Machine Learning, Artificial Intelligence, and Natural Language Processing

Marcelo C. Medeiros

Department of Economics
The University of Illinois at Urbana-Champaign

Big Data, Machine Learning, AI, and Inflation Forecasting

- Forecasting inflation is important
  - Monetary/fiscal policy
  - Pricing derivatives
- Forecasting inflation is hard

Traditional econometric model

$$\pi_{t+h} = \beta_0 + \beta_1 \pi_t + \beta_2 \pi^e_{t+h|t} + \beta_3 X_t + U_{t+h}$$

- π_t is the inflation rate at period t
- $\pi^e_{t+h|t}$ is a measure of future inflation expectation
- X_t is a measure of economic activity at period t

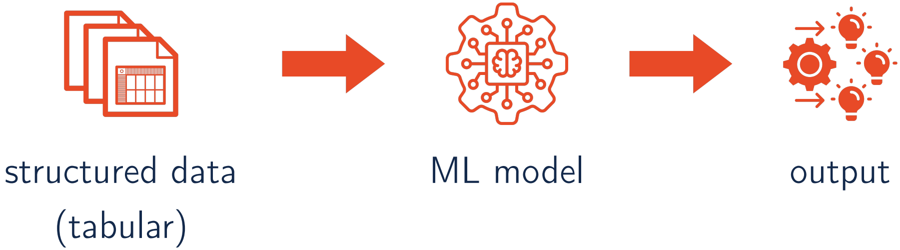
- Challenges
  1. Poor measures of future inflation expectation (especially at higher frequencies). Not the case of Brazil
  2. Many candidates for X_t
- Solutions
  - Drop $\pi^e_{t+h|t}$ from the model
  - X_t is a weighted average of a (small) set of candidate variables (PCA)
- Empirical results
  - Difficult to beat simple benchmark models such as the random walk
    - Atkeson and Ohanian (2001)
    - Stock and Watson (2008)

- New structured datasets with hundreds of variables
- Linear/nonlinear statistical models that are able to handle large datasets
- New age of models:

$$\pi_{t+h} = f(\boldsymbol{Z}_t) + U_{t+h}$$

where $\boldsymbol{Z}_t$ contains many variables including past inflation

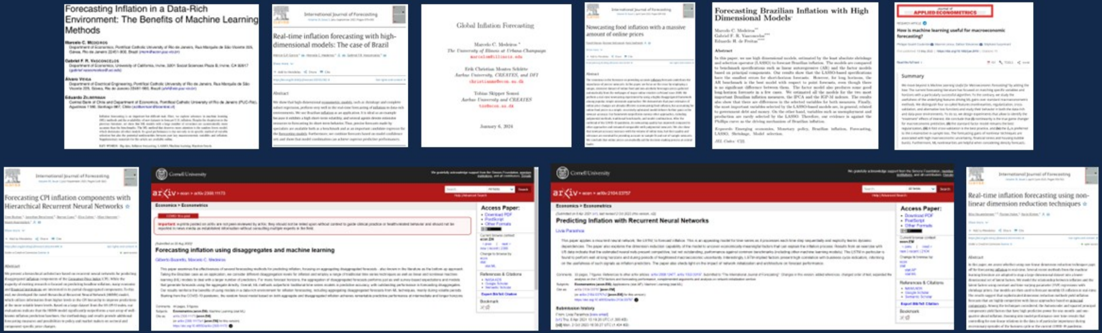
The boom of machine learning



Inflation Forecasting

Even structured (real-time) data can provide better forecasts…

… when carefully estimated ML models are considered



Big Data, Machine Learning, AI, and Inflation Forecasting

The age of "big" data, machine learning, and text data

- Most economic variables are low frequency: monthly, quarterly
- Some variables are only observed with delay
- However, news (text) data are high-frequency
- New methods to transform text into numbers

$$\pi_{t+h} = f(\boldsymbol{Z}_t, \mathsf{News}_t) + U_{t+h}$$

Transforming text into numbers (NLP methods)

- Dictionary methods, topic models, sentiment analysis, ...
- At this point, NLP methods were quite transparent

ECONOMETRICS LAB
QUANTITATIVE RESEARCH @ UIUC

Inflation Forecasting

However, news and social-media-based indexes can help further

The art of transforming text into numbers

Inflation forecasting using unstructured data: the benefits of indexes based on Twitter and newspapers

Gilberto Boaretto
Economics, PUC-Rio & Vos Radar

Marcelo Fernandes
FGV-EESP

Marcelo C. Medeiros
Economics, PUC-Rio

Thiago Milagres
Vos Radar

This draft: August 14, 2023
Preliminary. Please do not share.

Table 1: Out-of-sample RMSE – with respect to Focus (available)

	Focus (available)	Focus (benchmark)	Bias correction (OLS)	Bias correction (w/ indices)	adaLASSO (no indices)	adaLASSO (w/ indices)
infl12m	1.000	0.996	1.121	1.003	*0.703*	**0.614**
infl6m	1.000	0.984	0.957	*0.753*	0.810	**0.673**
infl3m	1.000	0.960	0.917	**0.766**	0.913	*0.902*
infl1m_30d	1.000	0.914	0.914	**0.861**	*0.871*	0.887
infl1m_5d	1.000	**0.878**	0.912	*0.903*	0.917	0.917

Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE | Full Access

Making text count: Economic forecasting using newspaper text

Eleni Kalamara, Arthur Turrell, Chris Redl, George Kapetanios, Sujit Kapadia

First published: 11 May 2022 | https://doi.org/10.1002/jae.2907 | Citations: 5

SECTIONS PDF TOOLS SHARE

Summary

This paper examines several ways to extract timely economic signals from newspaper text and shows that such information can materially improve forecasts of macroeconomic variables including GDP, inflation and unemployment. Our text is drawn from three popular UK newspapers that collectively represent UK newspaper readership in terms of political perspective and editorial style. Exploiting newspaper text can improve economic forecasts both unconditionally and when conditioning on other relevant information, but the performance of the latter varies according to the method used. Incorporating text into forecasts by combining counts of terms with supervised machine learning delivers the highest forecast improvements relative to existing text-based methods. These improvements are most pronounced during periods of economic stress when, arguably, forecasts matter most.

Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE | Open Access

News media versus FRED-MD for macroeconomic forecasting

Jon Ellingsen, Vegard H. Larsen, Leif Anders Thorsrud

First published: 26 July 2021 | https://doi.org/10.1002/jae.2859 | Citations: 9

SECTIONS PDF TOOLS SHARE

Summary

Using a unique dataset of 22.5 million news articles from the *Dow Jones Newswires Archive*, we perform an in depth real-time out-of-sample forecasting comparison study with one of the most widely used data sets in the newer forecasting literature, namely the *FRED-MD* dataset. Focusing on US GDP, consumption and investment growth, our results suggest that the news data contains information not captured by the hard economic indicators, and that the news-based data are particularly informative for forecasting consumption developments.

Text Selection

Bryan Kelly, Asaf Manela & Alan Moreira

Cite this article | https://doi.org/10.1080/07350015.2021.1947843

Full Article | Figures & data | References | Supplemental | Citations | Metrics | Reprints &

Read this article

Abstract

Text data is ultra-high dimensional, which makes machine learning techniques indispensable for textual analysis. Text is often selected—journalists, speechwriters, and others craft messages to target their audiences' limited attention. We develop an economically motivated high-dimensional selection model that improves learning from text (and from sparse counts data more generally). Our model is especially useful when the choice to include a phrase is more interesting than the choice of how frequently to repeat it. It allows for parallel estimation, making it computationally scalable. A first application revisits the partisanship of the U.S. congressional speech. We find that earlier spikes in partisanship manifested in increased repetition of different phrases, whereas the upward trend starting in the 1990s is due to distinct phrase selection. Additional applications show how our model can backcast, nowcast, and forecast macroeconomic indicators using newspaper text, and that it substantially improves out-of-sample fit relative to alternative approaches.

Journal of Econometrics

Can we measure inflation expectations using Twitter?

Cristina Angelico, Juri Marcucci, Marcello Miccoli, Filippo Quarta

Add to Mendeley | Share | Cite

Abstract

Drawing on Italian tweets, we employ textual data and machine learning techniques to build new real-time measures of consumers' inflation expectations. First, we select keywords to identify tweets related to prices and expectations thereof. Second, we build a set of daily measures of inflation expectations around the selected tweets, combining the Latent Dirichlet Allocation (LDA) with a dictionary-based approach, using manually labeled bi-grams and tri-grams. Finally, we show that Twitter-based indicators are highly correlated with both monthly survey-based and daily market-based inflation expectations. Our new indicators anticipate consumers' expectations, proving to be a good real-time proxy, and provide additional information beyond market-based expectations, professional forecasts, and realized inflation. The results suggest that Twitter can be a new timely source for eliciting beliefs.

- LLM: huge volume of data and processing capabilities
- Can we improve on traditional NLP methods?
  - Better indexes
  - Better sentiment analysis
- LLMs are black boxes with potential "data leakage"
  - Mullainathan (2025)
  - Ludwig, Mullainathan, and Rambachan (2025)

- Retrieval-augmented generation (RAG)
  - Enhances LLMs by enabling them to retrieve and integrate external information from knowledge bases or documents
  - Improves the accuracy and relevance of responses, addressing limitations in areas requiring up-to-date or domain-specific information
- LLM agents
  - Sophisticated AI systems built on top of LLMs
  - They are designed to go beyond simple text generation, exhibiting capabilities like planning, reasoning, using tools, and operating autonomously to complete complex, multi-step tasks

- Will we be able to trust LLM models at least as much as we trust our econometric models?
- How can we evaluate LLMs?
  - Out-of-sample validity
  - Stability
  - No hallucinations
  - ...

Many Thanks