrest at the end of 2019, which marked strong social discontent with concerns about the economy, prices, and unemployment. Historically, the index remains quite stable between 2015 and 2020. However, starting in 2021, there is an increase in the index, which can be associated with supply difficulties, increased shipping costs, and the adverse effects of the invasion of Ukraine on uncertainty and the supply of goods.

Figure 5: *N-gram indicator perception inflation index.*



This figure presents our N-gram indicator between January 2015 and March 2023, highlighting some real-time events that help explain the volatility during particular episodes.

Figure 6: *N-gram indicator perception inflation index.*



This figure presents our W2V indicator between January 2015 and March 2023, highlighting some real-time events that help explain the volatility during particular episodes.

5. A Comparative Analysis of our inflation perception indices, surveys, and market-based expectations.

In this section, we compare our inflation perception indicators with inflation and inflation expectation measures to determine whether common factors underlie their dynamics. In order to make a comparable monthly series, we compute the average of the balance for each month. First, we compare with annual headline inflation and inflation excluding food, this last measure eliminates volatility and is one of the main core inflation indicators.[6] Additionally, we compare our indices with two traditional measures of annual inflation expectations: The average response to a monthly survey that involves specialized forecasters from banks and other financial institutions and the expectation of breakeven inflation for one year (BEI), which measures the average expected inflation for one year implicit in the data of the sovereign bond market[7] Specifically, BEI is calculated as the difference between the yield of domestic currency bonds (TES in pesos) and that of inflation-linked bonds (TES in UVR). However, these estimates may be affected not only by inflation expectations, but also by uncertainty about expectations as well as by liquidity frictions, and there are alternative strategies to obtain the component related to expected inflation, e.g. Espinosa-Torres et al. (2017). In our exercises, we preferred to use unadjusted estimates. In this way, our results do not depend on the choice of a specific methodology to isolate the inflation signal and are based on information available to any market participant or professional forecaster.

5.1. Comparison with observed Inflation

Figure 7 compares our N-gram and W2V indicators with the inflation measures between 2015 and 2020. The left panel shows that our N-gram indicator exhibits a trend similar to that of the actual inflation rates between 2015 and 2020. A significant upsurge was observed during Q1-2016. The N-gram indicator effectively captures this spike, indicating a subsequent reduction in inflation levels during 2017 - 2018. However, as illustrated by the right panel graph, our W2V indicator shows only a small upward signal around June 2016, and remains stable between 2015 and the first quarter of 2018. In 2019, the N-gram and W2V indicators demonstrated distinct behaviors. Specifically, the former indicates a rising inflationary trend, whereas the latter suggests a constant level of inflation that aligns more closely with the observed measures. However, as we move into Q1 2020, both metrics display a marked increase in inflation rates; this uptick is not reflected by actual inflation measures, which indicates that any significant growth will only occur from 2021 onward. One factor behind this upward trend in both indicators is the increase in the number of tweets captured in both methodologies, where the average number of tweets captured in the N-Gram pass from 3787 in 2021 to 7673 in 2022, and from 6375 to 15227 for the W2V indicator. This fact reflects the higher importance of prices and inflation-related topics to Twitter users. As Weber et al. (2023) indicate, when inflation is low, users pay less attention to it. Therefore, users focus less on news and do not make important expectation adjustments, which may be reflected in their daily experiences. However, when inflation is high or rising, agents will see it in their expenditures; pay more attention, obtain more information, and communicate their daily experiences more frequently. Looking further ahead to late-2022, we observe interesting changes within these indicators: whereas n-grams signal stability for future periods, W2V may experience marginal reductions instead.

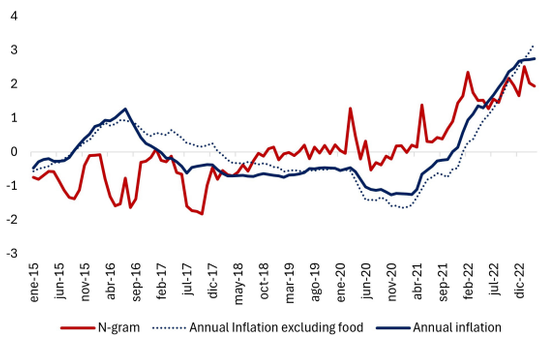
The similarity between our indices and observed inflation raises the question of whether we are measuring current perceptions or expectations or why there is a strong connection among them. Inflation expectations and actual inflation share many similarities due to various factors, including historical inflation experience,

[6]The Consumer Price Index is published monthly by DANE and can be accessed at https://www.dane.gov.co/index.php/estadisticas-por-tema/precios-y-costos/indice-de-precios-al-consumidor-ipc.

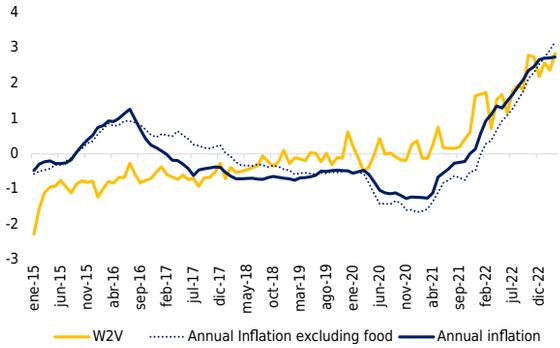
[7]The methodology and data from the survey for financial forecasters are available at: https://www.banrep.gov.co/es/estadisticas/encuesta-mensual-expectativas-analistas-economicos.

Figure 7: *Comparison of our indices Headline inflation, and inflation excluding food.*

(a) *N-gram indicator.*



(b) *W2V indicator.*



This figure compares the evolution of our two perception indices with annual inflation and annual inflation excluding food.

inflation perceptions, and anchoring to actual inflation rates. First, the historical inflation experience has a significant effect on future inflation rates expectations, and people who have experienced higher inflation rates in the past are likely to have higher inflation expectations in the future Xu et al. (2016). Second, empirical evidence indicates that both inflation perceptions and past inflation significantly affect the formation of inflation expectations. This finding suggests that people's perceptions of inflation and their experiences with past inflation rates can influence their expectations of future inflation rates Xu et al. (2016). Finally, Grant and Thomas (1999) found cointegration between survey measures of inflation expectations and actual inflation, indicating a strong relationship between the two variables, and inflation expectations have been anchored to both inflation targets and actual inflation, further emphasizing the similarity between these two variables Cicek and Akar (2014). Furthermore, studies find that inflation expectations tend to increase as actual inflation accelerates, highlighting the importance of actual inflation rates in shaping individuals' expectations Feldkircher and Siklos (2019). These findings suggest that individuals' expectations for future inflation rates are influenced by their past experiences, perceptions, and actual inflation rates.

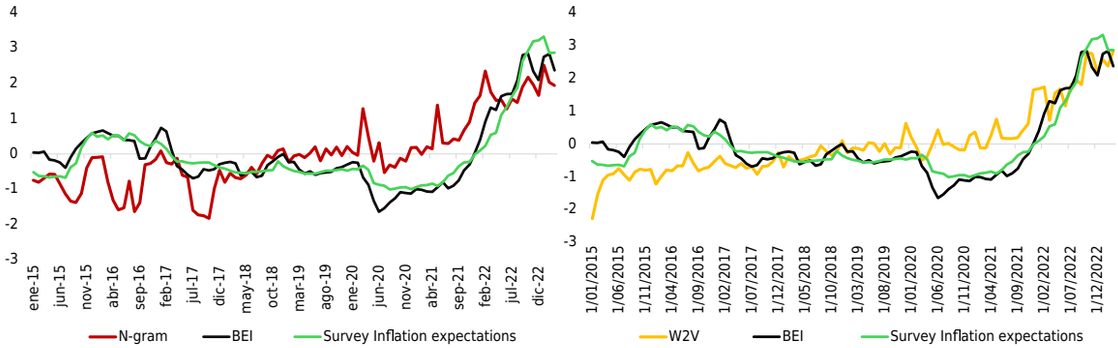
Our perception indicators have a weaker co movement with inflation, than the Twitter based measures in previous works for Italy, France, and Germany. Although there are some events like the increase in prices after the pandemics, that affect all countries similarly. There are idiosyncratic factors, affecting Colombia that may reflected in user's opinions. These differences may be supply factors, such as "El Niño","La Niña", and trucker strikes that may affect the perception from Twitter users more than it affect current inflation.

5.2. Comparison with Inflation Expectations

In this subsection we compare our indicators with annual inflation expectations measures, given that they are the most commonly known. Figure 8 presents our N-gram and W2V indicators and the traditional inflation expectation measures from January 2015 to March 2023. The upper panel shows our N-gram indicator, which increased in 2016, followed by subsequent declines in both 2017 and 2018, similar to that of break-even inflation and analyst surveys. However, in 2019, it deviated from the traditional indicators in terms of the observed increase in inflation perceptions. In contrast, our W2V indicator results showed stability similar to that of the two standard expected indication techniques. Finally, as mentioned in the previous subsection, our indicators indicate an increase in inflation perceptions by 2020. This finding is not consistent with the observed inflation or break-even inflation (BEI) expectations but may be consistent with the survey-based indicator, which shows that inflation expectations did not decrease during 2020. This difference could be motivated by the

Figure 8: *Comparison of our indices with expectations measures.*

(a) *N-gram indicator.* (b) *W2V indicator.*



This figure compares the evolution of our two perception indices with inflation expectations from a survey and those derived from the Break-Even Inflation.

fact that during pandemics, the reduction in inflation was led by recreation, culture, and other goods that were not being supplied, while the essential goods did not reduce their price. This fact is clearly captured in our indicator but may not be reflected in the analysts' opinion nor the financial market. Afterwards, we observe an increase in expectations and our perception indicators.

The similarity of our indices to these measures of inflation expectations may be caused by common underlying economic factors such as economic growth, monetary and fiscal policies, and global economic conditions. Furthermore, economic agents, such as investors, analysts, and consumers, use news releases, economic indicators, and financial market developments to form their expectations—information that is easily available on the Internet and even spread on Twitter. However, there could be factors that indicate differences in the information set for each agent. Expectations typically refer to headline inflation or exclude food inflation, but perceptions may refer to a smaller set of items that drive people's decisions and experiences. For example, a spike in fuel prices might heavily influence consumer perceptions but does not significantly alter broader inflation expectations. Another example is that expectations may adjust quickly to a new monetary policy, while perceptions might lag because of slower public dissemination of information. Finally, perceptions may be disproportionately affected by media coverage of specific goods such as food or housing, leading to a skewed view that does not align with overall inflation expectations.

An empirical exercise to show that our indicators π^T share common information with traditional inflation expectations π^E from break-even inflation and the expectation of monthly survey is to estimate the following regression: we regress the expectation measures with a constant and our two indicators. Thus, if there is a relationship among these variables, the associated parameter γ_1 will be significant.

$$\pi_t^E = \gamma_0 + \gamma_1 \pi_t^T + \eta_t \tag{1}$$

Table 4 presents the estimates for Equation 1. Columns 1 and 2 show the results for the N-gram indicator, while columns 3 and 4 show the results for the W2V indicator. As we can see, all estimates of γ_1 are significantly different from zero; thus, our measures relate to traditional inflation expectations.

In conclusion, our N-gram and W2V indicators show a close relationship between their dynamics and that of current inflation and inflation expectations. This finding is not surprising, given the strong connections we have discussed between actual events and how agents revise their inflation expectations based on the temporal nature of the shock. Specifically, whether the shock is expected to be permanent or transitory plays a significant role in shaping the agents' inflation expectations.

Table 4: *Relationship between our indicator and traditional inflation expectations.*

	N-gram		Word2Vec	
	BEI	EMEA	BEI	EMEA
γ_1	2.691***	0.015***	1.123***	0.007***
	(0.63)	(0.002)	(0.141)	(0.001)
γ_0	4.721***	0.033***	3.819***	0.041***
	(0.262)	(0.001)	(0.238)	(0.001)
Obs	96	96	96	96
R^2	0.165	0.181	0.394	0.479
R^2_{adj}	0.188	0.267	0.388	0.483

This table presents the estimates of equation 1, in which the coefficient γ_1 capture the linear relationship among expectations and our Twitter indicators. * indicates 10% significance, ** indicates 5% significance, and *** indicates 1% significance.

5.3. Robustness check - Perception Indicators for experts

One robustness check of our indicator is to construct it using only users with a better knowledge of economic conditions. Therefore, we classified them as experts if in their user description they could be identified as experts, analysts, or economists. Using this strategy, we obtained a new database of 293,264 tweets, which is approximately 8.5% of the initial database. We then follow our strategy and construct two indicators using only the experts' tweets, in which promotions and popular sayings are not necessary because these tweets are clearly associated with price information. Using these criteria, we checked and removed only five tweets. However, some tweets are related with crypto-currencies, so we remove them using the same LDA model used for the whole database.[8]

Figure 9 presents a comparison between our inflation perception indicators constructed from expert tweets and the results in the previous section. The left panel of the figure shows that until 2018, the expert indicator had a similar dynamic to the N-gram indicator but with a higher level. From 2019, the level and dynamics of the expert indicator became closer to those of the original indicator, and from 2021, the expert indicator once again had a higher value. Regarding the W2V indicator, the dynamics of the total database and those of experts are similar, except for the period between 2018 and 2020, where the expert indicator signals a higher inflation perceptions. However, our baseline indicators follow the behavior of inflation and traditional expectation measures more closely. It is worth noting that the two expert indices have volatility that is similar to the original indices.

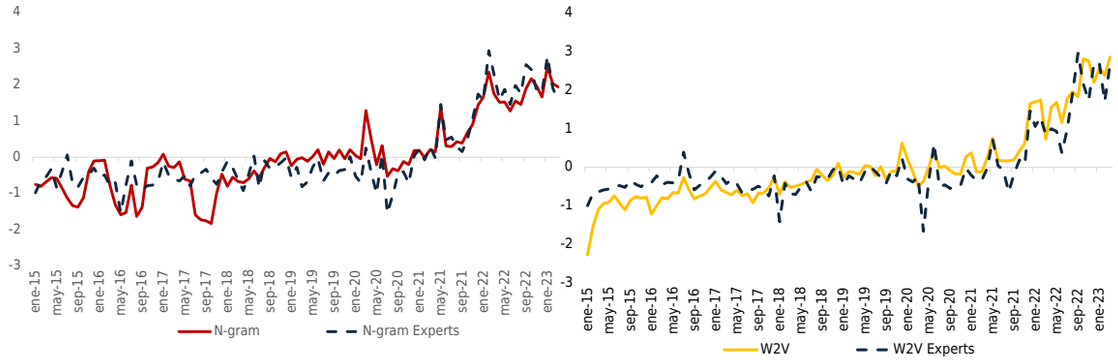
6. Predictive accuracy of our indices.

In this section, we analyze whether our Twitter indicators provide information for forecasting annual and monthly inflation rates and inflation expectations up to six months ahead. In order to do so, we test whether including our indicators improves the forecasting accuracy of the AR(p) model, which minimizes the Root

[8]When we provide the whole database to the LDA algorithm we show enough information so it can learn terms related to crypto-currencies than if we only use the experts database to estimate the LDA.

Figure 9: Comparison of our indices and an expert subsample.

(a) N-gram indicator. (b) W2V indicator.



This figure compares the evolution of our two perception indices with the evolution of the Expert indices, which are constructed by focusing on tweets from analysts and economic institutions.

Mean Squared Error (RMSE). Our strategy, which is based on the approach suggested by Fitchett and Robinson (2021), can be described in three steps.[9] In the first step, we estimate the AR(p) models up to a lag of six for each horizon h and the reference series Y_t and $h = 1, ..., 6$.

$$y_{t+h} = \alpha_1 y_t + \alpha_2 y_{t-1} + ...\alpha_p y_{t-p+1} + \epsilon_{h,t} \tag{2}$$

Then we compute the forecast error for each reference series and horizon in a rolling forecast evaluation using the last 39 months of our sample, that is, from January 2020 to March 2023. In this step, we select the model that minimizes the RMSE.

In the second step, we include into the previous specification our Twitter indicators π_t^T, compute the direct forecast, and the RMSE.

$$y_{t+h} = \alpha_1 y_t + \alpha_2 y_{t-1} + ...\alpha_p y_{t-p+1} + \beta \pi_t^T + \epsilon_{h,t} \tag{3}$$

In the third step, we compare the forecast accuracy of the two models. To do so, we compute the relative RMSE as the quotient of the RMSE of the augmented model and the RMSE of the reference model, equations 3 and 2. A value of one indicates that both models have the same accuracy, a value lower than one suggests that adding the Twitter indicator improves forecast accuracy, and a value greater than one suggests that it worsens accuracy. Finally, for models in which the relative accuracy is lower than one, we check whether this improvement is statistically significant using the Giacomini and White (2006) test. This test not only focuses on the mean but also on the variance and computes the difference between the squared forecast errors of the AR(p) model and those from the expanded model. For a particular horizon h, the test evaluates the null hypothesis that there is no difference between the errors from both models.

$$H_0 : E[\epsilon^2_{t+h,EQ2} - \epsilon^2_{t+h,EQ3}] = 0 \tag{4}$$

In our analysis, we use the direct forecast approach because it is less susceptible to bias resulting from incorrect model specifications, as suggested by Marcellino et al. (2006). Additionally, it does not require the

[9]The approach of these authors evaluate the predictive power of a series by comparing direct forecasts from the two nested models. The first model is an AR(1) of the reference series while the second includes the series of interest.

forecasting of explanatory variables, thus avoiding uncertainty in their estimations. Our forecasting period is marked by significant fluctuations in the inflation rates. Initially, there was a decrease due to the COVID pandemic in 2020, followed by a notable rise due to supply shocks, such as increased transport fees and Russia's invasion of Ukraine, which substantially increased costs globally.

To evaluate the forecast accuracy, we chose nine reference series, three from expectations: the expectations of total inflation derived from the Break Even Inflation and the total and excluding food expectations from the monthly survey to financial institutions. In addition, we include six inflation indicators that encompass various dimensions: headline inflation, core inflation, and four baskets that encapsulate diverse sources of inflationary pressures: administration, services, goods, and food.[10] In the administrative category, we include products that are subject to government price control, such as gasoline, electricity, and other utilities. The services category comprises goods whose pricing is less influenced by exchange rates and, therefore, is more contingent on internal demand dynamics. The goods basket includes items whose prices are shaped by international competition, thus reflecting both domestic and exchange rate-related factors. The food category encompasses products vulnerable to weather and transportation cost fluctuations, which predominantly react to demand-side shocks. In our analysis, core inflation excludes food classification, which corresponds to the items in the first three baskets. Headline inflation is the sixth indicator.

We develop our analysis for annual and monthly variations of our indicators. First, Table 5 presents the relative RMSE of our evaluation of the predictive accuracy for both perception indicators, covering a forecasting horizon of up to six months of annual change in the reference series. As previously mentioned, we first determined the AR(p) model with the lowest RMSE. For each of the nine reference series, we checked for up to six lags to determine the best autoregressive process in terms of forecast accuracy. In our analysis, the AR(2) model was chosen for the annual inflation of Food, Services, Core, and Total inflation rates, as well as for expectations derived from surveys and Break-even Inflation (BEI). For the remaining two series, the AR(1) model was selected as the best model. Each column corresponds to a distinct alternative measure of inflation or inflation expectations. The upper panel of the table corresponds to the N-gram indicator.

In terms of predicting food inflation, our findings indicate that incorporating this indicator improves the forecasting accuracy by 9% for the one-step-ahead forecast and by 20% for the six-month-ahead forecast. Moreover, in statistical terms, the improvement in forecast accuracy is significant at the 10% level for forecasts three to six months ahead. The forecasting accuracy of the remaining inflation measures is also improved by including our indicators. Generally, there are small and non-statistically significant gains for one-month-ahead forecasts, whereas the gains in forecasting accuracy become substantial for six-month-ahead predictions. Larger gains were observed in Food and Core Inflation, with reductions in the RMSE greater than 20%. Overall, the forecasting of headline inflation improves across all horizons and these gains are statistically significant. Regarding inflation expectations, we found an improvement across all horizons for break-even inflation, which is significant in almost all horizons, except for the two-month-ahead forecast. The forecast accuracy for measures derived from surveys did not improve for up to three months ahead. However, the forecast improves by up to 22% for the headline inflation.

Turning our attention to the results for our W2V indicator (lower panel of the table), we observe a consistent reduction in the relative RMSE when this indicator is included in the forecasting exercise. This reduction is pronounced for the basket of food and the total inflation, with reductions in the RMSE of up to 25%. Forecasts for the inflation of services and core baskets are reduced by up to 9% and are statistically significant. On the other hand, there is no improvement in forecasting regulated inflation. Importantly, the gains achieved in terms of forecasting accuracy for the W2V indicator tend to be lower than those observed for the N-gram indicator.

[10]The administrative category represents 17.3% and comprises 14 items, Services represents the highest percentage at 48.9% with 45 items, Goods includes 70 items, and Food includes 59 items that cover 15.1%. For a comprehensive grasp of this categorization, refer to González-Molano et al. (2020).

Table 5: *Relative forecast accuracy of the annual variation of reference series*

	Inflation measures						Expectations		
h	Food	Goods	Regulated	Services	Core	Total	BEI	Core	Total
					N-Gram				
1	0.91	0.99	0.94*	0.98	0.96	0.95*	0.98*	0.86	1.01
2	0.89	0.88*	0.92*	0.92**	0.87*	0.92**	0.94	0.99	1.01
3	0.92*	0.97	0.87**	0.93*	0.85*	0.90*	0.91*	0.94	0.97
4	0.90*	0.81**	0.88**	0.87*	0.80**	0.86**	0.89**	0.90*	0.87*
5	0.83**	0.76**	0.89*	0.81**	0.78**	0.81**	0.73**	0.87**	0.77**
6	0.80**	0.72**	0.89*	0.86*	0.77**	0.82*	0.70**	0.83*	0.78*
					W2V				
1	0.93*	0.98	1.00	0.94**	0.96*	0.96*	0.95*	0.89	1.01
2	0.83**	0.92**	0.99	0.91**	0.91**	0.92**	0.89*	0.96*	0.97*
3	0.79**	0.95	0.97	0.91**	0.91*	0.85**	0.83*	0.96*	0.94*
4	0.72**	0.91**	0.96	0.95*	0.95	0.88**	0.78*	0.93**	0.88**
5	0.74**	0.88**	0.94	0.96*	0.91*	0.83**	0.72**	0.90**	0.83**
6	0.75**	0.81**	0.93*	0.99	0.91**	0.85**	0.69**	0.86**	0.84**

This table presents the Relative RMSE computed as the RMSE of the specification including our Twitter indices equation relative to the RMSE of the AR(p) model with the best forecast accuracy for predicting the annual variation of alternative measures of inflation and inflation expectations. Additionally, we indicate whether the differences in forecasting accuracy are statistically significant using the test by Giacomini and White (2006). * indicates 10% significance, ** indicates 5% significance, and *** indicates 1% significance.

We also check the improvement in the forecast accuracy for monthly inflation and inflation expectations, Table 6. In this case, the AR(2) model was the best autoregressive process for the expectations derived from the surveys and BEI, while the AR(1) model was selected for the remaining series. As shown in the upper panel, despite the fact that the inclusion of the N-gram indicator reduces the RMSE, this difference is not statistically significant according to the test, primarily because of the higher volatility of monthly inflation. Nevertheless, we observe improvements in all inflation measures and expectations for forecasting inflation between four and six months ahead. The most significant reductions were observed in the inflation of food, goods, core, and total inflation as well as in expectations according to break-even inflation. However, statistical evidence does not suggest an improvement in forecasting regulated inflation and service inflation. The improvement in accuracy was similar for the W2V indicator across different measures. Finally, it is worth noting that the monthly improvement in break-even inflation may also be influenced by statistical factors related to its construction, which makes it highly volatile, specifically due to the fact that it interpolates at a low frequency, given the liquidity constraints.

7. Conclusions

Inflation expectation and perception measures are crucial in the analysis under the inflation-targeting regime. Having complementary indicators is useful for policymakers to portray different dimensions given their limitations and advantages. Social media platforms such as Twitter offer policymakers an opportunity to access real-time and detailed information on inflation expectations. This approach complements traditional mea-

Table 6: Relative forecast accuracy of the monthly variation of reference series

	Inflation measures						Expectations		
h	Food	Goods	Regulated	Services	Core	Total	BEI	Core	Total
					N-Gram				
1	0.86	1.00	0.97	1.04	1.00	0.97	0.98*	1.06	0.98
2	0.86	0.96	0.93	0.99	0.97	0.92	0.90*	0.94	0.93
3	0.87	0.82	0.92	0.93	0.89	0.85	0.89*	0.90	0.86
4	0.85*	0.88**	0.91	0.92	0.85*	0.80**	0.82**	0.92*	0.79**
5	0.74*	0.83**	0.91*	0.91*	0.84**	0.76**	0.66**	0.94*	0.80**
6	0.74**	0.80***	0.89**	0.89	0.81*	0.72***	0.31**	0.93*	0.79**
					W2V				
1	0.88	1.01	0.96	1.05	1.03	0.96	0.90*	1.02	1.00
2	0.85	0.98	0.94	1.01	0.97	0.89	0.80**	0.90	0.92
3	0.84*	0.83	0.93	0.94	0.90	0.84*	0.78*	0.86	0.85
4	0.83*	0.92**	0.93	0.97	0.92*	0.84*	0.78*	0.84*	0.88*
5	0.75**	0.90**	0.94	0.97	0.91**	0.82**	0.69**	0.84**	0.91*
6	0.73**	0.81**	0.91*	0.96	0.90**	0.82**	0.35**	0.84*	0.91*

This table presents the Relative RMSE computed as the RMSE of the specification including our Twitter indices relative to the RMSE of the AR(p) model with the best forecast accuracy for predicting the monthly variation of alternative measures of inflation and inflation expectations. Additionally, we indicate whether the improvement in forecasting accuracy are statistically significant using the test by Giacomini and White (2006). * indicates 10% significance, ** indicates 5% significance, and *** indicates 1% significance.

sures of inflation expectations in several ways. First, social media platforms provide real-time information and deliver insights faster than other indicators. Second, it is cost-effective because there is no need to design, develop, or conduct an entire survey. Additionally, social media platforms include a wide range of opinions from users with diverse backgrounds. Moreover, social media platforms are more reactive to short-term events and may be useful in providing early signals of changes in inflation.

Therefore, following Angelico et al. (2022); Bricongne et al. (2022); Born et al. (2023), we construct two inflation perception indices using over three million tweets collected from 2015 to 2023 via the Twitter API. The first indicator, N-gram, is based on a supervised approach using a dictionary created for Colombia, whereas the second uses the Word2Vec approach to determine whether a tweet conveys rising or decreasing price information.

Our indicators show that they follow real-time events in Colombia and have close dynamics with current inflation and traditional inflation expectation measures, confirming the connection between current inflation, perceptions, and expectations due to common factors. Moreover, our perception indicators provide new information useful for forecasting inflation and expectations over the short term. Our perception indicators notably improve the forecasting of headline inflation and different baskets of inflation from four to six months ahead, reducing the RMSE by up to 28%. Larger gains are achieved between four and six months ahead. The forecasting ability of our indicators is also significant for inflation expectations from surveys of financial analysts and break-even inflation, with improvements of up to 30%. These combined facts show that our indicators are connected to current inflation and expectations, making them a good measure of inflation perceptions.

Acknowledgements

The authors appreciate the valuable comments of Ana Maria Iregui, Franz Hamann, and the participants at the XX research workshop of Banco de la República and the Economics Seminar at Banco Central del Uruguay on previous versions of the document. Any remaining errors are the exclusive responsibility of the authors.

References

Angelico, Cristina, Juri Marcucci, Marcello Miccoli, and Filippo Quarta (2022), "Can we measure inflation expectations using Twitter?" *Journal of Econometrics*, 228, 259–277.

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro (2014), "Using Social Media to Measure Labor Market Flows." NBER Working Papers 20010, National Bureau of Economic Research, Inc.

Bailliu, Jeannine, Xinfen Han, Mark Kruger, Yu-Hsien Liu, and Sri Thanabalasingam (2019), "Can media and text analytics provide insights into labour market conditions in China?" *International Journal of Forecasting*, 35, 1118–1130.

Becerra, Juan Sebastián and Andrés Sagner (2020), "Twitter-Based Economic Policy Uncertainty Index for Chile." Working Papers Central Bank of Chile 883, Central Bank of Chile.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003), "Latent dirichlet allocation." *Journal of machine Learning research*, 3, 993–1022.

Born, Benjamin, Hrishbh Dalal, Nora Lamersdorf, and Sascha Steffen (2023), "Monetary Policy in the Age of Social Media: A Twitter-Based Inflation Analysis." Technical report, Frankfurt School.

Bricongne, Jean-Charles, Olivier de Bandt, Annabelle de Gaye, Julien Denes, Paul Hubert, and Pierre-Antoine Robert (2022), "New indicators of perceived inflation in France based on media data." Technical report, Banque de France, URL https://blocnotesdeleco.banque-france.fr/en/blog-entry/new-indicators-perceived-inflation-france-based-media-data.

Casella, George and Edward I. George (1992), "Explaining the Gibbs Sampler." *The American Statistician*, 46, 167–174.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei (2009), "Reading tea leaves: How humans interpret topic models." *Advances in neural information processing systems*, 22.

Cicek, Serkan and Cuneyt Akar (2014), "Do Inflation Expectations Converge Toward Inflation Target or Actual Inflation? Evidence from Expectation Gap Persistence." *Central Bank Review*, 14, 15–21.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020), "Unsupervised Cross-lingual Representation Learning at Scale." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451, Association for Computational Linguistics, Online.

Datareportal (2022), "DIGITAL 2022: COLOMBIA." URL https://datareportal.com/reports/digital-2022-colombia.

Espinosa-Torres, Juan Andrés, Luis Fernando Melo-Velandia, and José Fernando Moreno-Gutiérrez (2017), "Expectativas de inflación, prima de riesgo inflacionario y prima de liquidez: una descomposición del break-even inflation para los bonos del Gobierno colombiano." *Revista Desarrollo y Sociedad*, 1, 315–365.

Feldkircher, Martin and Pierre L. Siklos (2019), "Global inflation dynamics and inflation expectations." *International Review of Economics & Finance*, 64, 217–241.

Fitchett, Hamish and Finn Robinson (2021), "Down to business: Which QSBO measures are the best at forecasting?" Reserve Bank of New Zealand Analytical Notes series AN2021/01, Reserve Bank of New Zealand.

Gabrielyan, Diana, Jaan Masso, and Lenno Uusküla (2020), "Mining News Data for the Measurement and Prediction of Inflation Expectations." In *Theory and Applications of Time Series Analysis* (Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, Héctor Pomares, and Ignacio Rojas, eds.), 253–271, Springer International Publishing, Cham.

Geman, Stuart and Donald Geman (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, 721–741.

Giacomini, Raffaella and Halbert White (2006), "Tests of Conditional Predictive Ability." *Econometrica*, 74, 1545–1578.

Glaeser, Edward L., Hyunjin Kim, and Michael Luca (2019), "Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity." In *Big Data for Twenty-First-Century Economic Statistics*, NBER Chapters, National Bureau of Economic Research, Inc.

González-Molano, Eliana R., Ramón Hernández-Ortega, Edgar Caicedo-García, Nicolás Martínez-Cortés, Jose Vicente Romero, and Anderson Grajales-Olarte (2020), "Nueva Clasificación del BANREP de la Canasta del IPC y revisión de las medidas de Inflación Básica en Colombia." Borradores de Economia 1122, Banco de la Republica de Colombia.

Grant, Alan P. and Lloyd B. Thomas (1999), "Inflationary expectations and rationality revisited." *Economics Letters*, 62, 331–338.

Guío-Martínez, Daniela Valentina (2020), "Descripción de las Minutas e Informes de Política Monetaria a partir de herramientas de Lingüística Computacional." Technical report.

Indaco, Agustín (2020), "From twitter to GDP: Estimating economic activity from social media." *Regional Science and Urban Economics*, 85.

Larsen, Vegard H., Leif Anders Thorsrud, and Julia Zhulanova (2021), "News-driven inflation expectations and information rigidities." *Journal of Monetary Economics*, 117, 507–520.

Li, Guowen, Xiaoqian Zhu, Jun Wang, Dengsheng Wu, and Jianping Li (2017), "Using LDA Model to Quantify and Visualize Textual Financial Stability Report." *Procedia Computer Science*, 122, 370–376. 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson (2006), "A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series." *Journal of Econometrics*, 135, 499–526.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013), "Distributed Representations of Words and Phrases and their Compositionality." In *Proceedings of NIPS*.

Pérez, Juan Manuel, Damián A Furman, Laura Alonso Alemany, and Franco Luque (2021), "Robertuito: a pre-trained language model for social media text in spanish." *arXiv preprint arXiv:2111.09453*.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015), "Exploring the Space of Topic Coherence Measures." In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, 399–408, Association for Computing Machinery, New York, NY, USA.

Salvatore, Camilla, Silvia Biffignandi, and Annamaria Bianchi (2021), "Social Media and Twitter Data Quality for New Social Indicators." *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*, 156, 601–630.

Vydra, Simon and Jaroslaw Kantorowicz (2021), "Tracing Policy-relevant Information in Social Media: The Case of Twitter before and during the COVID-19 Crisis." *Statistics, Politics and Policy*, 12, 87–127.

Weber, Michael, Bernardo Candia, Hassan Afrouzi, Tiziano Ropele, Rodrigo Lluberas, Serafin Frache, Brent H Meyer, Saten Kumar, Yuriy Gorodnichenko, Dimitris Georgarakos, Olivier Coibion, Geoff Kenny, and Jorge Ponce (2023), "Tell Me Something I Don't Already Know: Learning in Low and High-Inflation Settings." Working Paper 31485, National Bureau of Economic Research.

Xu, Yingying, Hsu-Ling Chang, Oana-Ramona Lobonţ, and Chi-Wei Su (2016), "Modeling heterogeneous inflation expectations: empirical evidence from demographic data?" *Economic Modelling*, 57, 153–163.

A. Terms related with Colombian tweets

Table A.1: *Denonyms that classify a user or tweet as Colombian*

alvaradense	amazonas	amazonense	amazonense
antioquena	antioqueno	antioquia	arauca
araucana	araucano	armenia	armerita
asisena	asiseno	atlanticense	atlantico
barrameja	barramejo	barranquilla	barranquillera
barranquillero	bellanita	belumbrense	bogota
bogotana	bogotano	bolivar	bolivarense
bonaverense	boyaca	boyacense	bucaramanga
buguena	bugueno	bumangues	bumanguesa
cachaca	cachaco	cafetera	cafetero
caldas	caldense	calena	caleno
cali	canasgordense	caqueta	caquetena
caqueteno	carmeluna	carmeluno	cartagena
cartagenera	cartagenero	cartagüena	cartagüeno
casanare	casanarena	casanareno	cauca
caucana	caucano	cesar	cesarense
choco	chocoana	chocoano	colombia
colombiana	colombiano	cordoba	cordobes
costena	costeno	cucuta	cucutena
cucuteno	cundinamarca	cundinamarques	cuyabra
cuyabro	dabeibana	dabeibano	dosquebradense
envigadena	envigadeno	envigado	escudores
escudores	espinaluna	espinaluno	facatativena
facatativeno	falanense	falanense	florencia
floridena	florideno	fresnense	fusagasuguena
fusagasugueno	girardotena	girardoteno	guainarense
guainia	guainiana	guainiano	guajira
guajiro	guamaluna	guamaluno	guaviare
guaviarense	herreruna	herreruno	hincha del verde
hincha de millonarios	hincha del rojo	huila	huilense
ibague	ibaguerena	ibaguereno	icononzuna
icononzuno	ipialena	ipialeno	itagui
itagüisena	itagüiseno	la guajira	laboyana
laboyano	magdalena	magdalenense	magdalenico
manizalena	manizaleno	manizales	mariquitena
mariquiteno	medellin	melgarense	meta
metense	mituana	mituano	mocoana
mocoano	monteria	monteriana	monteriano
narinense	narino	neiva	neivana
neivano	norte de santander	norte santandereana	norte santandereano
opita	paisa	paisana	paisano
palmira	palmirense	palocabildense	parce
parcera	parcero	pasquena	pasqueno
pasto	pastuso	payanes	pereira
pereirana	pereirano	petrista	piedecuestana
piedecuestano	popayan	portena	porteno
putumayense	putumayo	quibdo	quindiana
quindiano	quindio	remediana	remediano
riohacha	riohachera	riohachero	rionegrera
rionegrero	rionegro	risaralda	risaraldense

Table A.1: *Continued*

rola	rolo	rovirense	samaria
samario	san andres	san andresana	san andresano
santamarta	santander	santandereana	santandereano
santoto	santo tomas	sergio arboleda	sibatena
sibateno	sincelejana	sincelejano	sincelejo
soachuna	soachuno	soledena	soledeno
sucre	sucrena	sucreno	tolima
tolimense	toludena	toludeno	tuluena
tulueno	tumaquena	tumaqueno	tunja
tunjana	tunjano	umbitana	umbitano
universidad de los andes	universidad del bosque	universidad del rosario	universidad del valle
universidad eafit	universidad eai	universidad ean	universidad icesi
universidad javeriana	universidad nacional	universidad tadeo	urabaense
uribista	valduparense	valle del cauca	vallecaucana
vallecaucano	valledupar	valluna	valluno
vaupense	vaupes	velena	veleno
vichada	vichadense	villavicencio	villavicense
yaguarena	yaguareno	yopalena	yopaleno
yumbena	yumbeno		

B. Terms related with Other countries

Table B.1: *Denonyms for other countries*

argentina	argentino	argentina
buenos aires	porteno	portena
cordoba	cordobes	cordobesa
rosario	rosarino	rosarina
mendoza	mendocino	mendocina
tucuman	tucumano	tucumana
la plata	platense	
mar del plata	marplatense	
salta	salteno	saltena
santa fe	santafesino	santafesina
san juan	sanjuanino	sanjuanina
resistencia	resistenciano	resistenciana
santiago del estero	santiagueno	santiaguena
corrientes	correntino	correntina
posadas	posadeno	posadena
neuquen	neuquino	neuquina
formosa	formoseno	formosena
san salvador de jujuy	jujeno	jujena
bahia blanca	bahiense	
parana	paranaense	
concordia	concordiense	
bolivia	boliviano	boliviana
santa cruz de la sierra	cruceno	crucena
la paz	paceno	pacena
cochabamba	cochabambino	cochabambina
el alto	alteno	altena
sucre	sucrense	
oruro	orureno	orurena
tarija	tarijeno	tarijena
potosi	potosino	potosina
chile	chileno	chilena
santiago	santiaguino	santiaguina
puente alto	puentealtino	puentealtina
maipu	maipucino	maipucina
antofagasta	antofagastino	antofagastina
vina del mar	vinamarino	vinamarina
valparaiso	porteno	portena
talcahuano	talcahuanino	talcahuanina
san bernardo	sanbernardino	sanbernardina
temuco	temuquense	
concepcion	penquista	
rancagua	rancagüino	rancagüina
la pintana	pintanino	pintanina
la serena	serenense	
costa rica	costarricense	
san jose	josefino	josefina
puerto limon	limonense	
san francisco	franciscano	franciscana

Table B.1 *Continued*

alajuela	alajuelense	
liberia	liberiano	liberiana
san carlos	carlovense	
paraiso	paraisino	paraisina
cuba	cubano	cubana
la habana	habanero	habanera
santiago de cuba	santiaguero	santiaguera
camagüey	camagüeyano	camagüeyana
holguin	holguinero	holguinera
guantanamo	guantanamero	guantanamera
santa clara	clarano	clarana
las tunas	tunero	tunera
bayamo	bayames	bayamea
cienfuegos	cienfueguero	cienfueguera
pinar del rio	pinareno	pinarena
republica dominicana	dominicano	dominicana
santo domingo	santiaguero	santiaguera
santiago de los caballeros	santiaguero	santiaguera
santo domingo este	santiaguero	santiaguera
santo domingo oeste	santiaguero	santiaguera
santo domingo norte	santiaguero	santiaguera
san pedro de macoris	macorisano	macorisana
la romana	romano	romana
san cristobal	cristobalense	
puerto plata	platanero	platanera
san francisco de macoris	macorisano	macorisana
ecuador	ecuatoriano	ecuatoriana
guayaquil	guayaquileno	guayaquilena
quito	quiteno	quitena
cuenca	cuencano	cuencana
santo domingo de los colorados	santodomingueno	santodominguena
machala	machaleno	machalena
duran	duraneno	duranena
manta	manteno	mantena
portoviejo	portovejense	
ambato	ambateno	ambatena
esmeraldas	esmeraldeno	esmeraldena
el salvador	salvadoreno	salvadorena
san salvador	salvadoreno	salvadorena
santa ana	santaneco	santaneca
soyapango	soyapaneco	soyapaneca
san miguel	migueleno	miguelena
mejicanos	mejicanense	
santa tecla	teclense	
apopa	apopense	
delgado	delgadense	
sonsonate	sonsonateco	sonsonateca
san marcos	marquense	
espana	espanola	espanol

Table B.1 *Continued*

madrid	madrileno	madrilena
barcelona	barcelones	barcelonea
valencia	valenciano	valenciana
sevilla	sevillano	sevillana
zaragoza	zaragozano	zaragozana
malaga	malagueno	malaguena
murcia	murciano	murciana
palma de mallorca	palmesano	palmesana
las palmas de gran canaria	palmeno	palmena
bilbao	bilbaino	bilbaina
guatemala	guatemalteco	guatemalteca
ciudad de guatemala	guatemalteco	guatemalteca
mixco	mixqueno	mixquena
villa nueva	villanovense	
honduras	hondureno	hondurena
tegucigalpa	hondureno	hondurena
san pedro sula	sampedrano	sampedrana
mexico	mexicano	mexicana
ciudad de mexico	mexicano	mexicana
guadalajara	tapatio	tapatia
monterrey	regiomontano	regiomontana
puebla	poblano	poblana
tijuana	tijuanense	
ciudad juarez	juarense	
leon	leones	leonea
zapopan	zapopano	zapopana
nicaragua	nicaragüense	
managua	nicaragüense	
panama	panameno	panamena
ciudad de panama	panameno	panamena
san miguelito	sanmigueliteno	sanmigueliteña
paraguay	paraguayo	paraguaya
asuncion	paraguayo	paraguaya
ciudad del este	esteno	estena
peru	peruano	peruana
lima	peruano	peruana
arequipa	arequipeno	arequipena
trujillo	trujillano	trujillana
chiclayo	chiclayano	chiclayana
piura	piurano	piurana
iquitos	iquiteno	iquitena
cusco	cusqueno	cusquena
chimbote	chimboteno	chimbotena
huancayo	huancaino	huancaina
tacna	tacneno	tacnena
puerto rico	puertorriqueno	puertorriquena

Table B.1 *Continued*

san juan	puertorriqueno	puertorriquena
bayamon	bayamones	bayamonea
uruguay	uruguayo	uruguaya
montevideo	uruguayo	uruguaya
venezuela	venezolano	venezolana
caracas	venezolano	venezolana
maracaibo	marabino	marabina
valencia	valenciano	valenciana
barquisimeto	barquisimetano	barquisimetana
maracay	maracayero	maracayera
ciudad guayana	guayanes	guayanea

C. Terms related with inflation and deflation

Table C.1: *N-grams for Downward Signal*

al mejor precio	costos bajando	poco paga
baja costo	costos caen	pocos los precios
baja el costo	costos cayendo	precio a la mitad
baja el precio	costos disminuyen	precio absequible
baja en el costo	costos menores	precio baja
baja en el precio	deflacion del salario	precio bajando
baja en los costos	deflacion en el precio	precio bajara
baja inflación	deflacion en los precios	precio bajo
baja la inflacion	deflacion en los salarios	precio barato
baja precio	demasiado barato	precio cae
bajada de precio	demasiado baratos	precio caera
bajada de precios	descuento en el precio	precio cayendo
bajan costos	descuento en los precios	precio de ganga
bajan los costos	desploma el petroleo	precio de huevo
bajan los precios	desploma el precio	precio descontado
bajando el costo	desploman los precios	precio diminuto
bajando el precio	desplome en el costo	precio disminuyendo
bajando los costos	desplome en el precio	precio especial
bajando los precios	desplome en los costos	precio favorito
bajar el costo	desplome en los precios	precio inferior
bajar el precio	disminuye costo	precio inmejorable
bajar la inflacion	disminuye el costo	precio justo
bajar los precios	disminuye el petroleo	precio mas bajo
bajar precios	disminuye el salario	precio mas competitivo
bajara el precio	disminuyen los costos	precio minimo
bajaran los precios	el mejor precios	precio razonable
bajaran los salarios	increible precio	precio reducido
bajaron de costo	inferior precio	precio va a caer
bajaron de precio	inferiores precios	precios a la baja
bajaron los costos	la deflacion golpea	precios bajan
baje el precio	los mejores precios	precios bajando
bajen los precios	mas barato	precios bajaran
bajo costo	mas baratos	precios bajos
bajo costo	mejor oferta	precios baratos
bajo el precio	mejor precio	precios caen
bajo precio	mejores precios	precios caeran
bajos costos	menor costo	precios cayendo
bajos los precios	menor inflacion	precios de ganga
bajos precios	menor precio	precios de oferta
barato el costo	menores costos	precios descontados
barato el precio	menores costos	precios diminutos
baratos los costos	menores precios	precios disminuyen
baratos los precios	menos caro	precios disminuyendo
buen precio	menos caros	precios inferiores
buena oferta	menos costo	precios inigualables
buenos precios	menos costoso	precios mas bajos
cae el costo	menos costosos	precios minimos
cae el precio	menos inflacion	precios razonables
caen los costos	minimo precio	rebaja de precio
caera el precio	minimos precios	rebajas de precio
caeran los precios	mitad de precio	reduccion en el precio

Table C.1: *Continued*

caidas en el precio	mitad de precios	reducen el precio
caidas en los precios	muy barato	reducen los costos
costo bajo	muy buen precio	reducen los precios
costo barato	muy buenos precios	reduciendo el precio
costo disminuye	no es caro	reduciendo los costos
costos a la baja	no es costoso	reduciendo los precios
costos bajan	oferta en los precios	reducir el precio
poco el precio	paga poco	salario a la baja
una ganga	poco costoso	salarios a la baja
se llama deflacion	disminuye la base monetaria	contraccion monetaria
disminuye la oferta monetaria	contrae la oferta monetaria	contraer base monetaria
contraer oferta monetaria	baja el precio de la energia	baja la energia
baja el costo de la energia	bajan los costos el costo de la energia	baja el precio de la electricidad
baja la electricidad	baja el costo de la electricidad	baja el arriendo
bajando el arriendo	bajan los alquileres	baja el alquiler
menor arriendo	menor alquiler	

Table C.2: *N-grams for Upward Signal*

absurdo el precio	elevado pago	por las nubes
absurdo los precios	elevado precio	precio absurdo
abusan con el precio	elevados los precios	precio abusivo
abusan con los precios	elevados pagos	precio al triple
abusivo precio	elevados precios	precio altisimo
abusivos precios	elevan los precios	precio alto
abusos precios	elevan precios	precio aumentara
acelera la inflacion	elevando el precio	precio caro
altas inflaciones	elevando los precios	precio crece
altisimo precio	elevando precio	precio creciente
altisimos precios	elevando precios	precio de oro
alto costo	escalada de los precios	precio disparado
alto el precio	escalada de precio	precio elevado
alto pago	escalada de precios	precio es altisimo
alto precio	escalada del precio	precio exorbitante
altos costos	expectativa de inflacion	precio impagable
altos los precios	explosion de los precios	precio inflado
altos pagos	factura cara	precio mas alto
altos precios	factura costosa	precio miserable
alza de los precios	facturas caras	precio muy alto
alza de precios	facturas costosas	precio nunca visto
alza del precio	fuerte alza	precio record
alza el precio	fuerte alzas	precio sube
alza en el precio	fuerte creciemiento	precio subira
alza en los precios	fuerte inflacion	precio tan alto
alzan los precios	gasolina cara	precio va a subir
aumenta el costo	gasolina costosa	precio va aumentando
aumenta inflacion	gasolina impagable	precios a la alza
aumenta precio	guerra a la inflacion	precios absurdos
aumentado el precio	guerra contra la inflacion	precios abusivos
aumentado los precios	hiper inflacion	precios al doble
aumentan los costos	hiperinflacion	precios al triple
aumentan precios	increiblemente caro	precios altisimos
aumentando precio	increiblemente caros	precios altos
aumentara precio	incremento de los precios	precios arriba
aumentaron precios	incremento de precios	precios aumentaran
aumento de la inflacion	incremento del costo	precios caros
aumento de los precios	incremento en el precio	precios crecen
aumento de precios	incrementos en los precios	precios crecientes
aumento del costo	inflación alta	precios disparados
aumento del precio	inflacion aumenta	precios elevados
aumento del salario	inflacion crece	precios están subiendo
aumento en el costo	inflacion creciente	precios estan super inflados
aumento en el precio	inflacion de dos digitos	precios exorbitantes
aumento en los precios	inflacion dura	precios impagables
aumento precio	inflacion elevada	precios mas altos
aumento precios	inflacion elevado	precios mas caros
aumentos de la gasolina	inflacion es un problema	precios muy altos
aumentos de los precios	inflacion fuerte	precios por las núbes
aumentos del petroleo	inflacion mas alta	precios record
aumentos del precio	inflacion muy alta	precios son altisimos
bajar el precio	inflacion no perdona	precios suben
cara factura	inflacion perdura	precios subiendo
cara la gasolina	inflacion persiste	precios subiran

Table C.2: *Continued*

caras facturas	inflacion se acelera	precios tan altos
caro el combustible	inflacion se dispara	precios tan altos
caro el precio	inflaciones altas	precios van a subir
caro pago	inflaciones muy altas	precios van arriba
caro precio	inflan el precio	precios van aumentando
caro recibo	inflan los precios	re caro
caros los precios	inflan precios	recibo caro
caros pagos	inflaran precio	recibos caros
caros precios	inflaran precios	repunte en costos
caros recibos	lo caro que esta	sale mas barato
combustible caro	lo caro que estan	sale mas caro
combustibles caros	los precios caros	sorprendentemente caro
como esta de caro	lucha contra inflacion	sorprendentemente caros
costo abismal	lucha contra la inflacion	suba de precios
costo impagable	mas caro	sube el costo
costo tan elevado	mas caros	sube el precio
costos crecientes	mas costoso	sube precio
costos impagables	mas costosos	suben los costos
costos por las nubes	mas han encarecido	suben los precios
costosa factura	mayor costo	suben precios
costosas facturas	mayor inflacion	subida de los precios
crea inflacion	mayor precio	subida de precios
crea la inflacion	mayores costos	subida del precio
crece el costo	mayores precios	subido de precio
crece el precio	miserable precio	subido el precio
crece inflacion	miserables precios	subido los precios
crece la inflacion	mucho mas caro	subiendo el precio
crecen el precio	mucho mas caros	subiendo los precios
crecen los costos	mucho mas costoso	subieron de precio
crecen los precios	mucho mas costosos	subir el precio
creciente inflacion	muchos mas caros	subir los precios
crecimiento fuerte	muy cara	subira precio
cuesta mucho	muy caro	subiran precios
culpa de la inflacion	muy costosa	subirnos el precio
demasiado caro	muy costoso	subirnos los precios
demasiado costoso	nace inflacion	super inflacion
demasiados caros	nace la inflacion	tan caro
demasiados costosos	pagan caro	todo esta caro
dispara costo	pagando caro	triple del precio
dispara precio	pagara caro	triplica precio
disparan costos	pago alto	triplicado su precio
disparan precios	pago caro	triplicado sus precios
doble de precio	pago elevado	un aumento
duplicado su precio	pagos altos	un crecimiento
duplicado sus precios	pagos caros	un precio altisimo
dura la inflacion	pagos elevados	unos aumentos
el precio caro	peor inflacion	unos crecimientos
eleva el precio	peor precio	veces el precio
eleva precio	peores precios	persiste la inflacion
elevado el precio	perdura la inflacion	expansion monetaria
aumenta la oferta monetaria	aumenta la base monetaria	aumenta oferta monetaria
expande la oferta monetaria	expandir base monetaria	expandir oferta monetaria
sube el precio de la energia	sube el costo de la energia	sube la energia
suben los costos de la energia	sube el precio de la luz	sube el costo de la luz

Table C.2: *Continued*

luz mas costosa	luz mas cara	luz impagable
sube la electricidad	sube el arriendo	suben los arriendos
arriendo sin techo	mayores arriendor	alquiler mas caro

D. Technical appendix

In this appendix, we discuss the machine learning techniques employed in the classification of topics using LDA and the construction of the Word to VEC indicator. Our analysis begins with the consideration that each tweet classified as Colombian represents a document d, for $d = 1, ..., D$. Collectively, these documents constitute a corpus C. Subsequently, we processed this corpus using conventional natural language processing (NLP) techniques to extract price signals. First, we remove stopwords, which are commonly used words that typically do not have a significant meaning. Second, to reduce the vocabulary size and enhance consistency, we performed lemmatization, which involves reducing words to their base or dictionary form, known as a lemma. A lemma represents the canonical form of a word that captures its meaning, eliminating inflectional or derivational changes, such as verb tense and pluralization, among other grammatical variations. Subsequently, we define the set of all words appearing in the corpus, which we define as V ("vocabulary"), and create the document term matrix (DTM), which relates the vocabulary to each document.

D.1. Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) algorithm was introduced by Blei et al. (2003) and its main objective is to uncover hidden topics within a corpus by breaking down the relationship between documents and words, making it easier to understand the context of a large corpus. In figure D1 we provide a visual explanation of the mechanics of the algorithm that decomposes the document term matrix (DTM), size (D,V), into two key matrices:

Document - Topic Matrix: This matrix shows the probability that each document belongs to a specific topic. This provides insights into the extent to which each document is related to different topics in the corpus, its size is (D,K)

Topic - Term Matrix: This matrix highlights the relationship between topics and the words in the vocabulary. It identifies the most important words for each topic and reveals what each topic is about, its size is (K,V).

In this setup, the words in each document $w_{i,j}$ are known for $i = 1, ..., D$, $j = 1, ..., V$, while the K topics are latent variables. The estimation is made through the Gibbs sampling algorithm introduced by Geman and Geman (1984), which requires some recursive assumptions that are updated in each iteration. Thus, the estimation proceeds as follows.[11]

•

i) For any document in a corpus choose the number of words N $\sim$ Poisson(ξ).

ii) choose the Dirichlet distribution of the document-topic matrix, $\Theta \sim$ Dir(α).

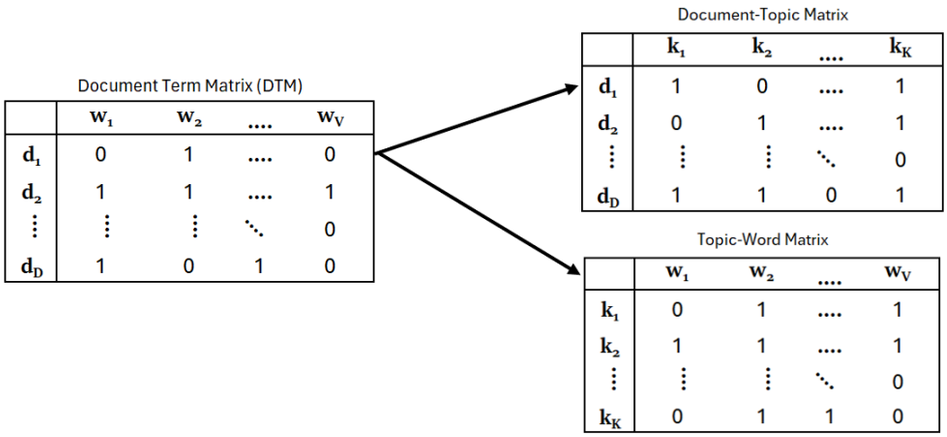
iii) initialize the Topic - Word matrix using a Dirichlet distribution $\beta_{i,j} = p(w^j = 1|z^i = 1)$

iv) For each of the N words w_n, select a topic (z_n) $\sim$ Multinomial (Θ).

v) choose w_n from p($w_n|z_n, \beta$), a multinomial probability conditioned on the topic z_n.

[11] A more detailed explanation of the Gibbs Sampler is found in Casella and George (1992).

Figure D1: LDA representation



This figure shows how the document term matrix is decomposed in the LDA algorithm.

A K-dimensional Dirichlet random variable Θ can thus take values in the (k-1) - simplex (a k-vector Θ lies in the (k-1)-simplex if $\Theta_i \geq 0$ and $\sum_{i=1}^{k} \theta_i = 1$) and has the following probability density on the simplex:

$$P(\Theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \Theta_1^{\alpha_1 - 1} \cdots \Theta_k^{\alpha_k - 1}$$

Where the parameter α is a K-vector with components $\alpha_i > 0$, and where $\Gamma(x)$ is the Gamma function. Given the parameters α and β, the joint distribution of a topic mixture Θ, a set of N topics **z**, and a set of N words **w** is given by:

$$P(\Theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = P(z_n|\Theta)P(w_n|z_n, \beta)$$

Where $P(z_n|\Theta)$ is simply Θ_i for the unique i such that $z_n^i = 1$. Integrating over Θ and summing over z, we obtain the marginal distribution of a document:

$$P(\mathbf{w}|\alpha, \beta) = \int P(\Theta|\alpha)(\prod_{n=1}^{N} \sum_{z_n} P(Z_n|\Theta)P(w_n|z_n, \beta))d\Theta$$

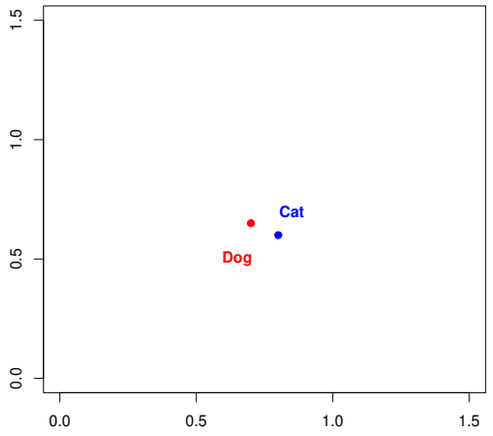
Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$P(D|\alpha, \beta) = \prod_{d=1}^{M} \int P(\Theta_d|\alpha)(\prod_{n=1}^{N_d} \sum_{z_{dn}} P(Z_{dn}|\Theta_d)P(w_{dn}|z_{dn}, \beta))d\Theta_d \tag{D.1}$$

vi) With the computed probabilities from equation D.1, of the document-topic and topic-word matrices the posterior probability that a word belongs to a particular topic is computed and the Topic is reassigned.

The process described in steps i) - vi) is repeated iteratively until convergence is reached, at which point the final distributions and output topics are estimated.

Figure D2: Two dimensional representation of Word Embeddings



This figure shows the two dimensional representation of the embeddings of the words cat and dog.

D.2. Mathematical Details of Word2Vec Methodology

In this methodology, the elements of the DTM are transformed into word embeddings, which are real-valued vectors in a high-dimensional space, usually between 200 and 300 dimensions, and are generated after training in a very large corpus. For example, the word cat can be represented as $[0.1823, 0.4839, -0.7291, ..., 0.1048]$, while dog will be $[0.2312, 0.4831, -0.6721, ..., 0.8765]$ which for illustrative purposes can be reduce to two dimensions vectors as [0.8, 0.6] and [0.7, 0.65], and portrait in figure D2.[12]

Due the word-embedding representation of our vocabulary, it is possible to compare how close semantically speaking are the words in our corpus by using similarity measures between two word vectors. The most commonly used metric for this endeavor is the cosine similarity, which is computed as:

$$\text{similarity}_{A,B} = \cos\theta = \frac{A \cdot B}{\|A\|\|B\|}$$

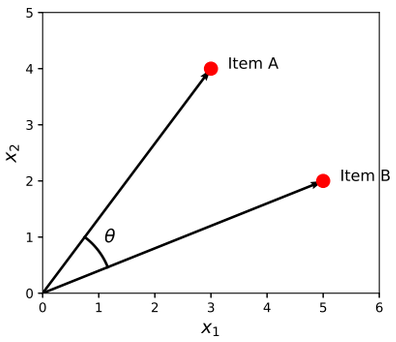
where A and B are the vector representations of the two words, $A \cdot B$ is the dot product of these vectors, and $\|A\|$ and $\|B\|$ are the norms of the vectors, as shown in Figure D3. A cosine similarity close to one indicates that the vectors are similar, whereas a cosine similarity close to zero indicates that they are orthogonal or unrelated.

In the construction of our W2V indicator, we focus on finding the closest terms to the center word that relate to price increases or decreases. For a given pair of center and context, that is, surrounding words within a certain window size for a target word, Word2Vec adjusts their vector representations such that words that frequently co-occur in similar contexts are closer in the vector space. We use the **Skip-Gram** algorithm of the W2V to find the words related to a given center for our price signal. Thus, the objective is to predict the surrounding context words $(w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2})$ given a target word or center w_t as depicted in Figure D4.

Mathematically, the algorithm maximizes the conditional probability of context words given a target word:

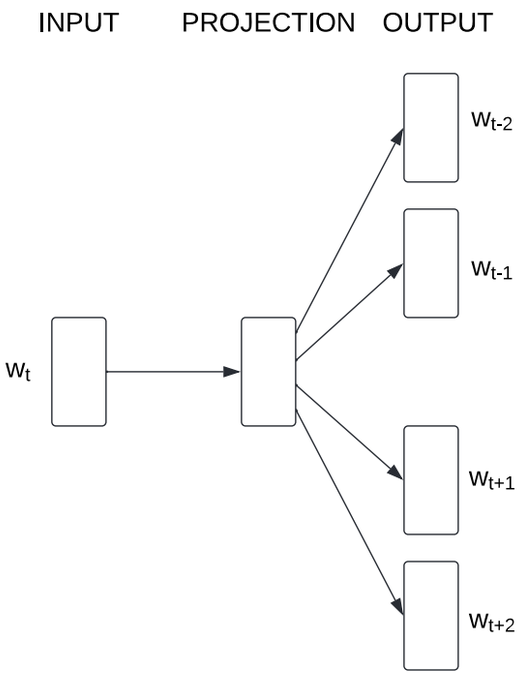
[12]However, this reduction of dimensionality lose most of the detailed information encoded in higher dimensions.

Figure D3: *Cosine Similarity*



This figure shows how the cosine similarity compares two vectors of word embeddings.

Figure D4: *Skip Gram algorithm*



This figure shows the logic behind the skip-gram model, in which we provide the center W_t and the algorithm projects to words in the surrounding context.

$$P(w_{t-n}, w_{t-n+1}, \ldots, w_{t-1}, w_{t+1}, w_{t+2}, \ldots, w_{t+n} \mid w_t)$$

Given a target word w_t, the skip-gram model predicts the surrounding words within a context window of size n. This is formalized as:

$$\mathcal{L} = \sum_{t=1}^{T} \sum_{-n \leq j \leq n, j \neq 0} \log P(w_{t+j} \mid w_t)$$

Where T represents all tokens in the corpus, $\mathcal{L}$ is the overall objective function that maximizes the likelihood of correctly predicting context words given the target word. Let us define v as the vector sum of a given word and its context.

$$v := \sum_{j \in N+i} w_j$$

Then, we take the dot-product-softmax with every other vector sum (this step is similar to the attention mechanism in transformers) to obtain the probability:

$$P(w_i | w_j : j \in n+i) := \frac{e^{v_w \cdot v}}{\sum_{v' \in V} e^{v' \cdot v}}$$

Using the softmax function, we transform the plain numbers of our output vector v into a probability distribution p. Thus, we can easily extract words with the highest probability of being in context for w_t. The quantity to be maximized is then, after simplification,

$$\sum_{i \in C} \sum_{j \in N+i} \left(v_{w_i} \cdot v_{w_j} - \ln \sum_{v' \in V} e^{v' \cdot v_{w_j}} \right)$$

The quantity on the left is fast to compute, but the quantity on the right is slow because it involves summing over the entire vocabulary set for each word in the corpus. Furthermore, the use of gradient ascent to maximize log-probability requires computing the gradient of the quantity on the right, which is intractable. Calculating the softmax function across an entire vocabulary is computationally expensive. To address this, Word2Vec employs two techniques:

Hierarchical Softmax: Instead of calculating the probability of a word across the entire vocabulary, the model uses a binary tree representation of the vocabulary. This reduces the computational complexity of computing the softmax function.

Negative Sampling: This approach modifies the objective by simplifying the softmax function. Instead of updating all words in the vocabulary, only a few "negative" samples (words not in the context) are selected for updating. The modified objective function for negative sampling is

$$\log P(w_o \mid w_t) + \sum_{i=1}^{k} \log P(w_i \mid w_t)$$

where w_o is the true context word and w_i is the k-negative samples (randomly chosen words).