Forecasting inflation: from twitter to mass media

J. Daniel, Aromí*
Martín, Llada†

August 30, 2023

Abstract

We use mass media news to generate a set of quantitative indexes related to the inflation rate in Argentina for the period 2017-2023. We distinguish the information content of different features of mass media news. First, an indicator of attention allocated to inflation is generated. Second, a set of indexes which captures signals about inflation direction is built. In-sample and out-of-sample exercises confirm that using mass media content leads to gains in forecast accuracy. We find that indexes that capture inflation direction achieve the strongest accuracy gains. The proposed indicators compare favorably with traditional macroeconomic indicators and those based on social media. Combining mass media and social media leads to interesting information gains. Furthermore, the forecast accuracy of indexes based on NLP techniques outperforms traditional approaches.

Keywords: inflation, mass media, social network, forecasting, text analysis
JEL Classification: E31, E37, E70

1 Introduction

The COVID-19 pandemic caused by the SARS-CoV-2 virus and the Russia-Ukraine conflict have triggered a rise in the inflation rate in many countries, including Argentina, a country with a chronic inflationary process lasting more than ten years. Given this uncertain context, the need for an early measure of inflation rate becomes highly relevant. In this paper, we develop a measure of how mass media cover the inflation topic, which anticipated the agents' inflation expectation behavior (Carroll, 2003; Lamla and Lein, 2008; Maag and Lamla, 2009). This information could be useful from a policy maker perspective, which provides a timely signal about the future inflation dynamic.

*IIEP-BAIRES (UBA-CONICET), Universidad de Buenos Aires, and Universidad Católica Argentina. aromi.daniel@gmail.com
†IIEP-BAIRES (UBA-CONICET) and Universidad de Alcalá. lladamartin@gmail.com

In this study, we implement an analysis of mass media content and inflation for the case of Argentina. Given that many mass media reporters do not have a repository, we collect the tweets have published in Twitter by a set of prominent newspaper of Argentina. A dataset of economic news is generated, which covers the period 2017-2022. Given a corpus of more than 240 thousand of news, we build a set of quantitative indicators that capture not only the volume of news regarding inflation, but also the frequency of statements about the increasing vs. decreasing price changes and the temporal orientation of those claims.

The empirical evidence indicates mass media content anticipates macroeconomic outcomes. More specifically, a simple indicator of the level of attention allocated to inflation provides valuable information regarding inflation levels. Also, our evidence shows that more precise indexes have useful information about inflation dynamic. In this sense, we build an index that captures whether the news denotes increasing or decreasing inflation, and an index that approximates whether the news is related to backward or forward orientation. Estimated forecasting models indicate that this set of indexes based on mass media anticipate expected inflation. Out-of-sample forecasts confirm that regularity, which provides evidence that mass media indicators allow for significant gains in forecast accuracy. More specifically, forecast models that incorporate regressors that approximate the frequency of statements about increasing inflation result in the strongest accuracy gains.

The estimated information gains provided by most of the mass media indexes compare favorably with the information provided by lagged inflation, the lagged devaluation rate, and the lagged inflation forecast published by professional forecasters. Also, analyses show that these information gains are substantial not only when we combined the set of indexes based on mass media, but also when we consider forecast combinations. Furthermore, the results not only persist but also evidence improvement when indexes are computed using a Natural Language Inference model (MacCartney and Manning, 2008), a technique in the field of Natural Language Processing that is less explored in an economic context.

This work contributes to three main strands of the literature. First, it is linked to a growing body of research that uses unstructured information to describe dynamics in macroeconomic settings (Tetlock, 2007; Baker et al., 2016; Thorsrud, 2016, 2020; Larsen et al., 2021; Altig et al., 2020). For example, there are studies which show that measures based on mass media content predict inflation phenomena. Carroll (2003) proposed an inflation attention index based on news published by newspaper the New York Times and the Washington Post. The author provides evidence that household inflation expectations are more accurate when there is more news about inflation, which also triggers an updating of their beliefs. Lamla and Lein (2008), in line with Carroll (2003), find an positive association between accuracy of expectations and a set of indexes based on German mass media news, which approximate the intensity of

reporting and the tone of news about inflation rate. Maag and Lamla (2009), show that indexes based on German news about inflation rate play a relevant role to explain for disagreement of consumers inflation expectations, but not for disagreement of professionals.

On the other hand, there is a set of studies that exploit social media content to understand the inflation dynamic. In this senses, Angelico et al. (2022) develop a set of Twitter-based inflation expectation indexes through combining a dictionary approach and unsupervised machine learning techniques. They evidence that twitter content has valuable information about Italian consumer inflation expectation. Denes et al. (2021) exploit an dictionary approach and supervised machine leaning techniques in order to generate an index that approximates French consumer inflation expectation. The authors evidence that their index has a high correlation with perceived inflation measures based on surveys and the inflation rate. Aromí and Llada (ming) develop an attention inflation index of Argentina based on Twitter data. The authors evidence that this index has predictive power for the future inflation rate of Argentina beyond what could be predicted based on traditional indicators.

In this sense, the present work contributes to this growing literature based on a set of exercises that show the usefulness of incorporating indicators based on unstructured information. In particular, this work confirms that mass media data constitutes a particularly valuable source of information regarding future inflation. Furthermore, we evidence stronger accuracy gains when mass media and Twitter indexes are combined. These results are useful given that the advantage of such data is that they are available at high frequency and in real time, but also because we provide evidence for a developing country.

Second, our work is also a methodological contribution on how to combine Twitter content and mass media news and use this corpus together with natural language processing techniques to extract meaningful information on macroeconomic variables. In this work, we combine a dictionary with a semantic approach to extract a direction and time orientation signal of inflation rate. The regularities persist when we employ an alternative classifier method.

Third, this paper is aligned with a literature that provide evidence in favor of information rigidities rather than full-information rational expectation (Andrade and Le Bihan, 2013; Coibion and Gorodnichenko, 2015; Fuhrer, 2018). In this sense, economic agents are rational inattentive given that that updating and processing of information is costly (Sims, 2003). Considering the latter, agents rely on certain common sources not only to reduce costs of information acquisition, but also because they do not have the resources to monitor all the events potentially relevant for their decision. The current study contributes to this strand of research providing evidence that common sources of economic information as mass media news contain valuable to understand macroeconomic outcomes.

The rest of the paper is organized as follows. Section 2 describes the data and the methodology. Section 3 presents the results and Section 4 concludes.

2 Data & Methodology

As previously mentioned, we want to evaluate if quantitative indices based on mass media newspapers contain valuable information regarding the evolution of inflation and whether these data can be used to forecast inflation in Argentina. To analyze this issue, we need data of traditional indicators (such as past inflation and exchange rates) as well as a measure for the intensity and content of reporting on inflation in the media in a given period. The sample period is January 2017 through December 2022.

2.1 Mass media data: from tweets to newspaper articles

For media data we rely on tweets published by a set of prominent mass media of Argentina publishes. Taking into account that most of media press do not have a digital archive, Twitter offers an important opportunity to recover their historic publications. Mass media uses Twitter as alternative channel to spread their news article. In general, a tweet could contain a brief text of the news article, an image, a video, and/or the URL (Uniform Resource Locators) of the news article.

In a first step, we use the Twitter API for Academic Research in order to collect the tweets published by each media press during the sample period. The data comprises articles published by A24 (its twitter account is @A24COM), Ámbito Financiero (@Ambitocom), Clarín (@clarincom), El Cronista (@Cronistacom), Infobae (@infobae), La Nación (@LANACION), Página 12 (@pagina12) and Perfil (@perfilcom). Given the corpus of tweets, in the next step we extract the URL from each tweet. Once we have the link of each article, we collect, if available, the title, the subtitle and the body of each news published in the Economic section. From this corpus a set of quantitative indicators are generated.

We build a news inflation attention index ($AttentionIndex$), which is equal to the number of inflation articles published within in a month. This indicator of attention is built computing the frequency of articles where appears at least one time the noun "inflation" or the adjective "inflationary"[1][2]. More specifically,

[1] In Spanish, inflación, and inflacionario/a/s.

[2] The results are robust whether extend the list of terms related to inflation. On the one hand, we stretch out the list to incorporate the terms "prices", "cpi", "cost of living" ("precios", "ipc", "costo de vida" in Spanish). On the other hand, we extend the last list in order to incorporate the term "dollar" ("dólar" and "dolar" in Spanish).

let i_t represents the number of articles in which a keyword is detected in articles corresponding to month t and n_t represents the total number of articles corresponding to month t. Then, the corresponding value of the inflation attention index is given by $AttentionIndex_t = i_t/n_t$. A higher number is interpreted as more attention being allocated to inflation by mass media during a given month.

Also, we construct an indicator that captures the content of the news. We build an inflation news direction index which considers the number of sentences in an article regarding regarding both rising and falling inflation. A sentence is about inflation if contains at least one time the noun "inflation" or the adjective "inflationary". Once we have identified the sentences about inflation, we can identify the direction of these sentences. A sentence is about rising inflation if contains at least one time a word in the following list: expensive, costly, prohibitive, high, exorbitant, affordable, inaccessible, excessive, abnormal, scam, ruinous, scandalous, out of reach, inconceivable, growth, more expensive, rise, increase, positive, boom, higher[3]. On the other hand, a sentence is considered to to be discussing falling inflation if it contains at least one word from the following list: low, modest, advantageous, discounted, unbeatable, derisory, attractive, bargain, bargain price, affordable, reasonable, competitive, accessible, acceptable, normal, fair, interesting, suitable, negligible, less expensive, decrease, negative, minor[4]. More specifically, let s_t^r represents the sum of sentences regarding rising inflation corresponding to month t, while s_t^f represents the sum of sentences regarding falling inflation in month t. Since these definitions, we build a set of quantitative indicators to capture the content direction of news stories. The inflation news direction index is given by $DirectionIndex_t = (s_t^r - s_t^f)/s_t$, where s_t represents the total number of sentences corresponding to month t. A positive number is interpreted as more attention being allocated to rising inflation by mass media during a given month, while a negative number indicates more attention to falling inflation. The increasing inflation direction index is given by $IncreasingDirection_t = s_t^r/s_t$, while the decreasing inflation direction index is given by $DecreasingDirection_t = s_t^f/s_t$. Lastly, a neutral inflation direction index ($NeutralDirection_t$) is generated. This index captures the statements that do not contain information regarding news on rising or falling inflation.

Additionally, a set of temporal orientation indexes are built. These indicator capture not only the time structure of the article, i.e. whether the news is related to past or future inflation dynamics, but also the direction of the news (i.e. increasing or descending). We follow a similar strategy as we previously

[3]In Spanish, caro, costoso, prohibitivo, alto, exorbitante, inasequible, inaccesible, excesivo, anormal, estafa, ruinoso, escandaloso, fuera de alcance, inconcebible, crecimiento, más caro, cara, más cara, suba, alza, positiva, aumento, auge, mayor.

[4]In Spanish, bajo, modesto, ventajoso, descontado, imbatible, irrisorio, atractivo, oferta, precio de oferta, atractiva, promoción, asequible, razonable, competitivo, accesible, aceptable, normal, justo, interesante, adecuado, insignificante, menos caro, baja, caída, negativa, disminución, descenso, merma, menor.

presented. Once we have identified the sentences related to inflation, we count the sentences that talk about past ($s_t^{backward}$) or future inflation ($s_t^{forward}$). In this regard, we utilize a list of over 27 thousand words that encompass past tense forms of verbs, and similarly, 14 thousand words for future tense forms of verbs. Next, we identify the sentences related to rising and falling inflation in those sentences related with past and future. Then, the corresponding value of the news backward inflation index is given by $BackwardIndex_t = s_t^{backward}/s_t$, while the news forward inflation index is given by $ForwardIndex_t = s_t^{forward}/s_t$.

Also, we build a orientation index for each temporal dimension proposed. In this sense, the backward inflation news direction index is given by $BackDir_t = (s_t^{IncrBack} - s_t^{DecrBack})/s_t$, where $s_t^{IncrBack}$ represents the count of sentences that talk about past and increasing inflation, while $s_t^{DecrBack}$ represents the count of sentences that talk about past and decreasing inflation dynamics. In a similar way, we build a forward inflation news direction index is given by $ForwDir_t = (s_t^{IncrForw} - s_t^{DecrForw})/s_t$. A higher value of these indexes is interpreted as more attention being allocated to rising inflation during a given month and related on a specific time dimension, while a negative number indicates more attention to falling inflation.

2.2 Traditional indicators

We use a set of data given by the consumer price index, the exchange rate and the inflation professional forecasters' inflation expectations. The consumer price index data is from the National Institute of Statistics and Census (INDEC)[5]. The Argentine peso-US dollar exchange rate time series is from the Argentine Central Bank (Banco Central de la República Argentina)[6]. We compute the inflation rate on month t (Δcpi_t) as the monthly variation: $\Delta cpi_t = cpi_t/cpi_{t-1}–1$. In turn, the peso-dollar exchange rate is given a monthly variation computed as the log-difference between month t and month $t-1$ values: $\Delta er_t = log(er_t) - log(er_{t-1})$. Lastly, Professional forecasts correspond to those collected in the Relevamiento de Expectativas de Mercado (REM), a survey produced by the Argentine Central Bank. The first release of the survey corresponds to June 2016. We use median monthly inflation forecasts for one-month-ahead forecasts. Consider $F_{t-1}\Delta cpi_t$ to represent the one-month-ahead forecast at time $t-1$ of the inflation rate at time t.

2.3 Descriptive statistics

Table 1 shows descriptive statistics for the inflation rate, the devaluation of the exchange rate, the professional forecasters inflation expectations, and the news

[5]The data are from INDEC. `https://www.indec.gob.ar/indec/web/Nivel3-Tema-3-5`.

[6]We use the monthly average Wholesale Foreign Exchange Rate (ARS/USD) Com. A 3500 exchange rate. `http://www.bcra.gov.ar/`.

inflation attention index. The table shows that the average monthly inflation rate for the period was 3.4%. The period was characterized by high volatility as indicated by the standard deviation of, approximately, 1.5%. The average monthly devaluation rate for the sample period is approximately 3.3%. This period was also volatile in terms of the foreign exchange rate with a standard deviation of approximately 5%, a maximum value of 25% and a minimum of -4%. The professional forecasters seem to report an expectation that was reasonably align to the target variable during the analyzed period.

According to the inflation attention news indexes, the inflation topic is covered, on average, by 27% of articles published by mass media during the analyzed period in the economic section. The maximum value of the index, 45%, corresponds to May 2022. This is a month of high volatility in the inflation rate, whose standard deviation along the first four months of 2022 was 1.3%. During this period began the conflict between Ukraine and Russia. The minimum value of the index corresponds to March 2020. During this month, the government of Argentina confirmed its first SARS-CoV-2 virus cases and deaths and announced a nationwide lockdown.

Concerning the variables that exploit the content of news that cover the inflation topic, we see that articles regarding rising inflation are more frequent than those articles regarding decreasing inflation. On the other hand, the set of indices that capture the temporal structure of the news suggests that backward orientation is more frequent than forecast orientation along the analyzed period.

Table 1: Descriptive statistics

Sample period is 2017-2022. Data frequency is monthly. Δcpi_t: first difference of Consumer Price Index. Δer_t: log difference peso-dollar exchange rate. $F_{t-1}\Delta cpi_t$: one-month-ahead forecast at time $t-1$ of the inflation rate at time t. $AttentionIndex_t$: inflation attention news index. $DirectionIndex_t$: inflation news direction index. $IncreasingDirection_t$: increasing inflation direction index. $DecreasingDirection_t$: decreasing inflation direction index. $BackwardIndex_t$: backward inflation news index. $ForwardIndex_t$: forward inflation news index. $BackDir_t$: backward inflation news direction index. $ForwDir_t$: forward inflation news direction index.

Variable	Mean	Median	St. Dev.	Q1	Q3	Minimum	Maximum	N
Δcpi_t	0.034	0.032	0.015	0.023	0.041	0.012	0.074	72
Δer_t	0.033	0.026	0.046	0.011	0.041	-0.039	0.248	72
$F_{t-1}\Delta cpi_t$	0.032	0.030	0.013	0.024	0.040	0.013	0.063	72
$AttentionIndex_t$	0.274	0.260	0.071	0.231	0.304	0.139	0.445	72
$DirectionIndex_t$	0.006	0.006	0.007	0.001	0.010	-0.010	0.020	72
$IncreasingDirection_t$	0.024	0.024	0.006	0.020	0.028	0.012	0.040	72
$DecreasingDirection_t$	0.019	0.017	0.005	0.015	0.023	0.011	0.031	72
$BackwardIndex_t$	0.011	0.011	0.003	0.010	0.013	0.006	0.018	72
$ForwardIndex_t$	0.002	0.002	0.001	0.002	0.003	0.001	0.004	72
$BackDir_t$	0.001	0.001	0.001	0.000	0.001	-0.001	0.003	72
$ForwDir_t$	0.000	0.000	0.000	0.000	0.000	-0.001	0.000	72

Figure 1 provides further evidence on the co-movement of inflation, the inflation attention news index, the inflation news direction index, the backward inflation news index, and the forward inflation news index. We can see that increments in the indices based on news coincide with increments in the inflation rate. Additionally, the estimated coefficient of correlation among inflation and the attention indices ranges between 0.36 and 0.71. When we compute the correlation using the one-month-lagged indices based on mass media, the estimated coefficient of correlation ranges between 0.42 and 0.66. This result suggests that this set of indicators in general captures forward-looking information.

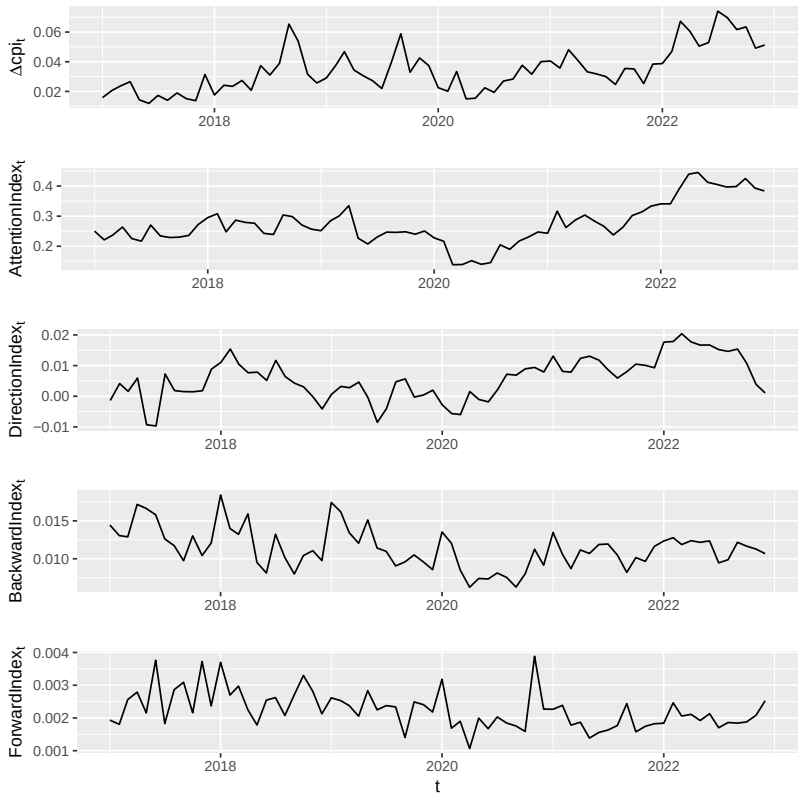

Figure 1: Inflation rate (Δcpi_t), and indexes based on mass media articles

2.4 Methodology

To see whether media articles have useful information regarding the future evolution of the inflation rate, we estimate a series of forecasting models. These models are given by an autoregressive specification that, depending on the specification, is complemented with a lagged indicator which belongs to the set of variables that has been presented in the previous section. The estimation of these models allows us to test the predictive power of the non-traditional indicators by evaluating the information content regarding the inflation dynamic beyond the traditional variables. In this sense, we consider three benchmark models. The first benchmark model (AR Benchmark) is given by:

$$\Delta cpi_{t+1} = \alpha + \sum_{s=0}^{p} \beta_s \Delta cpi_{t-s} + \mu_{t+1} \tag{1}$$

A second benchmark model (AR-EX Benchmark) is estimated, which involves the estimation of the following equation:

$$\Delta cpi_{t+1} = \alpha + \sum_{s=0}^{p} \beta_s \Delta cpi_{t-s} + \beta_{er} \Delta er_t + \mu_{t+1} \quad (2)$$

Finally, the third benchmark model (AR-EX-PF Benchmark) proposes the following specification:

$$\Delta cpi_{t+1} = \alpha + \sum_{s=0}^{p} \beta_s \Delta cpi_{t-s} + \beta_{er} \Delta er_t + \beta_{pf} F_t \Delta cpi_{t+1} + \mu_{t+1} \quad (3)$$

where μ_{t+1} is the error term. To examine the ability of indexes based on mass medias to predict Δcpi_{t+1}, we extended the benchmark models by adding I_t as a predictor, which could be one of the indices related to inflation dynamics. The number of lags is selected minimizing the Bayesian Information Criterion.

3 Results

To ascertain whether the suggested inflation attention indexes based on mass media content are capturing future inflation dynamics, we estimate a set of forecasting models. On the one hand, an in-sample forecast analysis is carried out. On the other hand, we evaluate if the documented regularities remain when we extend the exercises to an out-of-sample analysis. Lastly, we compare the informative capacity of mass media indices respect with an alternative indicator of subjective states.

3.1 Inflation forecast: in-sample analysis

This section presents and discusses the estimations of the different exercises related on in-sample forecasting models. The three baseline and extended models are represented along the columns of Table 2, 3, and 4. The baseline models indicate that lagged monthly inflation, lagged devaluation, and lagged professional forecast expectation are statistically and economically significant predictors of inflation. Conversely, the extended models, which incorporate mass media news content, demonstrate that it contributes valuable information regarding future inflation levels. Overall, the estimated coefficients signify that these predictors are economically significant. Furthermore, the adjusted R^2 values suggest that these variables contain substantive information regarding subsequent levels of inflation.

Considering the estimated coefficients of the inflation attention news index, the result are in line with Carroll (2003), Lamla and Lein (2008), and Maag and Lamla (2009). In this sense, more media reporting about inflation anticipates future inflation dynamics. Additionally, set of regressors that exploit the content of news regarding inflation seems to be informative as a regressor. In this

sense, the sign of the inflation news direction index is positive and statistically significant, which suggests that more media reporting about increasing inflation anticipates an increment (on average) of the inflation rate. This results seems robust when we control by the effect of news that reports on falling and rising inflation separately. The estimated coefficients of the increasing and decreasing inflation direction index are in line with expected results. These results do not change significantly when we control by the neutral inflation direction index. These results are consistent with the work Lamla and Lein (2008) and Maag and Lamla (2009).

However, we also examined the relevance of the time dimension in relation to specific news stories about inflation. While the sign of the estimated coefficients for the backward and forward inflation indexes align with our expectations, it appears that only the indexes capturing backward-looking content provide informative insights into the future dynamics of inflation

Table 2: Inflation forecast model: AR Benchmark

Sample period is 2017-2022. Data frequency is monthly. Δcpi_t: first difference of Consumer Price Index. $AttentionIndex_t$: inflation attention news index. $DirectionIndex_t$: inflation news direction index. $IncreasingDirection_t$: increasing inflation direction index. $DecreasingDirection_t$: decreasing inflation direction index. $NeutralDirection_t$: neutral inflation direction index. $BackwardIndex_t$: backward inflation news index. $ForwardIndex_t$: forward inflation news index. $BackDir_t$ backward inflation news direction index. $ForwDir_t$ forward inflation news direction index.

	lead(pi, 1)						
Δcpi_t	0.011^{***}	0.009^{***}	0.010^{***}	0.009^{***}	0.009^{***}	0.012^{***}	0.010^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$AttentionIndex_t$		0.004^{***}					
		(0.001)					
$DirectionIndex_t$			0.003^{***}				
			(0.001)				
$IncreasingDirection_t$				0.003^{**}	0.005^{**}		
				(0.001)	(0.002)		
$DecreasingDirection_t$				-0.003^{***}	-0.002^{*}		
				(0.001)	(0.001)		
$NeutralDirection_t$					-0.003		
					(0.002)		
$BackwardIndex_t$						-0.0003	
						(0.001)	
$ForwardIndex_t$						0.001	
						(0.001)	
$BackDir_t$							0.002
							(0.001)
$ForwDir_t$							0.002^{*}
							(0.001)
Constant	0.035^{***}	0.035^{***}	0.035^{***}	0.035^{***}	0.035^{***}	0.035^{***}	0.035^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	71	71	71	71	71	71	71
Adjusted R^2	0.573	0.597	0.600	0.599	0.601	0.562	0.598

Note: standard errors in parentheses are estimated following Newey and West (1987, 1994).*p<0.1; **p<0.05; ***p<0.01

Table 3: Inflation forecast model: AR-EX Benchmark

Sample period is 2017-2022. Data frequency is monthly. Δcpi_t: first difference of Consumer Price Index. Δer_t: log difference peso-dollar exchange rate. $AttentionIndex_t$: inflation attention news index. $DirectionIndex_t$: inflation news direction index. $IncreasingDirection_t$: increasing inflation direction index. $DecreasingDirection_t$: decreasing inflation direction index. $NeutralDirection_t$: neutral inflation direction index. $BackwardIndex_t$: backward inflation news index. $ForwardIndex_t$: forward inflation news index. $BackDir_t$: backward inflation news direction index. $ForwDir_t$: forward inflation news direction index.

	lead(pi, 1)							
Δcpi_t	0.011^{***}	0.011^{***}	0.008^{***}	0.009^{***}	0.009^{***}	0.009^{***}	0.011^{***}	0.009^{***}
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Δer_t		0.002^{***}	0.003^{***}	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$AttentionIndex_t$			0.004^{***}					
			(0.001)					
$DirectionIndex_t$				0.003^{**}				
				(0.001)				
$IncreasingDirection_t$					0.003^{**}	0.005^{*}		
					(0.001)	(0.003)		
$DecreasingDirection_t$					-0.003^{***}	-0.002^{*}		
					(0.001)	(0.001)		
$NeutralDirection_t$						-0.003		
						(0.003)		
$BackwardIndex_t$							0.0002	
							(0.001)	
$ForwardIndex_t$							0.0004	
							(0.001)	
$BackDir_t$								0.002^{*}
								(0.001)
$ForwDir_t$								0.001
								(0.001)
Constant	0.035^{***}	0.035^{***}	0.035^{***}	0.035^{***}	0.035^{***}	0.035^{***}	0.035^{***}	0.035^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	71	71	71	71	71	71	71	71
Adjusted R^2	0.573	0.591	0.620	0.620	0.616	0.615	0.580	0.614

Note: standard errors in parentheses are estimated following Newey and West (1987, 1994).*p<0.1; **p<0.05; ***p<0.01

Table 4: Inflation forecast model: AR-EX-PF Benchmark

Sample period is 2017-2022. Data frequency is monthly. Δcpi_t: first difference of Consumer Price Index. Δer_t: log difference peso-dollar exchange rate. $F_{t-1}\Delta cpi_t$: one-month-ahead forecast at time $t-1$ of the inflation rate at time t. $AttentionIndex_t$: inflation attention news index. $DirectionIndex_t$: inflation news direction index. $IncreasingDirection_t$: increasing inflation direction index. $DecreasingDirection_t$: decreasing inflation direction index. $NeutralDirection_t$: neutral inflation direction index. $BackwardIndex_t$: backward inflation news index. $ForwardIndex_t$: forward inflation news index. $BackDir_t$: backward inflation news direction index. $ForwDir_t$: forward inflation news direction index.

	lead(pi, 1)								
Δcpi_t	0.011***	0.011***	0.004**	0.001	0.002	0.002	0.002	0.004**	0.003**
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Δer_t		0.002***	0.001	0.002*	0.001	0.002	0.001	0.001	0.001*
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$F_t\Delta cpi_{t+1}$			0.009***	0.009***	0.009***	0.010***	0.010***	0.010***	0.009***
			(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.003)	(0.002)
$AttentionIndex_t$				0.004**					
				(0.002)					
$DirectionIndex_t$					0.003***				
					(0.001)				
$IncreasingDirection_t$						0.004***	0.004*		
						(0.001)	(0.002)		
$DecreasingDirection_t$						−0.001	−0.001		
						(0.001)	(0.001)		
$NeutralDirection_t$							−0.001		
							(0.003)		
$BackwardIndex_t$								0.001	
								(0.001)	
$ForwardIndex_t$								0.001	
								(0.001)	
$BackDir_t$									0.002**
									(0.001)
$ForwDir_t$									0.001
									(0.001)
Constant	0.035***	0.035***	0.035***	0.035***	0.035***	0.035***	0.035***	0.035***	0.035***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	71	71	71	71	71	71	71	71	71
Adjusted R^2	0.573	0.591	0.688	0.713	0.722	0.725	0.721	0.693	0.707

Note: standard errors in parentheses are estimated following Newey and West (1987, 1994).*p<0.1; **p<0.05; ***p<0.01

Summarizing, the evidence reported above suggests that indices based on mass media content have valuable information regarding future levels of inflation. It is worth noting that these observations do not depend on the baseline model under consideration or the specification of the mass media content indicator.

3.2 Inflation forecast: out-of-sample analysis

In this subsection, we provide further insights on the associations between information content of mass media and inflation rate. In particular, we implement out-of-sample forecasts exercises in which models are trained recursively with past information. The performance of forecast generated by the baseline autoregressive models are compared to forecasts produced by models that incorporate an additional predictor.

Each forecast model is evaluated computing the root-mean-square prediction error (RMSE). For extended models these measure of accuracy is also expressed as a fraction of the RMSE of the baseline model. A ratio below one suggest that the accuracy of the extended model overcome the performance of the baseline model. Additionally, the performance of the models is assessed using two alternative starting dates for the pseudo out-of-sample forecast exercise. The starting dates are selected so that the smallest training subsample represents 60% and 80% of the full sample respectively.

Table 5 shows the results for out-of-sample forecast exercises. For most of the mass media, the estimated forecast accuracy is higher than that observed in the case of the baseline model. When we considered the forecasts generated by a single predictor, the improvement in performance is observed not only in the inflation attention index, but also when we incorporate more precise indices (such as the direction index and the index that capture an increasing inflation dynamic). More specifically, the strongest accuracy gains is observed when baseline models are extended by the inflation news direction index $DirectionIndex_t$ and increasing inflation news direction index ($IncreasingDirection_t$). These results are in line with the regularities have reported from in-sample forecast models. Furthermore, most of extended forecast model (we use the AR-EX-PF as a benchmark model) achieve a favorable performance respect with expert forecast, when the latter are considered as benchmark model. Additionally, forecast combination allows for considerable gains in accuracy.

Considering there is a very high level of common information in these indices, we build a composite mass media measure based on the set of indicators exhibited in Table 5 from column 4 to 11. A Principal Component technique is implemented, which is estimated recursively[7]. The first principal component is

[7]The eigenvector associated with the first principal component is 0.42, 0.52,

the latent variable that captures the common information provided by this set of indicators. Column 13 shows the performances of forecast generated by an extended model which incorporates the latent variable as regressor. This forecast model achieves one of the strongest performance, which evidences valuable gains derived from combining the information based on mass media news.

In summary, these out-of-sample forecast exercises provide further support to the idea that social media content provides valuable information regarding future levels of inflation.

Table 5: Inflation forecast models: out-of-sample

Sample period is 2017-2022. Data frequency is monthly. Δcpi_t: first difference of Consumer Price Index. Δer_t: log difference peso-dollar exchange rate. $F_{t-1}\Delta cpi_t$: one-month-ahead forecast at time $t-1$ of the inflation rate at time t. $AttentionIndex_t$: inflation attention news index. $DirectionIndex_t$: inflation news direction index. $IncreasingDirection_t$: increasing inflation direction index. $DecreasingDirection_t$: decreasing inflation direction index. $BackwardIndex_t$: backward inflation news index. $ForwardIndex_t$: forward inflation news index. $BackDir_t$: backward inflation news direction index. $ForwDir_t$: forward inflation news direction index. $ForecastComb_t$: Forecast combinations are implemented through simple averages. $LatentIndex_t$: first principal component of mass media indexes.

			$AttentionIndex_t$	$DirectionIndex_t$	$IncreasingDirection_t$	$DecreasingDirection_t$:	$BackwardIndex_t$	$ForwardIndex_t$	$BackDir_t$	$ForwDir_t$	$ForecastComb_t$	$LatentIndex_t$
Forecasts begin: 07/2020 (60%)												
Δcpi_{t-1}	RMSE	0.0125	0.0105	0.0101	0.0124	0.0117	0.0124	0.0124	0.0113	0.0118	0.0095	0.0105
	Ratio		0.843	0.811	0.995	0.939	0.995	0.995	0.907	0.947	0.761	0.843
$\Delta cpi_{t-1} + \Delta er_t$	RMSE	0.0126	0.011	0.0106	0.0123	0.0116	0.0126	0.0126	0.0116	0.0119	0.0101	0.011
	Ratio		0.875	0.843	0.979	0.923	1.003	1.003	0.923	0.947	0.803	0.875
$\Delta cpi_{t-1} + \Delta er_t + F_t\Delta cpi_{t+1}$	RMSE	0.0096	0.0086	0.0082	0.0083	0.0098	0.0095	0.0097	0.0087	0.0095	0.0086	0.0078
	Ratio		0.896	0.854	0.866	1.023	0.991	1.009	0.909	0.989	0.895	0.817
$F_t\Delta cpi_{t+1}$	RMSE	0.0095										
	Ratio		0.899	0.857	0.868	1.026	0.994	1.012	0.911	0.992	0.897	0.819
Forecasts begin: 10/2021 (80%)												
Δcpi_{t-1}	RMSE	0.0165	0.0131	0.013	0.0164	0.0155	0.0168	0.0164	0.0147	0.0156	0.0122	0.0133
	Ratio		0.796	0.789	0.996	0.941	1.02	0.996	0.893	0.947	0.742	0.808
$\Delta cpi_{t-1} + \Delta er_t$	RMSE	0.0164	0.0139	0.0121	0.0138	0.0166	0.0158	0.0164	0.0138	0.0158	0.0145	0.0138
	Ratio		0.848	0.823	0.976	0.945	1.006	1.000	0.909	0.958	0.776	0.842
$\Delta cpi_{t-1} + \Delta er_t + F_t\Delta cpi_{t+1}$	RMSE	0.0128	0.0115	0.0106	0.0109	0.013	0.0125	0.0128	0.0113	0.0125	0.0113	0.0101
	Ratio		0.898	0.828	0.851	1.015	0.976	0.999	0.882	0.976	0.879	0.789
$F_t\Delta cpi_{t+1}$	RMSE	0.0128										
	Ratio		0.901	0.831	0.854	1.019	0.98	1.003	0.886	0.98	0.882	0.792

0.38, -0.24, 0.03, -0.12, 0.47, 0.35. where each element represents the loading of $AttentionIndex_t$, $DirectionIndex_t$, $IncreasingDirection_t$, $DecreasingDirection_t$, , $BackwardIndex_t$, $ForwardIndex_t$, $BackDir_t$, and $ForwDir_t$, respectively

3.3 Evaluation of other inflation attention indexes

Along this subsection, we evaluate how inflation attention indexes based on mass media data compares with other alternative indicators of subjective states. Aromí and Llada (ming) propose a novel inflation attention index based on information published in Twitter. The value of their monthly index is given by the ratio between the mentions of inflation and the number of words in tweets of the corresponding month. Using a sample for Argentina during the period 2012-2019, the authors show that mass media content and mass media tweets fail to provide information regarding future inflation beyond the information that capture the lagged inflation rate and the inflation attention index based on tweets. In this regard, an open question arises: Does mass media content contain valuable information about future inflation when we consider a broader range of mass media sources?

In order to answer this question, we estimate the out-of-sample forecasts models by incorporating in the extended models the inflation attention index based on Twitter[8]. The results are reported in Table 6. Column 4 evidences gains in accuracy when we extend the AR-EX baseline model by incorporating an index based on Twitter content. However, there is no gain in the predictive performance when the expert forecasts are considered.

In a second evaluation, we compare the the performance of baseline models versus the performance of combination forecasts in which model inflation forecasts extended by mass media and twitter indicators are averaged. Column 5 in Table 6 allows to infer that forecast combination allows for considerable gains in accuracy.

Lastly, the last column of Table 6 shows that a combination of subjective indexes allow for further gains in accuracy. More specifically, we propose a new regressor which is given by the first principal component considering the set of indicators based on mass media and the inflation attention index based on twitter.[9] As we can in column 6, a extended forecast model that incorporates a regressor that combine indexes based on mass media and social media content report one of the strongest performance, even when we compare its performance with those forecast models showed in Table 5.

In summary, these exercises provide further support to the idea that mass media and social media content provides valuable information regarding future levels of inflation.

[8]The data can be download from `https://sites.google.com/view/inflacion-y-redes-sociales`.

[9]The eigenvector associated with the first principal component is 0.42, 0.40, 0.46, 0.34, -0.21, 0.04, -0.11, 0.43, 0.30. where each element represents the loading of $AttentionIndex_t$, itw_t, $DirectionIndex_t$, $IncreasingDirection_t$, $DecreasingDirection_t$, , $BackwardIndex_t$, $ForwardIndex_t$, $BackDir_t$, and $ForwDir_t$, respectively

Table 6: Inflation forecast models: out-of-sample and Indexes based on Twitter

Sample period is 2017-2022. Data frequency is monthly. Δcpi_t: first difference of Consumer Price Index. Δer_t: log difference peso-dollar exchange rate. $F_{t-1}\Delta cpi_t$: one-month-ahead forecast at time $t-1$ of the inflation rate at time t. itw_t : inflation attention index based on Twitter. $ForecastComb_t$: Forecast combinations are implemented through simple averages. $LatentIndex_t$: first principal component of mass media indexes and inflation attention index based on Twitter.

			itw_t	$ForecastComb_t$	$LatentIndex_t$
Forecasts begin: 07/2020 (60%)					
$\Delta cpi_{t-1} + itw_t$	RMSE	0.0125	0.012	0.0093	0.0096
	Ratio		0.963	0.749	0.769
$\Delta cpi_{t-1} + \Delta er_t + itw_t$	RMSE	0.0126	0.0112	0.0099	0.0102
	Ratio		0.891	0.791	0.812
$\Delta cpi_{t-1} + \Delta er_t + F_t\Delta cpi_{t+1} + itw_t$	RMSE	0.0096	0.0095	0.0084	0.0077
	Ratio		0.991	0.879	0.8
$F_t\Delta cpi_{t+1}$	RMSE	0.0095			
	Ratio		0.994	0.882	0.802
Forecasts begin: 10/2021 (80%)					
$\Delta cpi_{t-1} + itw_t$	RMSE	0.0165	0.0142	0.0135	0.0125
	Ratio		0.862	0.822	0.759
$\Delta cpi_{t-1} + \Delta er_t + itw_t$	RMSE	0.0164	0.0133	0.0124	0.0125
		0.811	0.757	0.762	
$\Delta cpi_{t-1} + \Delta er_t + F_t\Delta cpi_{t+1} + itw_t$	RMSE	0.0128	0.0127	0.0111	0.0099
	Ratio		0.992	0.866	0.771
$F_t\Delta cpi_{t+1}$	RMSE	0.0128			
	Ratio		0.995	0.87	0.774

3.4 Evaluation of indexes based on NLI

In the previous section, we provided evidence that indexes capturing different aspects of inflation based on news contain valuable information for anticipating its dynamics. These indexes were constructed using a dictionary approach, which relies on a list of words that depend on user's ad-hoc decisions. In this context, it is important to assess whether the observed regularities persist when

using another classifier method. Given our categorical classification problem, we will employ Natural Language Inference (NLI), a model belonging to the field of Natural Language Processing (NLP). The NLI model determines which discrete label from a predefined set is applicable to an inference pair, composed of a premise (p) and a hypothesis (h). The Figure 2 shows an example for an application of a NLI model to classify inflation features. In this sense, the NLI model must determine whether the hypothesis (a prompt which represents the class label) entails the premise (which corresponds to the instance to be classified). For more details, see MacCartney and Manning (2008).

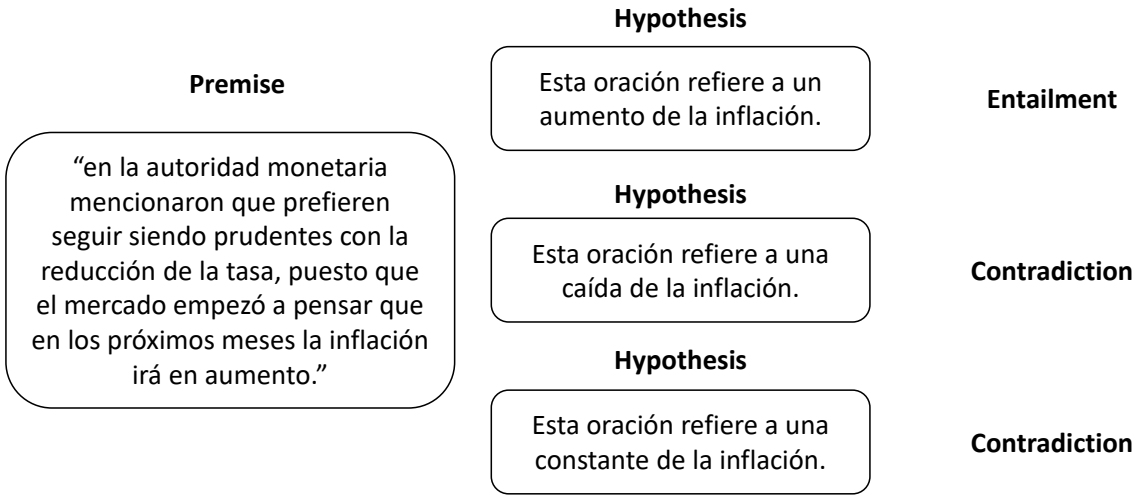

Figure 2: An example of the application of NLI inflation features classification. Given the premise "In the monetary authority, they mentioned that they prefer to remain cautious about reducing the interest rate, as the market has started to believe that inflation will increase in the coming months.", three hypotheses represent the inflation direction (increase, decrease, constant). The representation of increase is entailed and therefore predicted.

We use the pretrained sentence transformer model "Recognai/bert-base-

spanish-wwm-cased-xnli", which is publicly available within the Hugging Face Transformers Python library.[10]. This model contains 110M of parameters. This classifier model is asked to estimate how likely the situation described in the hypothesis sentence would be true given the premise. Formally, given a premise $p \in P$ and a hypothesis $h \in H$, a NLI model is a mapping $F : P \times H \rightarrow [0, 1]$. Through this model, the set of indices are given by $p(h, x_t) * n_t$, where $p(.)$ is the average entailment probability across all sentences in t and n_t represents the total number if articles corresponding to month t.

Table 7 shows the results for out-of-sample forest exercises following the procedure explained in subsection 3.2. On average, the accuracy of indexes based on NLI modelme are higher than the performance of indexes build using traditional approaches. These out-of-sample forecast results provide evidence of significant accuracy improvements gained from using machine learning models for performing natural language processing task, such as computing precise indexes to approximate a complex economic phenomenon like inflation dynamics. Nevertheless, a more comprehensive analysis is required to achieve a more informative evaluation of these results. This entails identifying the sensitivity of the results to a set of different prompts (hypotheses) and different NLI models.

[10] `https://huggingface.co/sentence-transformers`.

Table 7: Inflation forecast models: Out-of-sample and Indexes based on NLI

Sample period is 2017-2022. Data frequency is monthly. Δcpi_t: first difference of Consumer Price Index. Δer_t: log difference peso-dollar exchange rate. $F_{t-1}\Delta cpi_t$: one-month-ahead forecast at time $t-1$ of the inflation rate at time t. $AttentionIndex_t$: inflation attention news index. $DirectionIndex_t$: inflation news direction index. $IncreasingDirection_t$: increasing inflation direction index. $DecreasingDirection_t$: decreasing inflation direction index. $BackwardIndex_t$: backward inflation news index. $ForwardIndex_t$: forward inflation news index. $BackDir_t$: backward inflation news direction index. $ForwDir_t$: forward inflation news direction index. $ForecastComb_t$: Forecast combinations are implemented through simple averages. $LatentIndex_t$: first principal component of mass media indexes.

			$AttentionIndex_t$	$DirectionIndex_t$	$IncreasingDirection_t$	$DecreasingDirection_t$:	$BackwardIndex_t$	$ForwardIndex_t$	$BackDir_t$	$ForwDir_t$	$ForecastComb_t$	$LatentIndex_t$
Forecasts begin: 07/2020 (60%)												
Δcpi_{t-1}	RMSE	0.0125	0.0105	0.01	0.0102	0.0109	0.0106	0.0102	0.0101	0.0105	0.01	0.0101
	Ratio		0.843	0.802	0.819	0.875	0.851	0.819	0.811	0.843	0.803	0.811
$\Delta cpi_{t-1} + \Delta er_t$	RMSE	0.0126	0.011	0.0099	0.0102	0.0111	0.0106	0.0102	0.01	0.0104	0.0102	0.01
	Ratio		0.875	0.789	0.812	0.883	0.843	0.812	0.795	0.828	0.814	0.794
$\Delta cpi_{t-1} + \Delta er_t + F_t\Delta cpi_{t+1}$	RMSE	0.0096	0.0086	0.0083	0.0087	0.0091	0.0089	0.0088	0.0083	0.0087	0.0085	0.0086
	Ratio		0.896	0.865	0.909	0.949	0.935	0.923	0.871	0.905	0.884	0.904
$F_t\Delta cpi_{t+1}$	RMSE	0.0095										
	Ratio		0.899	0.867	0.911	0.951	0.937	0.926	0.874	0.907	0.886	0.906
Forecasts begin: 10/2021 (80%)												
Δcpi_{t-1}	RMSE	0.0165	0.0131	0.0125	0.0127	0.014	0.0133	0.0128	0.0124	0.0134	0.0126	0.0124
	Ratio		0.796	0.759	0.771	0.85	0.808	0.777	0.753	0.814	0.767	0.753
$\Delta cpi_{t-1} + \Delta er_t$	RMSE	0.0164	0.0139	0.0124	0.0127	0.014	0.0134	0.013	0.0125	0.0129	0.0128	0.0125
	Ratio		0.848	0.756	0.775	0.854	0.817	0.793	0.762	0.787	0.781	0.762
$\Delta cpi_{t-1} + \Delta er_t + F_t\Delta cpi_{t+1}$	RMSE	0.0128	0.0115	0.0108	0.0113	0.0122	0.0117	0.0116	0.0107	0.0114	0.0112	0.0111
	Ratio		0.898	0.843	0.882	0.952	0.913	0.906	0.835	0.89	0.875	0.867
$F_t\Delta cpi_{t+1}$	RMSE	0.0128										
			0.901	0.846	0.886	0.956	0.917	0.909	0.839	0.894	0.878	0.87

4 Conclusions

This paper examines the information content of mass media news in economic contexts. More specifically, we analyze a news have published by set of prominent newspaper of Argentina for the period 2017-2022. The evidence indicates that mass media news provide valuable information regarding future inflation dynamic. The information content is economically significant. Also, we evidence a huge accuracy gain when forecast model incorporate more precise indexes that capture the price change. In this sense, we show interesting information gains derived from exploiting the news content in a deep way. Also, the information content of the indexes is different from that provided by traditional macroeconomic indicators and indexes based on social media. These findings are robust to changes in the specification of the forecast exercise.

There are several directions in which these exercises can be extended. First, in this work, we use a simple strategy to summarize unstructured information. The use of natural language processing models could allow for gains in the capacity to extract information from mass media news. In a similar direction, we could evaluate whether a certain group of newspapers (e.g., business-focused) could be more informative. Finally, this study evaluated regularities using monthly time series. Analyses at higher frequencies can provide further insights regarding the relationship between mass media content and inflation dynamics.

Acknowledgement

This paper was written while the second author was visiting the IELAT (UAH), Spain. He thanks the financial support received from Santander Bank and University of Alcalá.

References

Altig, D., Baker, S., Barrero, J. M., Bloom, N., Bunn, P., Chen, S., Davis, S. J., Leather, J., Meyer, B., M. E., Mizen, P., Parker, N., Renault, T., Smietanka, P., and Thwaites, G. (2020). Economic uncertainty before and during the covid-19 pandemic. *Journal of Public Economics*, 191.

Andrade, P. and Le Bihan, H. (2013). Inattentive professional forecasters. *Journal of Monetary Economics*, 60(8):967–982.

Angelico, C., Marcucci, J., Miccoli, M., and Quarta, F. (2022). Can we measure inflation expectations using twitter? *Journal of Econometrics*, 228(2):259–277.

Aromí, D. J. and Llada, M. (forthcoming). Forecasting inflation with twitter. *Económica*.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636.

Carroll, C. D. (2003). Macroeconomic expectations of households and professional forecasters. *The Quarterly Journal of Economics*, 118(1):269–298.

Coibion, O. and Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–78.

Denes, J., Lestrade, A., and Richardet, L. (2021). Using twitter data to measure inflation perception. Working papers, Irving Fisher Committee on Central Bank Statistics.

Fuhrer, J. C. (2018). Intrinsic expectations persistence: evidence from professional and household survey expectations. Working Papers 18-9, Federal Reserve Bank of Boston.

Lamla, M. J. and Lein, S. M. (2008). The role of media for consumers' inflation expectation formation. KOF Working Papers 201, ETH Zurich.

Larsen, V. H., Thorsrud, L. A., and Zhulanova, J. (2021). News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, 117:507–520.

Maag, T. and Lamla, M. J. (2009). The role of media for inflation forecast disagreement of households and professional forecasters. KOF Working Papers 223, ETH Zurich.

MacCartney, B. and Manning, C. D. (2008). Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.

Newey, W. K. and West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653.

Sims, C. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

Thorsrud, L. A. (2016). Nowcasting using news topics, big data versus big bank. Working Papers 20, Norges Bank.

Thorsrud, L. A. (2020). Words are the new numbers: a newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, 38(2):393–409.