Manufacturing Sentiment

Forecasting Industrial Production with Text Analysis

Tomaz Cajner
Norman Morin
Leland D. Crane
Paul E. Soto
Christopher Kurz
Betsy Vrankovich

Federal Reserve Board of Governors

Outline

What we do: Use unique survey data on manufacturing to:

1. Evaluate NLP for classifying sentiment
2. Forecast industrial production
3. Better understand *why* text matters

Outline for today

- Review ISM and IP data
- Measure sentiment
- Forecasting
- Interpreting transformer-based results

Preview of Results

Classifying Sentiment

- Dictionaries exhibit poor performance
  - Short comments contain none of the words in dictionaries
- Transformer-based models are better
  - Particularly "fine-tuned" models

Forecasting Industrial Production

- **Text improves forecasting, but context matters!**
- Sentiment Indices based on Dictionaries
  - Indices based on general dictionaries do not improve forecasting
    - Curated dictionaries (Stability) do
- Sentiment Indices based on Deep Learning (Transformers)
  - Fine-Tuned models perform best
  - Larger gains for forecasting during GFC
- Few words drive variation in Deep Learning models

The ISM Data: Overview

Our main data are from the Institute of Supply Management (ISM)

- Survey has been running since the 1930s
- "The earliest available information for the national economy on any given quarter..."

The ISM Data: Overview

- Survey contacts $\sim$ 100-300 purchasing managers monthly
- Covers manufacturing sector, representative by 3 digit NAICS
- Survey asks about operations, economic conditions
  - Focused on **this month** relative to **last month**
- Purchasing managers complete form
  - **categorical response**
    - 3 Choices: increased, decreased, or stayed the same
  - **open-ended response**
    - written explanation in comment fields

The ISM Data: Details

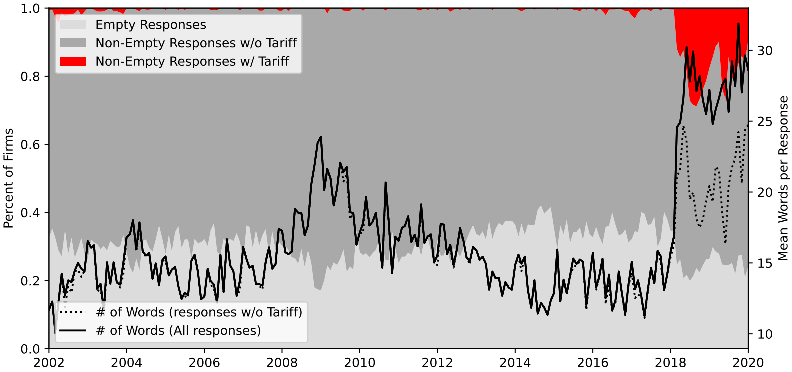
Questions about operations in the ISM survey:

- production levels
- new orders
- orders backlog
- employment
- supplier delivery times
- input inventories
- exports
- imports

The ISM Data: Text form explanations from public data

- "A slowdown in new housing construction and concerns of a slowing economy have customers delaying purchases in an effort to destock." (Chemical Products)
- "While there are lingering concerns about a recession, we are not expecting a large drop-off in manufacturing this year. Worst case is flat." (Nonmetallic Mineral Products)
- The text responses are excerpted in data release but not released publicly

Time series

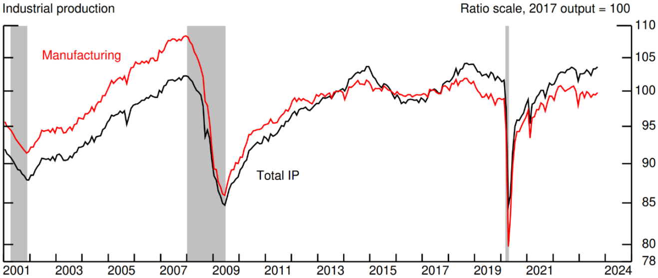


- Black line: Firms averaged 15 words per month, jumped up with 2018 tariffs and stayed high

Industrial Production (IP)

We will be forecasting manufacturing output growth

- IP is a monthly output index
  - Assembled using micro-level data sources
- Highly cyclical:
  - Watched by NBER Business Cycle Dating Committee
- Long time series of monthly data (1919 onwards)



Forecasting Exercises: Data flow

The real time data flow is important:

- The ISM data for month t are typically released on the first business day of month t + 1
- The first IP data for month t are typically released around the 15th of month t + 1
- The IP estimates for a month t are revised several time over the subsequent months and years, as more product data becomes available and benchmark revisions are incorporated.

Forecasting Exercises

Predicting IP once ISM publishes at the beginning of the month

$$\Delta IP_t^{current} = \alpha + \beta_1 \Delta IP_{t-1}^{t^*} + \beta_2 \Delta IP_{t-2}^{t^*} + \beta_3 \Delta IP_{t-3}^{t^*} + \delta x_t^{t^*} + \epsilon_t$$

- $\Delta IP_t^{current}$ is the fully revised, current-vintage growth rate of manufacturing output in month t
- $\Delta IP_{t-h}^{t^*}$ is revised (h times) IP growth from month $t - h$
- $x_t^{t^*}$ collects the ISM metrics for month t
  - Baseline contains only composite PMI index (an average of five ISM diffusion indexes)
- Similar results if we assume econometrician is in 3rd week of month (after IP publishes) link

Measuring sentiment: Overview

Goal:

- Extract positive/neutral/negative sentiment from comments
- Get aggregate sentiment index
- Forecast with it

Dictionary-based methods

Each comment is treated as a *bag of words*

Using specific dictionary, each word is coded as -1/0/+1

- **Harvard** and **AFINN** dictionaries:
  - general purpose/social media focus
  - Economics might have different interpretation of words
- **Loughran and McDonald (LM)**:
  - Specialized dictionary for finance/accounting
  - Based on examination of SEC filings and earnings calls
- **Financial Stability** (Correa et al):
  - Based on central bank financial stability reports

Average word scores to get a sentiment score for the comment

Large Language Models (LLMs): Background

LLMs account for grammar, context-dependent meanings, etc

We mostly use **BERT**, published in 2018 (ancient!)

LLMs are neural networks, mostly transformers

- Each token (word) represented as a vector: embedding.
  - One dimension sentiment, another past/future tense, etc.
- "Attention mechanism":
  - Allows interaction of words, shifting focus, etc.

Transformer-Based: Pretrained Models

FinBERTv1:

- Original BERT weights, fine-tuned on SEC filings
- Sentiment classifier: AnalystTone dataset

FinBERTv2:

- Original BERT weights, fine-tuned on Reuters financial news
- Sentiment classifier: FinancialPhrasebank dataset

Both models are trained on data from the financial world
ISM comments are mostly about backlogs, inventories, production disruptions, weather, shipping times, etc.

Transformer-Based: Human Labelled Data

- Two economists hand-label 1,000 comments for sentiment: positive (+1), neutral (0), and negative (-1)
- "Is this comment consistent with manufacturing IP rising month over month?"
- We agreed on about 700 comments, train on most of these, keep a hold-out sample for evaluation

Models:

- Fine-tuned BERT: Human Labelled
  - Original BERT weights, fine-tuned classifier on our labels
- Transformer-small ("TF-small"):
  - Encoder-only transformer trained from scratch on our labels

Transformer-Based: Production Data

- Naturally occurring labels
- Exploits panel structure of survey data
- Predict firm f's $PRODVAL_{t+1}$ using $Text_t$

Models:

- Fine-tuned BERT: Production Data
    - Original BERT weights, fine-tuned classifier on PROD labels

Transformer-Based:

Our fine-tuned models respect forecasting timing

- 2018M1-2020M1 Out of Sample
  - 2001-2017 used for fine-tuning
- 2007M12-2009M6 Out of Sample
  - 2001-2007M11 used for fine-tuning

Types of Sentiment Measures

Dictionary-Based:

- Harvard, AFINN, Loughran/McDonald, Financial Stability

Transformer-Based:

- FinBERT
- Small Transformer
  - Using Human Labelled Data
- Fine-Tuned BERT
  - Using Human Labelled Data
  - Using Production Data

Summary of Sentiment and Activity Measures

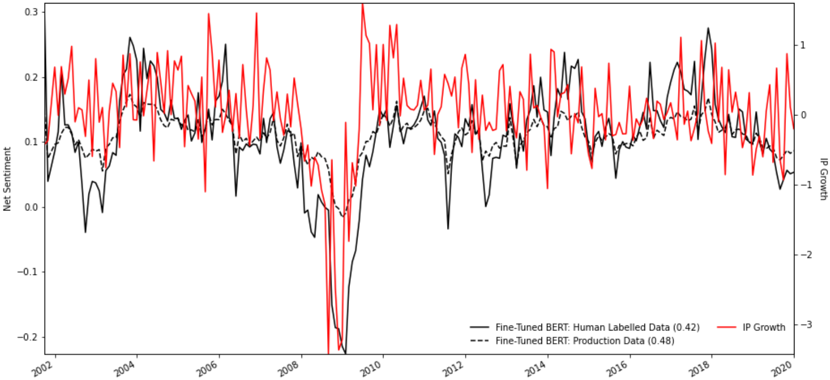
N=219	Mean	Std. Dev.	Min	Median	Max
Text Measures					
LM	-0.0096	0.0064	-0.0399	-0.0097	0.0046
Harvard	0.0005	0.0048	-0.0216	0.0009	0.0138
AFINN	0.0123	0.0109	-0.0300	0.0115	0.0374
Stability	-0.0012	0.0040	-0.0233	-0.0012	0.0095
FinBERT (v1)	-0.0379	0.1111	-0.4454	-0.0313	0.2019
FinBERT (v2)	-0.0633	0.1024	-0.4882	-0.0442	0.1561
TF-Small	0.1864	0.1403	-0.2559	0.1954	0.9813
Fine-Tuned BERT: Human Labeled Data	0.1082	0.0810	-0.2261	0.1156	0.3141
Fine-Tuned BERT: Production Data	0.1100	0.0310	-0.0162	0.1128	0.1733
Macro Variables					
IP $Growth_t$	0.0335	0.7041	-3.4210	0.0406	1.5950
ISM_PMI_t	53.0959	4.6551	34.5000	53.2000	61.4000
$ISM\_NewOrders_t$	55.8511	6.6517	25.9000	56.6000	71.3000
$ISM\_Inventories_t$	47.9950	4.2814	33.5000	48.6000	56.8000

Accuracy Scores on Unseen Human Labeled Data (Test Set: 2018M1-2020M1)

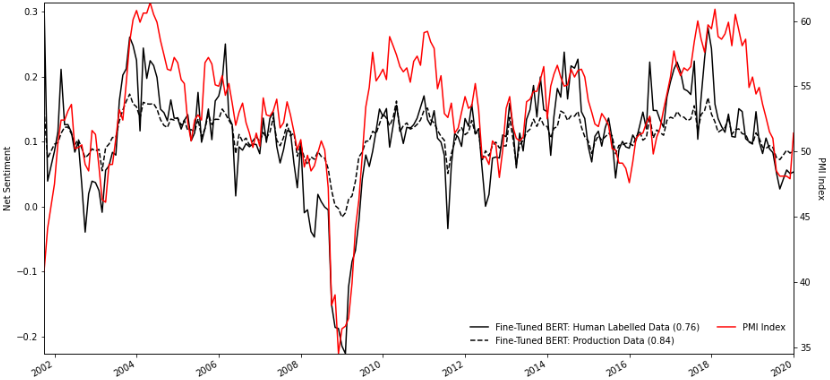
Model	Accuracy (percent)	Rescaled
AFINN	27.9	68.5
Harvard	24.3	65.8
LM	20.7	75.7
Stability	11.7	70.3
FinBERTv1	70.3	73.0
FinBERTv2	56.8	72.1
TF-Small	67.6	73.0
Fine-Tuned BERT: Human Labelled Data	82.9	-
Fine-Tuned BERT: Production Data	4.5	87.4

Sentiment-based Transformers do better...

Industrial Production and Sentiment



PMI and Sentiment



Forecasting Exercises: In Sample Results

	(1)	(2)	(5)	(6)	(8)	(9)	(10)
	Dependent Variable: IP $Growth_t$						
		Dictionary Based Methods		*Deep Learning Methods*			
Text Measure	*Baseline*	LM	Stability	FinBERT (v1)	TF-Small	Fine-Tuned BERT: Human Labelled Data	Fine-Tuned BERT: Production Data
$ISM\_Sentiment_t$		0.0917	0.163***	0.159*	0.0995*	0.138*	0.244***
		(0.0583)	(0.0575)	(0.0829)	(0.0541)	(0.0727)	(0.0929)
ISM_PMI_t	0.0660***	0.0611***	0.0673***	0.0500***	0.0614***	0.0518***	0.0346**
	(0.0147)	(0.0140)	(0.0148)	(0.0152)	(0.0139)	(0.0141)	(0.0159)
IP $Growth_{t-1}$	-0.0303	-0.0468	-0.0491	-0.0539	-0.0389	-0.0488	-0.0497
	(0.0908)	(0.0882)	(0.0866)	(0.0894)	(0.0897)	(0.0884)	(0.0863)
IP $Growth_{t-2}$	0.0611	0.0437	0.0245	0.0246	0.0434	0.0370	0.0253
	(0.0947)	(0.0905)	(0.0874)	(0.0953)	(0.0930)	(0.0926)	(0.0922)
IP $Growth_{t-3}$	0.0248	0.00621	0.000205	-0.000883	-0.00317	0.00462	-0.0161
	(0.0963)	(0.0954)	(0.0947)	(0.0992)	(0.0959)	(0.0955)	(0.0987)
Observations	219	219	219	219	219	219	219
R-squared	0.219	0.228	0.244	0.234	0.231	0.230	0.245

Stability and transformer models do well in sample

Forecasting Exercises: Out of Sample Results (2018-2020)

In-Sample: 2001M11-2017M12

Out-of-Sample: 2018M1-2020M1

The Stability and transformer models mostly do well OOS

Interpretability

- BERT has many advantages: picks up on word context, good forecasting performance.
- But, it is a black box.
- Can we approximate BERT with something like a dictionary?

Quick answer: Yes!

Interpretability

Step 1: Get contribution of each word to a comment's score

- Use Shapley decompositions; good properties, additive

Step 2: Get time-invariant average contributions for each word

- Simple average of the Shapley scores, decent approximation

Step 3: Only keep extreme-valued words

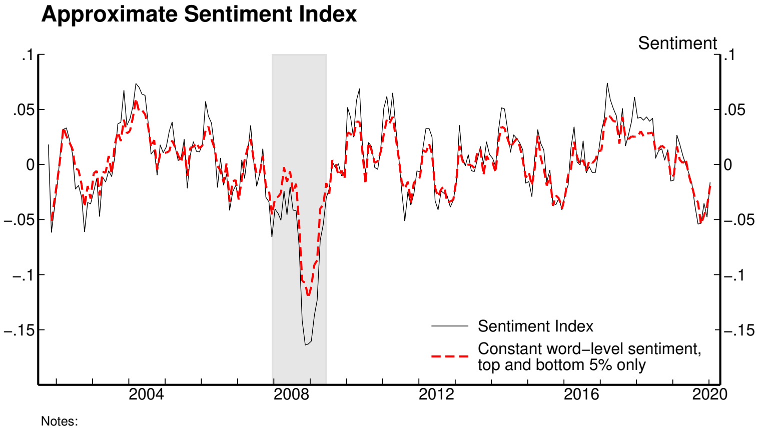
- Top and bottom 5% of words account for most of the action

Most positive/negative words

Table 7: Average Net Positive Scores

Positive Words	Score	Negative Words	Score
specials	0.055	weak	-0.063
improved	0.053	inability	-0.064
excellent	0.051	fragile	-0.064
booming	0.049	decline	-0.066
upbeat	0.048	downward	-0.066
improves	0.048	declining	-0.068
improvement	0.047	downs	-0.069
improve	0.046	weakening	-0.070
increase	0.045	depressed	-0.071
good	0.044	weaken	-0.072
rum	0.043	discontinued	-0.073
launch	0.041	slow	-0.075
brisk	0.040	offs	-0.075
increased	0.040	insufficient	-0.076
increasing	0.036	instability	-0.080
heightened	0.033	slowing	-0.081
upgrade	0.033	slug	-0.084
advantages	0.033	erosion	-0.085
lift	0.032	errors	-0.093
doubled	0.032	unstable	-0.105

Interpretability



Dictionary-based approximation (red) tracks the BERT-based index well. **We can get back to an interpretable index**

Conclusions and Next Steps

- New, useful data
  - Text covering the operations of manufacturers
- Transformers do well classifying comments
  - Especially after fine-tuning
- Aggregate sentiment index has forecasting power
  - Reduce OOS MSE $\sim$ 2-6%
  - Particularly important during GFC

Thank You!

christopher.j.kurz@frb.gov

Appendix: The ISM Data Details

ISM publishes diffusion indexes summarizing the categorical responses

- Ranges between 0 and 100, 50 is neutral. Formula:

$$
\begin{aligned}
D_t = & 100 \times (\text{Fraction saying production is higher}) \\
& + 50 \times (\text{Fraction saying production is the same})
\end{aligned}
$$

- Rescaled (more intuitive?) version, range (-1,1):

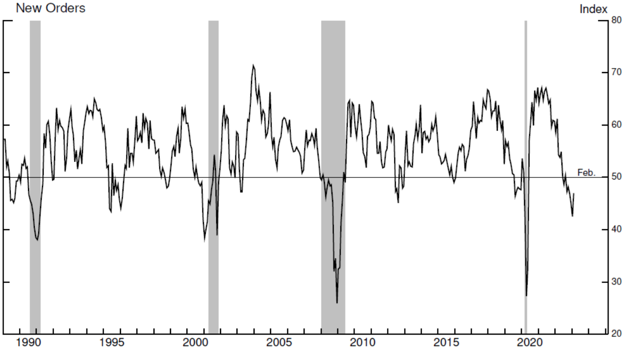
$$
\begin{aligned}
D_t' = & (\text{Fraction saying production is higher}) \\
& - (\text{Fraction saying production is lower})
\end{aligned}
$$

  ranges between -1 and +1.

- Closely watched for signs of recession/recovery

Appendix: The ISM Data Details

- Our dataset covers the roughly 42,000 firm-month observations
- Dates covered: 2001 to 2020



Appendix: The ISM Comment Details

Survey asks for free-response comments, typically 1-2 sentences

Two types of comments:

- General Remarks
- Comments on individual survey questions: why is X higher/the same/lower?

Date	NAICS3	General Remarks	Production higher/same/lower	Production comments	New orders higher/same/lower	New orders comments
10/1/2008	332	Business activity has decreased noticeably due to economic conditions.	Lower than a month ago	economy	Lower than a month ago	economy
4/1/2018	311	Labor shortage in our area is our biggest concern	Same as a month ago	Labor constraints	Higher than a month ago	New orders are coming in. Export demand is solid
9/1/2018	327	Distributors and Manufactures are pushing increase due to tariffs.	Same as a month ago		Same as a month ago	

Appendix: Text summary statistics

Table 2: Survey Summary Statistics

Field	(1) Fraction W/ Text	(2) Mean Word Count	(3) Mean Word Count Cond. on Text
General Remarks	0.49	8.21	16.73
Production	0.27	1.47	5.53
New Orders	0.26	1.50	5.70
Backlog	0.19	1.20	6.46
Employment	0.01	0.07	5.10
Supplier Speed	0.12	0.92	7.72
Input Inventories	0.23	1.58	6.81
Exports	0.11	0.63	6.01
Imports	0.12	0.81	6.64
All Text (Appended)	0.68	16.40	24.27

Appendix: Sentiment Example

Look at a comment and sentiment according to different methods

- Overall class: Is the sentence classified as positive ($+1$), neutral (0), or negative (-1)?
- Classification of individual words:
  - Dictionary methods: Overall class is a direct function of the individual words
  - BERT models: Not so simple. E.g. all negative/neutral words can be positive when put into a sentence

Appendix: Sentiment Example: Continued

"Demand has been higher than capacity", human-coded as positive

Method	demand	has	been	higher	than	capacity	**Overall class**
Harvard	0	0	0	0	0	0	0
Afinn	-1	0	0	0	0	0	-1
Stability	0	0	0	0	0	0	0
FinBERT_v1	0	0	0	1	0	0	1
FinBERT_v2	0	0	0	0	0	0	1
Fine-tuned BERT	0	1	1	1	1	0	1

BERT models classify sentences well, but hard to interpret in terms of individual words

Appendix: Large language models (LLMs)

Popular types of transformers:

- GPT3: ("decoder only") trained on next word prediction, given the text to that point
  - Good for generating text given a prompt
- BERT: ("encoder only") Trained on predicting missing words, and guessing whether pairs of sentences match
  - Produces a good embedding summarizing the meaning of a sentence

Since 2022: Instruction tuning

- Generative transformers just try to continue the prompt text: okay, but not great.
- GPT3.5/chatGPT: Collect human responses to prompts, fine-tune model to mimic human responses.
- Flan, Alpaca: other approached to instruction tuning

Appendix: Forecasting Exercises: Specification at End of Month

Assessing predictive power of sentiment measures day after IP publishes to predict next month's IP (3rd week of the month)

$$\Delta IP_{t+1}^{current} = \alpha + \beta_1 \Delta IP_t^{t^*} + \beta_2 \Delta IP_{t-1}^{t^*} + \beta_3 \Delta IP_{t-2}^{t^*} + \delta x_t^{t^*} + \epsilon_t$$

- where $\Delta IP_t^{current}$ is the fully revised, current-vintage growth rate of manufacturing output in month t
- $\Delta IP_t^{t^*}$ is the initial estimate of IP
- $x_t^{t^*}$ collects the ISM metrics for month t
- For the baseline model x_t contains only the the composite PMI index, an average of five of the ISM diffusion indexes

return