Text Mining for Economics and Finance

Sentiment and Forecasting

Stephen Hansen
University of Oxford

Introduction

Forecasting with unstructured data is key area of interest for central banks.

In recent survey of 52 central banks, ‘forecasting’ and ‘nowcasting’ are two most-cited potential applications of Big Data.

Jyry Hokkanen, the head of Statistics at Sveriges Riksbank was asked “what are the key big data research questions—what needs to be answered?” His answer was:

> *“I like text-mining techniques. I’d like that to become stable, because there’s huge potential in it. We do communicate and central banks produce text, but we also receive a lot of text, and you can monitor society and the economy via tweets, speeches, articles, and company statements.”*

This is a very new area, and here I will give some highlights for text data in particular.

Approaches

1. Use text to measure sentiment, which is then related to macroeconomic outcomes.

2. Use 'off-the-shelf' dimensionality reduction algorithm, treat low-dimensional representation of text as data input into standard forecasting model.

3. Gap in the literature: joint model of unstructured data and macroeconomic observables.

Why Sentiment?

There are two motivations for measuring sentiment to predict economics conditions:

1. It may reflect hard information about economic conditions held by the public but not yet present in data.
2. Sentiment may be a cause of business cycle fluctuations independent of hard information.

If the goal is prediction, we can stay silent on the precise channel, and simply exploit the correlation between sentiment and outcomes.

How to Measure Sentiment?

The first step in measuring sentiment is to identify relevant texts.

Two common sources are traditional print media and social media. Trade-off between quality of content, timeliness, and population sample size.

The second step is to extract sentiment information from the texts.

Distinction between dictionary methods and (supervised) machine learning approaches.

Tetlock (2007)

Tetlock (2007) is a highly cited paper that applies dictionary methods to the Wall Street Journal's "Abreast of the Market" column.

Uses Harvard IV-4 dictionaries
`http://www.wjh.harvard.edu/~inquirer`.

Large number of categories: positive, negative, pain, pleasure, rituals, natural processes, etc. 77 in all.

Count number of words in each dictionary in each column from 1984-1999.

Principal components analysis shows most variation on dimensions that reflect pessimism: negative, weak, fail, fall.

Tetlock (2007)

Tetlock (2007) is a highly cited paper that applies dictionary methods to the Wall Street Journal's "Abreast of the Market" column.

Uses Harvard IV-4 dictionaries `http://www.wjh.harvard.edu/~inquirer`.

Large number of categories: positive, negative, pain, pleasure, rituals, natural processes, etc. 77 in all.

Count number of words in each dictionary in each column from 1984-1999.

Principal components analysis shows most variation on dimensions that reflect pessimism: negative, weak, fail, fall.

Main result: pessimism predicts low short-term returns (measured with the Dow Jones index) followed by reversion.

Loughran and McDonald (2011)

Following Tetlock (2007), popular to use just negative word dictionary from Harvard IV-4.

This includes words like ‘tax’, ‘cost’, ‘capital’, ‘liability’, and ‘vice’.

Unclear that these are appropriate for describing negative content in financial context.

Loughran and McDonald (2011) use 10-K filings to define their own finance-specific word lists, available from `http://www3.nd.edu/~mcdonald/Word_Lists.html`.

Negative list includes words like ‘restated’, ‘litigation’, ‘termination’, ‘unpaid’, ‘investigation’, etc.

Loughran and McDonald (2011)

Following Tetlock (2007), popular to use just negative word dictionary from Harvard IV-4.

This includes words like ‘tax’, ‘cost’, ‘capital’, ‘liability’, and ‘vice’.

Unclear that these are appropriate for describing negative content in financial context.

Loughran and McDonald (2011) use 10-K filings to define their own finance-specific word lists, available from `http://www3.nd.edu/~mcdonald/Word_Lists.html`.

Negative list includes words like ‘restated’, ‘litigation’, ‘termination’, ‘unpaid’, ‘investigation’, etc.

Main result: the context-specific list has greater predictive power for return regressions than the generic one.

Social Media Data

Social media data is another data source for measuring sentiment.

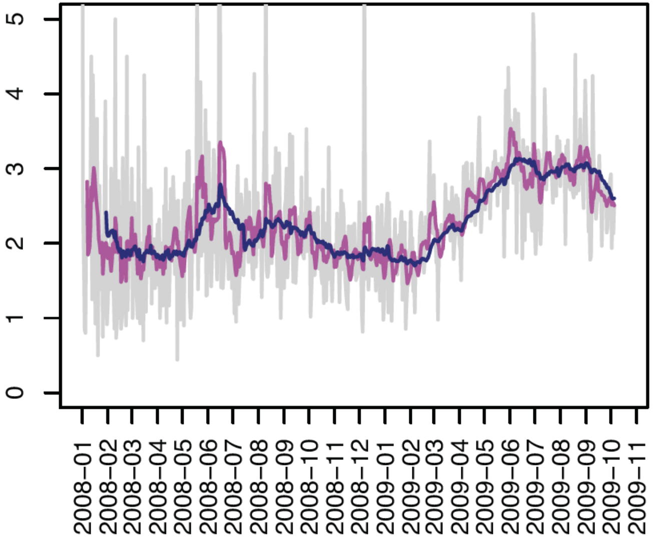
O'Connor et. al. (2010) use Twitter data to track consumer confidence, as measured by the US Index of Consumer Sentiment.

Two challenges: (1) identify relevant tweets; (2) measure sentiment within relevant tweets.

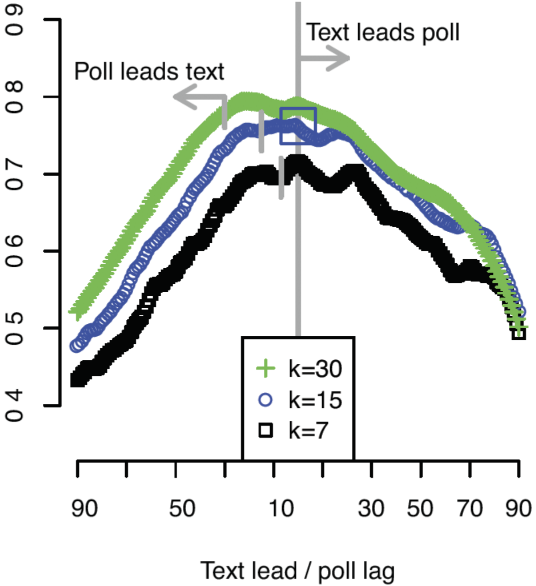
For (1), use all tweets that contain word 'economy', 'job', and 'jobs'.

For (2), use positive and negative words from OpinionFinder. Tweet is positive (negative) if it contains any positive (negative) word; day t sentiment score is ratio of positive to negative messages.

Sentiment Index (Daily, Weekly and Monthly Smoothing)



Correlation with ICS



Theoretically Grounded Dictionaries

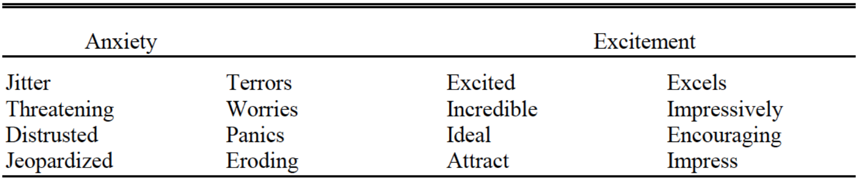
Nyman et. al. (2018) use dictionaries grounded in psychological theory to characterize emotional states that ground people's actions.

The index is applied to three different sources of text:

1. Bank of England market commentary.
2. Broker reports.
3. Reuters news archive.

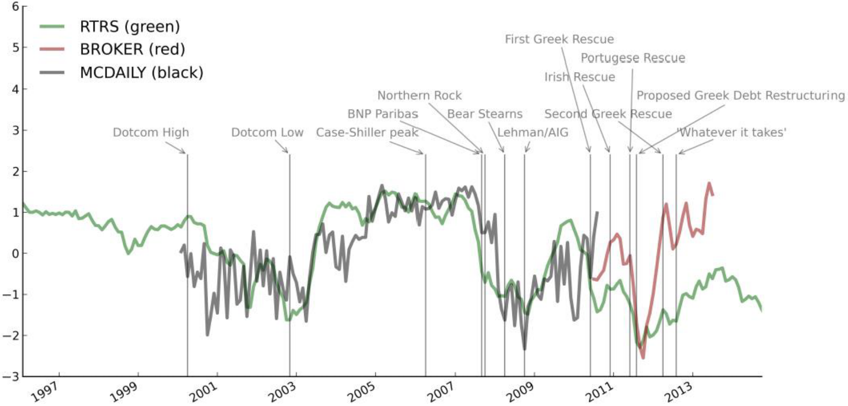
Dictionary

Table 1: Emotion dictionary samples

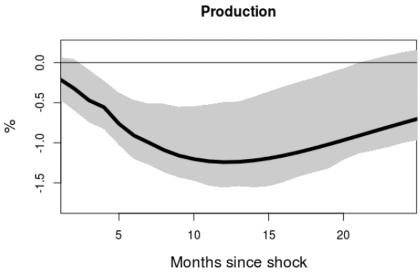


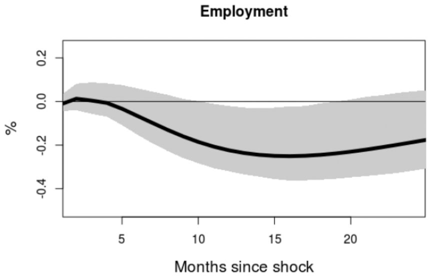
Anxiety		Excitement	
Jitter	Terrors	Excited	Excels
Threatening	Worries	Incredible	Impressively
Distrusted	Panics	Ideal	Encouraging
Jeopardized	Eroding	Attract	Impress

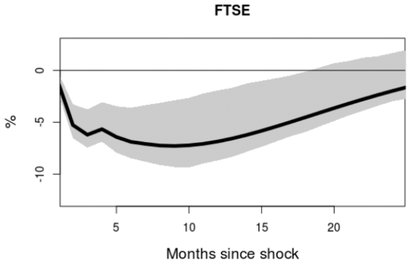
Index


RTRS (green)
BROKER (red)
MCDAILY (black)
First Greek Rescue
Portugese Rescue
Irish Rescue
Proposed Greek Debt Restructuring
Northern Rock
BNP Paribas
Bear Stearns
Second Greek Rescue
Dotcom High
Dotcom Low
Case-Shiller peak
Lehman/AIG
'Whatever it takes'
6
5
4
3
2
1
0
−1
−2
−3
1997
1999
2001
2003
2005
2007
2009
2011
2013

VAR Results


Production
%
0.0
-0.5
-1.0
-1.5
5
10
15
20
Months since shock


Employment
%
0.2
0.0
-0.2
-0.4
5
10
15
20
Months since shock


FTSE
%
0
-5
-10
5
10
15
20
Months since shock

Machine Learning Approach

Shapiro et. al. (2018) use a machine learning algorithm to estimate sentiment in news articles.

They solve the problem of where to get objective labels by using the ‘Experience Project’ social network website, which operated from 2007-2016.

This website had extensive free-form writing about various objectively labeled emotional states.

The authors then use a proprietary machine learning algorithm to build a mapping from words to emotions.

The algorithm is then applied out-of-sample to a collection of news texts from the US.

The resulting sentiment index is again correlated with macroeconomic outcomes.

Direct Forecasting

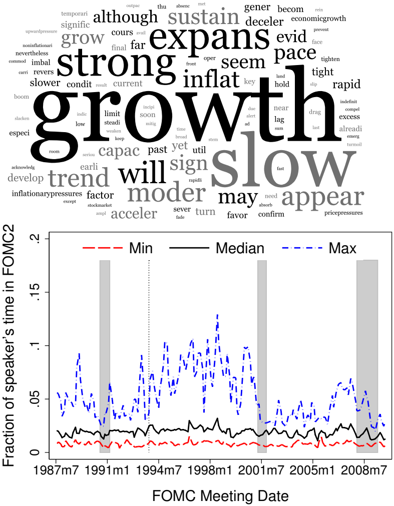
The measurement of sentiment for forecasting takes a very specific stance on what information is likely to predict fluctuations.

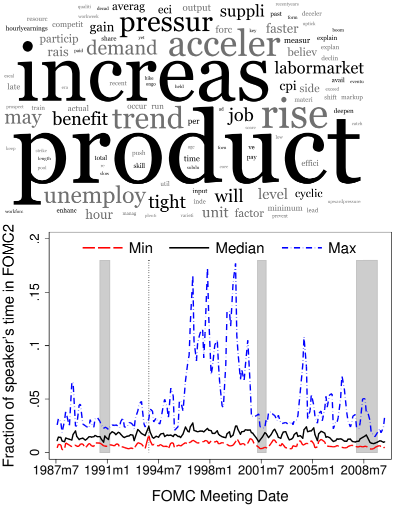
If aspects of content besides sentiment contain relevant information, they will be missed.

One strategy growing in popularity is to reduce the dimensionality of text with LDA, then use the topic shares as covariates in a standard time series model.

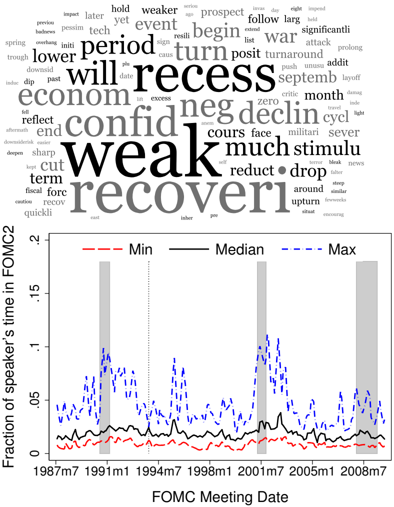
Examples include Thorsrud (2018) and Larsen and Thorsrud (2018) in macroeconomic forecasting; and Mueller and Rauh (2018) in conflict forecasting.

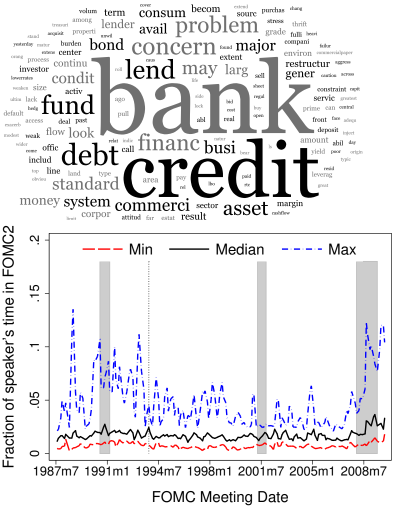
Pro-Cyclical Topics



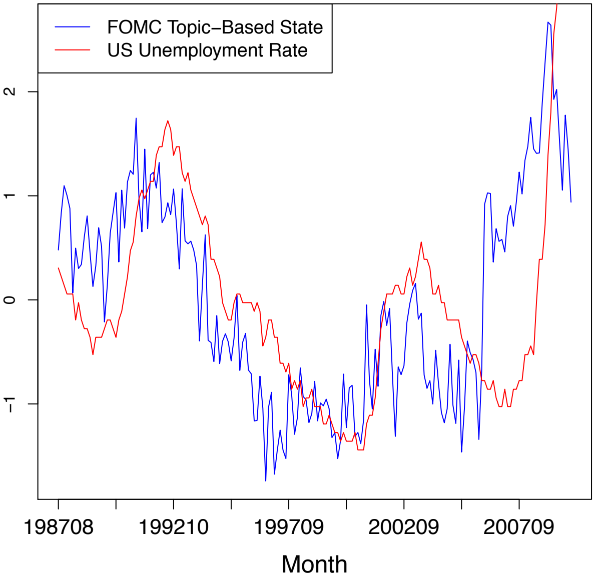


Counter-Cyclical Topics

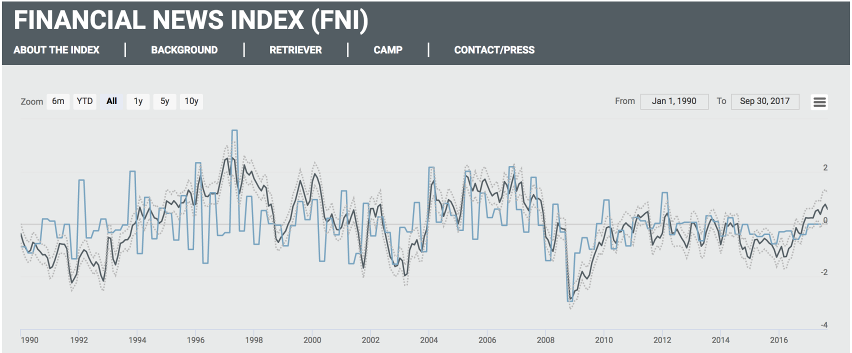




Derived State-of-the-World



Financial News Index



FINANCIAL NEWS INDEX (FNI)

From Jan 1, 1990 To Sep 30, 2017

An Inherent Tension

This way of generating forecasts is clearly a natural and important first step, but has some limitations.

The algorithm used to construct the topic shares treats all documents as independently drawn at each point in time, which is precisely what we think is *not* true if text moves with the business cycle.

Another issue is that treating the topic shares as data ignores that they are derived from an auxiliary statistical model whose uncertainty is not addressed.

Need for joint model of unstructured data and macro observables?

Conclusion

High demand for forecasting with unstructured data, likely to grow in the future.

Basic tools from our course, combined with standard time series machinery, represents the state-of-the-art.

This has already led to the development of new and useful indicators.

We are still in the world of ‘small data’ insofar as there are a limited number of observations to predict, even with new, rich data sources.

Opens need for new modeling tools in my opinion.