Discussion of:
"Word2Prices: embedding central bank communications for inflation prediction"

by Douglas Araujo, Nikola Bokan, Fabio Alberto Comazzi, and Michele Lenza

---

Michael McMahon[1]

Banco de Espana Annual Research Conference, November 2025

[1]University of Oxford and CEPR

Welcome to Economic Data Science!

From an accidental Economic Data Scientist!

Welcome to Economic Data Science!

From an accidental Economic Data Scientist!

Already converging on a DS norm – short paper!

- Not quite the 8 pages of DS...
- But 29 for economics is fantastic!

Comment 2: Economics vs Data Science

Economics use of NLP

NLP typically is a measurement device that improves the inputs that we can use.
Often, a new measure is shown to have some external validation, but there is not the effort to find "best fit" in the measurement phase.

- Methods used are often not complex
- General preference, at least in applied monetary economics, for transparency
- Contribution is rarely the NLP!

Comment 2: Economics vs Data Science

Economics use of NLP

NLP typically is a measurement device that improves the inputs that we can use.

Often, a new measure is shown to have some external validation, but there is not the effort to find "best fit" in the measurement phase.

NLP Research

- The focus is on showing how, in a downstream task, the algorithm improves on our ability to model / understand the language.
- Train and Test use of the data set is fundamental.
- There are explicit measures for assessing fit.

Comment 3: The Exercise and The Channel

$$\left[\ \pi_t\ \right] = \Phi_1(L)\left[\ \pi_{t-1}\ \right] + v_t \tag{1}$$

VS

$$\begin{bmatrix} \pi_t \\ m_t \end{bmatrix} = \Phi_2(L)\begin{bmatrix} \pi_{t-1} \\ m_{t-1} \end{bmatrix} + \nu_t \tag{2}$$

Comment 3: The Exercise and The Channel

Table 1: Results for the full sample

	H=1	H=2	H=3	H=4
	Language Models			
Word2Vec	0.9685	0.9687	0.8593	0.8318
Bert	0.8075	0.7728	0.6756	0.6440
OpenAI	0.7746	0.7479	0.6714	0.7425
	Placebo			
Count Inflation	1.0336	1.0835	1.1016	1.1091
Statement length	1.0157	1.0195	1.0049	1.0030
	Sentiment			
Sent. Inflation	0.9408	0.9639	0.9389	0.9627
Sent. GC	0.9820	0.9805	0.9621	0.9695

What is the source of the value added?

Comment 3: The Exercise and The Channel

$$\begin{bmatrix} \pi_t \end{bmatrix} = \Phi_1(L) \begin{bmatrix} \pi_{t-1} \end{bmatrix} + v_t \tag{1}$$

VS

$$\begin{bmatrix} \pi_t \\ m_t \end{bmatrix} = \Phi_2(L) \begin{bmatrix} \pi_{t-1} \\ m_{t-1} \end{bmatrix} + \nu_t \tag{2}$$

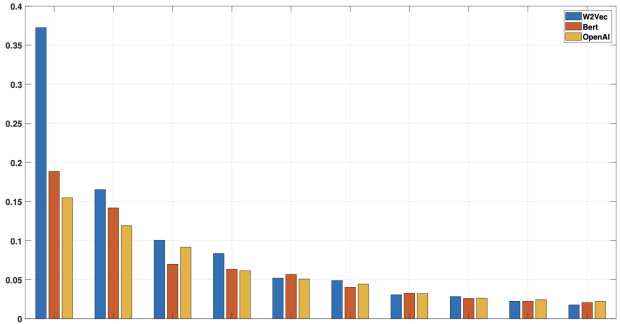
I WANT TO ADD

VS

$$\begin{bmatrix} \pi_t \\ y_t \\ m_t \end{bmatrix} = \Phi_3(L) \begin{bmatrix} \pi_{t-1} \\ y_{t-1} \\ m_{t-1} \end{bmatrix} + \nu_t \tag{3}$$

Comment 4: Measuring Text and Dimensionality Reduction

Figure 3: Variance explained by first ten principal components



Still held back by the dimensionality of the VAR!

Byrne et al (2023): ECB Yield Curve

Regress event news on "narrative signals" - Elastic net regression

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \mathbf{X}_i \beta)^2 + \lambda \left[\frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1 \right] \right\}$$

- $\alpha = 0.99$; estimation via a non-parametric bootstrap - 5000 draws
- Estimate λ by 10-fold cross-validation
- Adjusted R^2 from OLS on the subset of LASSO-selected variables

Byrne et al (2023): ECB Yield Curve

Adjusted R^2, Yield Curve News, Intraday Dependent Variables

Specification	OIS 1M	OIS 1Y	OIS 2Y	DE 5Y	DE 10Y
Controls Only	0.030	0.072	0.072	0.008	0.014
Topics	0.256	0.288	0.282	0.187	0.150
Topics, Past Topics	0.356	0.368	0.357	0.220	0.194
Topics, Future Topics	0.322	0.362	0.368	0.251	0.195
Topics, Future and Past Topics	0.409	0.430	0.427	0.274	0.236

Byrne et al (2023): ECB Yield Curve

MSE, Yield Curve News, Intraday Dependent Variables

Specification	OIS 1M	OIS 1Y	OIS 2Y	DE 5Y	DE 10Y
Topics	100.00	100.00	100.00	100.00	100.00
Topics, Past Topics	93.01	94.95	95.88	99.76	99.19
Topics, Future Topics	97.16	95.53	93.99	97.03	98.39
Topics, Future and Past Topics	90.68	90.74	91.06	97.50	97.55

Byrne et al (2023): ECB Yield Curve

Adjusted R^2, Yield Curve News, Intraday Dependent Variables

Specification	OIS 1M	OIS 1Y	OIS 2Y	DE 5Y	DE 10Y
Controls Only	0.030	0.072	0.072	0.008	0.014
Topics	0.046	0.086	0.087	0.024	0.030
Topics, Past Topics	0.058	0.092	0.093	0.029	0.038
Topics, Future Topics	0.058	0.092	0.092	0.028	0.037
Topics, Future and Past Topics	0.068	0.095	0.096	0.032	0.042

Byrne et al (2023): ECB Yield Curve

Adjusted R^2, Yield Curve News, Intraday Dependent Variables

Specification	Target	Timing	FG	QE	INFO	MPOL
Controls Only	0.020	0.020	0.033	0.096	0.007	-0.023
Topics	0.217	0.280	0.289	0.208	0.165	0.167
Topics, Past Topics	0.275	0.360	0.329	0.262	0.262	0.238
Topics, Future Topics	0.330	0.352	0.342	0.263	0.271	0.251
Topics, Future and Past Topics	0.373	0.415	0.372	0.312	0.339	0.325

Comment 5: Which models?

- Why not consider multiple forecasting models? The objective is prediction so do we care about the structural form of the model?
- Nonlinear treatment of dictionaries?

Comment 5: Which models?

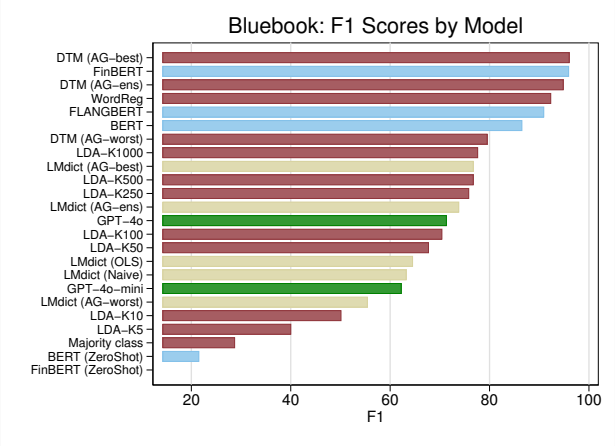
- Why not consider multiple forecasting models? The objective is prediction so do we care about the structural form of the model?
- Nonlinear treatment of dictionaries?

$\Rightarrow$ "EcoFinBench - A Natural Language Processing Benchmark for Economics and Finance" by Ahrens, Gorduza, & McMahon

Bringing the Data Science Approach to Economic Data Science

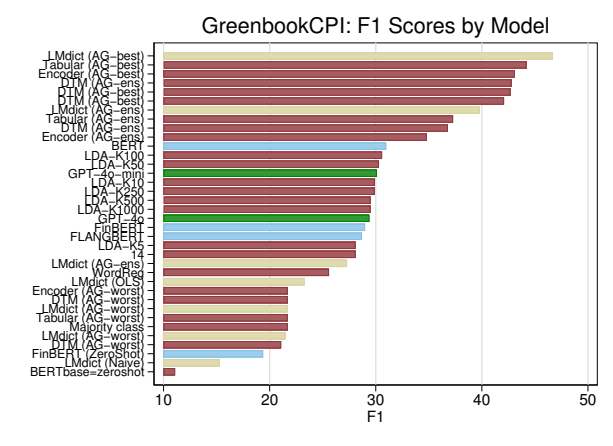
Compare the performance of different NLP approaches across a series of downstream tasks using typical economics or financial datasets.

Results: FOMC Bluebook (#train: 327, #test: 82)



dataset	type	total	train	val	test	neg	neut	pos	mean len	max len	min len
Bluebooks	text	418	64%	16%	20%	6%	75%	17%	2,716	5,934	666
Greenbooks	text+tab	144	64%	16%	20%	48%	6%	47%	3,940	13,063	292

FOMC Greenbook CPI Multimodal (#train: 112, #test: 28)



dataset	type	total	train	val	test	neg	neut	pos	mean len	max len	min len
Bluebooks	text	418	64%	16%	20%	6%	75%	17%	2,716	5,934	666
Greenbooks	text+tab	144	64%	16%	20%	48%	6%	47%	3,940	13,063	292

Comment 6: Real-time Data

"In so doing, we approximately mirror the problem faced by the economists of policy institutions in real-time."

- How important is this? Showing this would be a nice contribution.
- "although it is not possible to exactly attribute this gain to the models' sophistication or to the (partial) in-sample nature of the embedding estimation which, as explained in the previous section, could impart a 'look ahead' bias to the estimates of m_t."
  - It is - by using the word embeddings for the full sample.
  - Bring the problem to all of the measures!

Comment 7: Others

- Placebos: *the count of the word "inflation" in each introductory statement* and *"the length, in words, of each statement.*
  - Not pure placebos: could contain information.
  - Some papers specifically used these!
- Why quarterly and not monthly?

Cross-country applicability

"assessing whether our results hold for other institutions and economies would allow us to unveil whether some specific policy and/or communication practices are more conducive to extract useful information from text, for the purpose of macroeconomic forecasting."

Cross-country applicability

Adjusted R^2, FED Structural Surprises, Intraday Dependent Variables

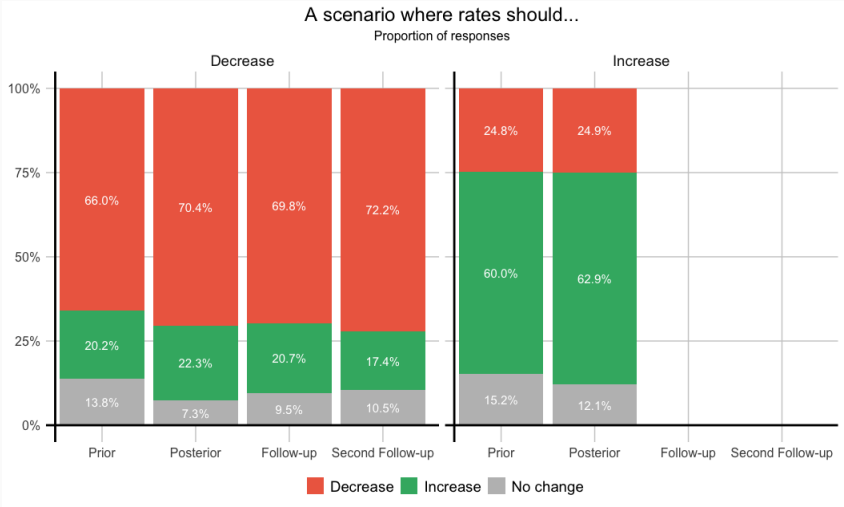
Specification	Target	FG	LSAP	INFO	MPOL
Controls Only	0.175	0.020	0.054	0.115	0.259
Topics	0.317	0.233	0.166	0.234	0.390
Topics, Past Topics	0.373	0.308	0.208	0.275	0.441
Topics, Future Topics	0.380	0.313	0.187	0.273	0.493
Topics, Future and Past Topics	0.431	0.373	0.228	0.309	0.526

Cross-country applicability

Specific question

Scenario: Unexpected rise in inflation and decline in unemployment... inflation has risen unexpectedly from 2 to 4% and the unemployment rate has declined from 5 to 4%. We would like you to provide a forecast for how the Federal Reserve would set interest rates...

Cross-country applicability



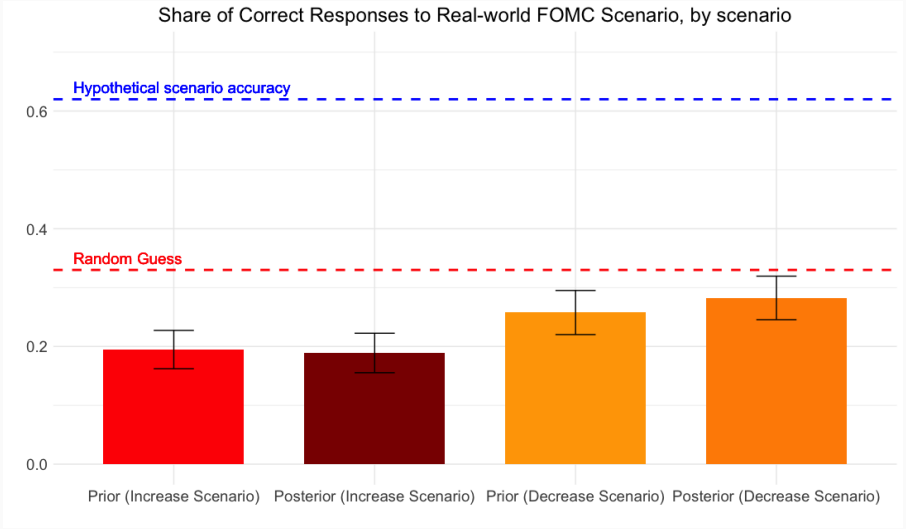
Cross-country applicability

Information received since the Federal Open Market Committee met in December indicates that the labor market has continued to strengthen and that economic activity has continued to expand at a moderate pace. Job gains remained solid and the unemployment rate stayed near its recent low. Household spending has continued to rise moderately while business fixed investment has remained soft. Measures of consumer and business sentiment have improved of late. Inflation increased in recent quarters but is still below the Committee's 2 percent longer-run objective.

Consistent with its statutory mandate, the Committee seeks to foster maximum employment and price stability...[FILL IN THE OPTION].

UP, NO CHANGE, DOWN.

Cross-country applicability



Cross-country applicability

How well can AI do the job?

Type	Method	Overall	A	B	C
Baselines	Human Expert	81.25	90.63	81.25	90.63
	Random	16.67	33.33	33.33	33.33
LLMs	GPT-4	75.00	84.38	75.0	87.5
	FOMC-RoBERTa	53.13	65.63	56.25	71.88
	FOMC-RoBERTa NG Text	34.38	46.88	43.75	53.13
Ranking Methods	RoBERTa Ranking	56.25	71.88	59.38	78.13
	NarrativeGraph	18.75	34.38	31.25	31.25

Conclusion

Conclusion

1. Great to have more people interested!
2. Machine learning models, with fine-tuning, can improve performance of even of simple measurements.
3. Plenty of scope to harness the power of unstructured data.

Overall Take Away

Econ needs to be a bit more like NLP!

Appendix

Classification and the Confusion Matrix

Confusion Matrix:

		Actual	
		0	**1**
Predicted	**0**	tn	fn
	1	fp	tp

- Define the following predictive outcomes from the downstream task:
  - tp $\equiv$ true positives
  - fp $\equiv$ false positives
  - fn $\equiv$ false negatives
  - tn $\equiv$ true negatives
- Define:

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \qquad \text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

- The F1 score is the harmonised mean over precision and recall:

END