Forecasting inflation with twitter

J. Daniel Aromí [1,3]   Martín Llada [1, 2]

[1] IIEP-Baires (UBA-Conicet) [2]FCE

[3]Universidad Católica Argentina

December 6, 2019

Overview

1. Motivation
2. Data & methodology
3. In-sample Forecast
4. Out-of-sample Forecast
5. Extended Analysis
6. Conclusions

Motivation: Inflation as an emergent process

The general price level evolves in a context of:

- **Nominal arbitrariness/indeterminacy.**
  (Ascari&Ropele 2009, Lubik&Schorfheide 2004, Beyer&Farmer 2004)
- **Forward-looking adaptive behavior.**
  (Heymann&Leijonhufvud 1995, Arifovic 1995, De Grauwe&Ji 2019)
- **Variable policy regimes.**
  (Ascari&Ropele 2009, McCallum 2001, Cukierman&Meltzer 1986, Hommes&Lustenhouwer 2019)

Hence,...

- **Traditional indicators** (interest rates, monetary aggregates, fiscal deficits, ...) might miss relevant aspects.
- Potential gains linked to proxies of **subjective states**.

Motivation

This work:
Social media content as an indicator of **unobservable states/factors** controlling the evolution of inflation in Argentina (2012-2019).

Specific evaluations:

- Does social media contain valuable information regarding the evolution of inflation?
- How does the performance of the resulting index compare with other proxies of subjective states? (Google trends, surveys, newspaper content, mass media tweets)

Data & methodology

Twitter Data:

- 2012-2019: Approx. 70 million tweets.
- Sample Stream (representative 1%) + web-scrapped tweets (for selected months).
- Argentine tweets identified by user self-reported location.

Simple indicators of Twitter content:

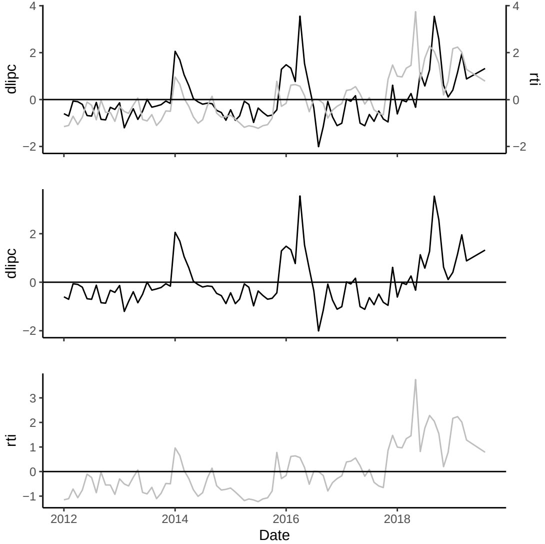
- **Level of attention:**

$$I_t = \frac{\#\ mentions\ of\ \text{``}inflation\text{''}}{\#\ of\ tweets}$$

- **Relative level of attention:**

$$\hat{I}_t = I_t - \frac{\sum_{k=1}^{12} I_{t-k}}{12}$$

Inflation rate and Indicator of Attention based on Twitter



Descriptive statistics

Variable	Mean	Median	St. Dev.	Q1	Q3	Minimum	Maximum
$\triangle$ *ipc*	0.02	0.02	0.01	0.02	0.03	0.00	0.06
$\triangle$ *tcn*	0.03	0.01	0.05	0.01	0.03	-0.04	0.25
#Tweets	690424.46	677686.00	281433.58	450239.00	929539.00	110712.00	1392608.00
Mentions of "inflation"	209.39	151.00	153.42	102.00	302.00	24.00	739.00
I_t ($\times 10^4$)	3.15	2.58	1.91	1.71	4.18	0.81	9.75

Sample period is 2012-2019. Data frequency is monthly.

Results: In-sample Forecasts

	$\triangle$ ipc$_{t+1}$					
$\triangle$ ipc$_t$	0.008***	0.006***	0.006***	0.006***	0.005***	0.006***
	(0.001)	(0.0004)	(0.001)	(0.001)	(0.001)	(0.0004)
$\triangle$ tcn		0.003***			0.002***	0.002***
		(0.001)			(0.001)	(0.001)
I_t			0.004***		0.004***	
			(0.001)		(0.001)	
$\hat{I}_t$				0.004***		0.003***
				(0.001)		(0.001)
Constant	0.025***	0.025***	0.025***	0.025***	0.025***	0.025***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	92	92	92	80	92	80
R^2	0.433	0.495	0.531	0.516	0.560	0.539
Adjusted R^2	0.427	0.483	0.520	0.504	0.545	0.521
F Statistic	68.678***	43.580***	50.305***	41.095***	37.360***	29.614***

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

Other proxies of subjective states

$$\triangle\ ipc_{t+1} = \alpha + \beta_0 \triangle\ ipc_t + \beta_{ind} ind_t + u_t$$

	Baseline	I_t	GT-inflation	GT-dollar	Newspaper	Mass media tweets	Cons. Surv.
$\hat{\alpha}$	0.025***	0.025***	0.025***	0.024***	0.025***	0.024***	0.025***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$\hat{\beta}_0$	0.008***	0.005***	0.006***	0.005***	0.007***	0.007***	0.007***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$\hat{\beta_{ind}}$		0.004***	0.002**	0.004***	-0.001	0.000	-0.000
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Adj. R^2	0.427	0.511	0.451	0.494	0.428	0.410	0.414

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

Out-of-sample Forecast

Methodology:

- Baseline autoregressive model:

$$\triangle\ ipc_{t+1} = \alpha + \beta_0 \triangle\ ipc_t + u_t$$

- Evaluated model:

$$\triangle\ ipc_{t+1} = \alpha + \beta_0 \triangle\ ipc_t + \beta_{ind} ind_t + u_t.$$

  Where $ind_t \in \{I_t, \hat{I}_t, \hat{I}_t^{+}, \triangle\ tcn_t\}$.

- Gains in forecast accuracy: ratio of model RMSE vs baseline RMSE.

Details:

- Expanding window for training dataset.
- First forecast exercise with 60% and 80% of the sample in the training dataset.
- Statistical inference using bootstrap methodology (Faust et al. 2013).

Results: Out of sample forecasts

Forecasts begin	11/2016 (60%)		04/2018 (80%)	
	RMSE	Ratio	RMSE	Ratio
Baseline	0.0099		0.0125	
$\triangle\ tcn_t$	0.0093	0.93 [0.07]	0.0114	0.91 [0.06]
I_t	0.0091	0.91 [0.03]	0.0104	0.83 [0.01]
$\hat{I}_t$	0.0094	0.94 [0.11]	0.0112	0.90 [0.04]
$\hat{I}_t^+$	0.0090	0.90 [0.03]	0.0101	0.81 [0.01]
Forecast combination	0.0089	0.89 [0.01]	0.0106	0.84 [0.01]

Note: Forecast combination is implemented through simple averages. p-values in brackets.

Professional Forecasters: Central Bank survey (REM)

- Comparing performance of model vs. expert forecasts.
- Evaluating complementarities between model & expert forecasts.

RMSE

Forecasts begin	11/2016 (60%)	04/2018 (80%)
REM	0.0080	0.0102
Without twitter content		
Model forecast	0.0095	0.0117
Forecast combination (REM+Model)	0.0081	0.0103
With twitter content		
Model forecast	0.0089	0.0106
Forecast combination (REM+Model)	0.0077	0.0096

Conclusions

Results:

- Twitter content provides valuable information regarding the evolution of inflation.
- The combination of traditional economic indicators and indices based on Twitter allows for gains in prediction accuracy.

Further research:

- NLP (Topic models, word embeddings)
- Network topology, communities, classification of users,...