Read between the headlines: Can news data predict inflation?

ALAN CHESTER ARCIN, MA. ELLYSAH JOY GULIMAN, GENNA PAOLA CENTENO, JACQUELINE MARGAUX HERBO, SANJEEV PARMANAND, **CHERRIE MAPA**
*Department of Economic Research, Bangko Sentral ng Pilipinas**

4th BIS-IFC– Bank of Italy Workshop on Data Science in Central Banking
18-20 February 2025

* The usual institutional disclaimer applies.

Motivation

Leverage nontraditional data sources

Support inflation nowcasting

Inflation News Index (INI)

The INI is a quantitative summary of digital news coverage on inflation and prices of goods and services

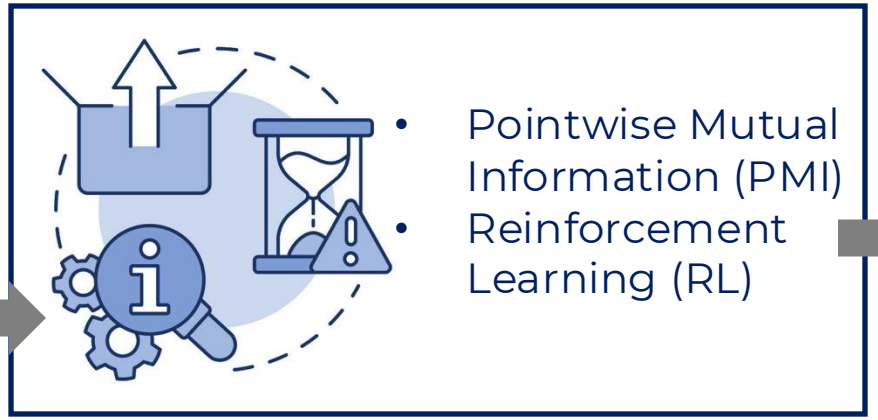
Constructing the INIs: An overview

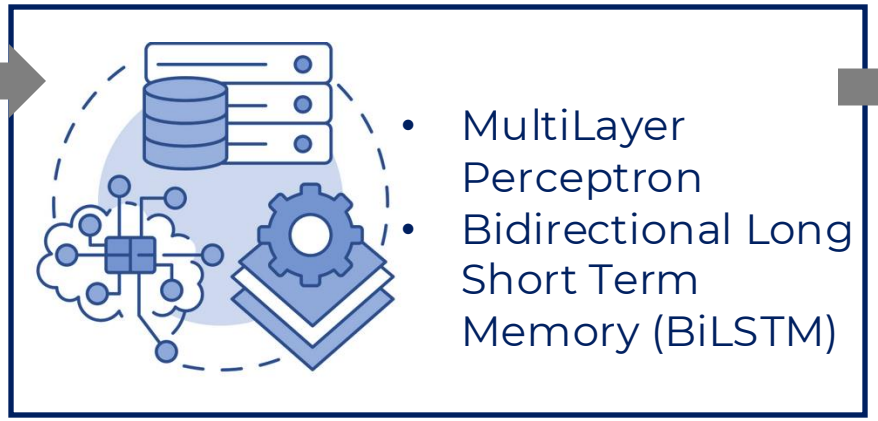
Web scraping

Keyword Filtering

Manual Annotation

Deriving Lexicons





Machine Learning

INI-RL

INI-BiLSTM

Data sources and annotation

- ✓ Data were collected from multiple local news sources. [Sections of interest: economy, banking, finance].
- ✓ Articles are then filtered with the keywords: “inflation”, “price”, “prices”
- ✓ A random sample of 3,000 sentences were annotated

Sample annotated sentences and their classes/labels[

News Text	Label
Aside from flour, other ingredients also went up including LPG, and sugar which substantially increase prices.	1
This was a slight decline from last weeks price level of P37.83/kg.	-1
Some will not pass on the fuel price increases on passengers so they can keep their ticket prices low and haul in more customers.	0

Source: Various media websites, authors’ estimates

Data pre-processing

Pre-processing steps include:

- Removal of extra spaces in the text
- Case normalization
- Removal of punctuations marks
- Unicode conversion
- Tokenization
- Vectorization

Additional steps include:

- Part of speech tagging
- Named entity recognition

Methodology I: Lexicon refined with reinforcement learning

Dictionary based method involves a creating a list of keywords. Pointwise mutual information (PMI) was used to infer word associations with increasing and decreasing inflation

Overall PMI score for a word w:

$$Score(w) = PMI(w|increase) - PMI(w|decrease)$$

Initial lexicon: Words with Score > 0 are classified as increase and words with Score < 0 are classified as decrease

Reinforcement learning was used to improve the initial lexicon

Methodology II: Machine Learning models

Machine learning based method utilizes two artificial neural networks to predict association of text

- MultiLayer Perceptron (MLP)
  - A fully connected feedforward neural network, notable for finding nonlinear relationships.

- Bidirectional Long Short Term Memory (BiLSTM)
  - Contains two LSTM layers in opposite directions. LSTMs are used for finding long term dependencies in sequential data such as time series, text and audio.

Evaluation of methodologies

Both dictionaries and machine learning models were evaluated against test set of the manually-labelled sentences sampled from online news articles

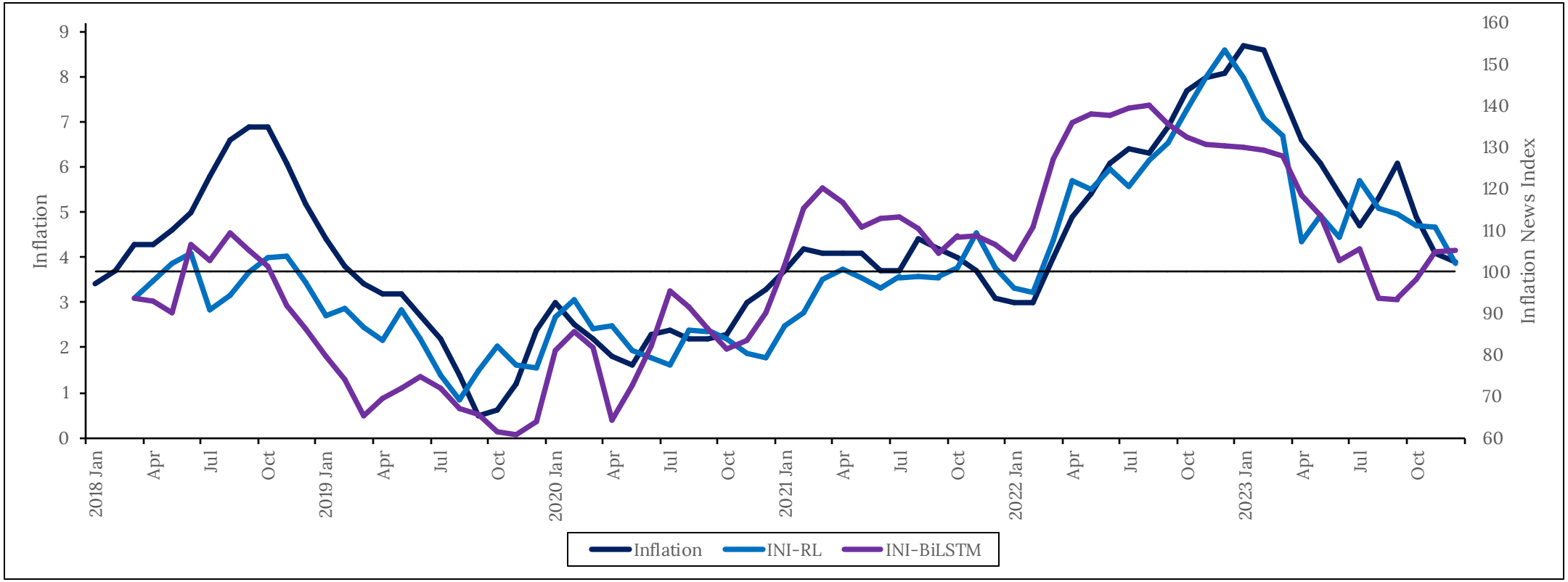
Lexicons	Accuracy (in percent)	Macro F1 (in percent)
Initial lexicon	64.3	41.8
Lexicon with RL	89.3	69.3
Machine Learning Models	**Accuracy (in percent)**	**Macro F1 (in percent)**
MLP	68.9	45.7
BiLSTM(16)	78.4	56.2
BiLSTM(32)	79.8	57.9

Note: Number in parenthesis for BiLSTM refers to number of hidden units.
Source: Authors' estimates

Results

Estimates of INIs strongly co-move with inflation.

INIs vis-à-vis inflation
Q1 2018 to Q4 2023



Source: Authors' estimates

Nowcast evaluation

On average, forecast errors have declined in models augmented with INI-RL compared to baseline models

Nowcast evaluation: Mean Absolute Error

January-December 2023

	Baseline AR	ARX+INI
NCR	0.24	0.24
CAR	0.28	0.26
R1	0.27	0.24
R2	0.36	0.33
R3	0.40	0.34*
R4A	0.33	0.30
R4B	0.30	0.24*
R5	0.31	0.28
R6	0.41	0.37
R7	0.30	0.28
R8	0.22	0.21
R9	0.39	0.40
R10	0.26	0.29
R11	0.29	0.27
R12	0.38	0.38
BARMM	0.38	0.23*
R13	0.16	0.15
Philippines	0.30	0.31
Average	**0.31**	**0.28**

	Baseline SVR	SVR+INI
NCR	0.29	0.28
CAR	0.28	0.04*
R1	0.46	0.33*
R2	0.47	0.32
R3	0.47	0.08*
R4A	0.36	0.14*
R4B	0.30	0.06*
R5	0.38	0.22
R6	0.41	0.40
R7	0.28	0.22
R8	0.27	0.30
R9	0.61	0.59
R10	0.30	0.21
R11	0.29	0.29
R12	0.33	0.35
BARMM	0.37	0.13*
R13	0.20	0.14*
Philippines	0.37	0.03*
Average	**0.36**	**0.23**

Source: Authors' estimates

*Passed Diebold-Mariano test

Nowcast evaluation

On average, forecast errors have declined in models augmented with INI-BiLSTM compared to baseline models

Nowcast evaluation: Mean Absolute Error

January-December 2023

	Baseline AR	ARX+INI
NCR	0.24	0.27
CAR	0.28	0.26
R1	0.27	0.24*
R2	0.36	0.33*
R3	0.40	0.32*
R4A	0.33	0.29
R4B	0.30	0.23*
R5	0.31	0.27
R6	0.41	0.37
R7	0.30	0.30
R8	0.22	0.21
R9	0.39	0.38
R10	0.26	0.27
R11	0.29	0.29
R12	0.38	0.37
BARMM	0.38	0.25*
R13	0.16	0.16
Philippines	0.30	0.29
Average	**0.31**	**0.28**

	Baseline SVR	SVR+INI
NCR	0.29	0.02*
CAR	0.28	0.24*
R1	0.46	0.22*
R2	0.47	0.39
R3	0.47	0.03*
R4A	0.36	0.02*
R4B	0.30	0.02*
R5	0.38	0.12*
R6	0.41	0.45
R7	0.28	0.24
R8	0.27	0.29
R9	0.61	0.60
R10	0.30	0.23
R11	0.29	0.28
R12	0.33	0.33
BARMM	0.37	0.25*
R13	0.20	0.22
Philippines	0.37	0.18*
Average	**0.36**	**0.23**

Source: Authors' estimates

*Passed Diebold-Mariano test

Key findings

- INIs are quantitative summary of digital news coverage on inflation and prices of goods and services
- INIs are found to co-move with actual inflation
- INIs exhibit strong correlation with survey-based inflation expectations of firms and professional forecasters
- INIs contain information that can help nowcast inflation

Sources

- Beck, G. W., Carstensen, K., Menz, J. O., Schnorrenberger, R., & Wieland, E. (2023). Nowcasting consumer price inflation using high-frequency scanner data: Evidence from Germany (No. 34/2023). Deutsche Bundesbank Discussion Paper.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
- Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 16(1), 22-29
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
- Gabriel M., Bautista D., & Mapa C. (2020). Forecasting regional inflation in the Philippines using machine learning techniques: A new approach. Bangko Sentral ng Pilipinas Working Paper Series No. 2020-10.
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63, 3-42
- Macias, P., & Stelmasiak, D. (2019). Food inflation nowcasting with web scraped data (p. 302). Warsaw: Narodowy Bank Polski, Education & Publishing Department.
- Mahadevaswamy, U. B., & Swathi, P. (2023). Sentiment analysis using bidirectional LSTM network. Procedia Computer Science, 218, 45-56
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.
- Serena, J. M., Tissot, B., Doerr, S., & Gambacorta, L. (2021). Use of big data sources and applications at central banks (No. 13). Bank for International Settlements.
- Watkins, C. J., & Dayan, P. (1992). Q-learning. Machine learning, 8, 279-292

Thank you!