Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC – Bank Indonesia International Workshop and Seminar on *"Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data"*

Bali, Indonesia, 23-26 July 2018

# Measuring market and consumer sentiment and confidence[1]

Stephen Hansen,

University of Oxford

[1] This presentation was prepared for the meeting. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

# Measuring Market and Consumer Sentiment and Confidence

## Bank Indonesia—IFC Workshop

Stephen Hansen
University of Oxford

# Introduction

The first lecture focused mainly on recovering the underlying structure of text.

The methods we discussed did not seek to directly explain any label associated with text.

In many cases, we are interested in mapping the content of text into some outcome variable of interest.

One prominent example of this is sentiment analysis, in which we wish to associate text with the sentiment it reflects.

We will again see a distinction between word-count exercises and 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.

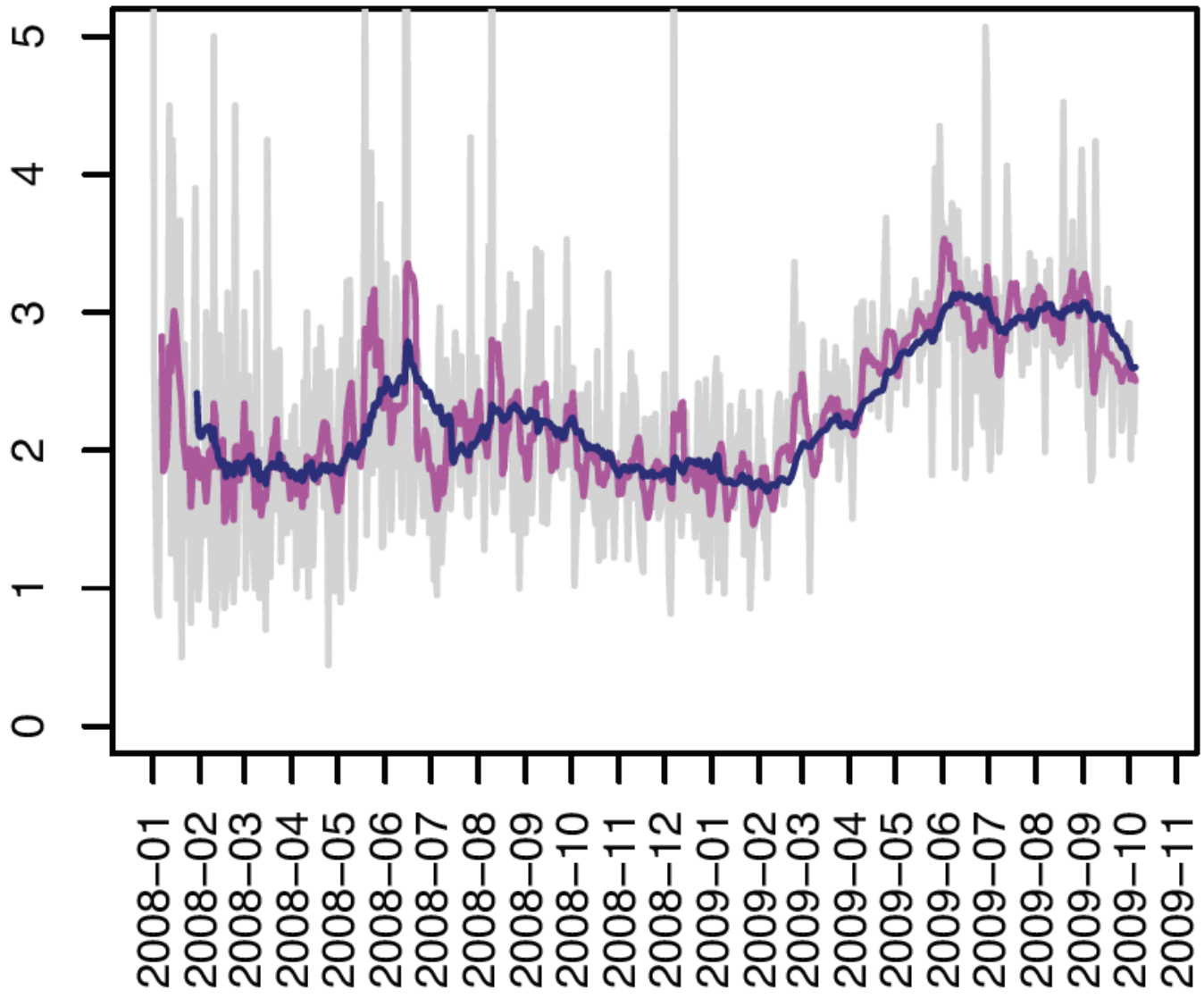
O'Connor et. al. (2010) use Twitter data to track consumer confidence, as measured by the US Index of Consumer Sentiment.

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

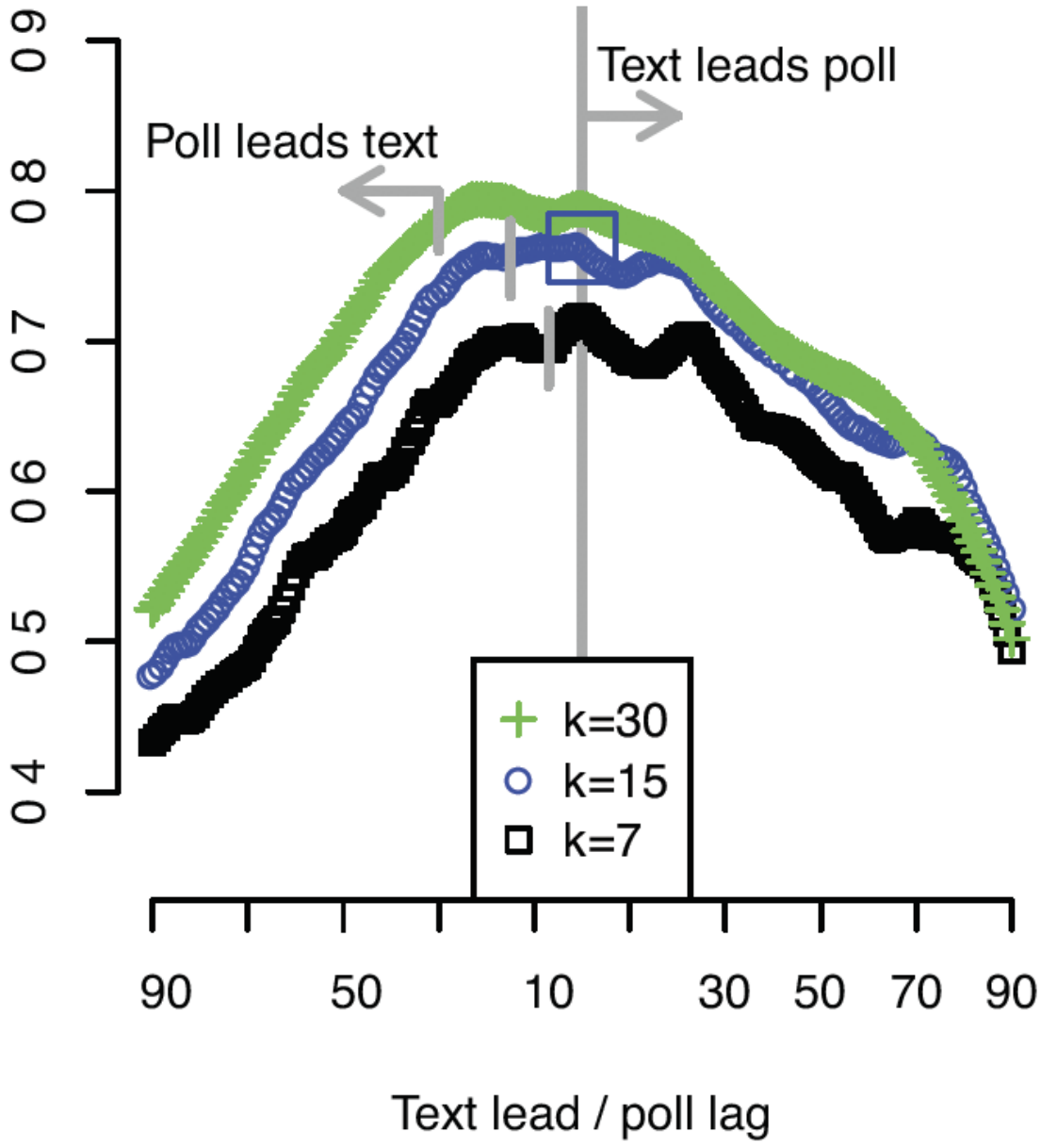
For (1), use all tweets that contain word 'economy', 'job', and 'jobs'.

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:

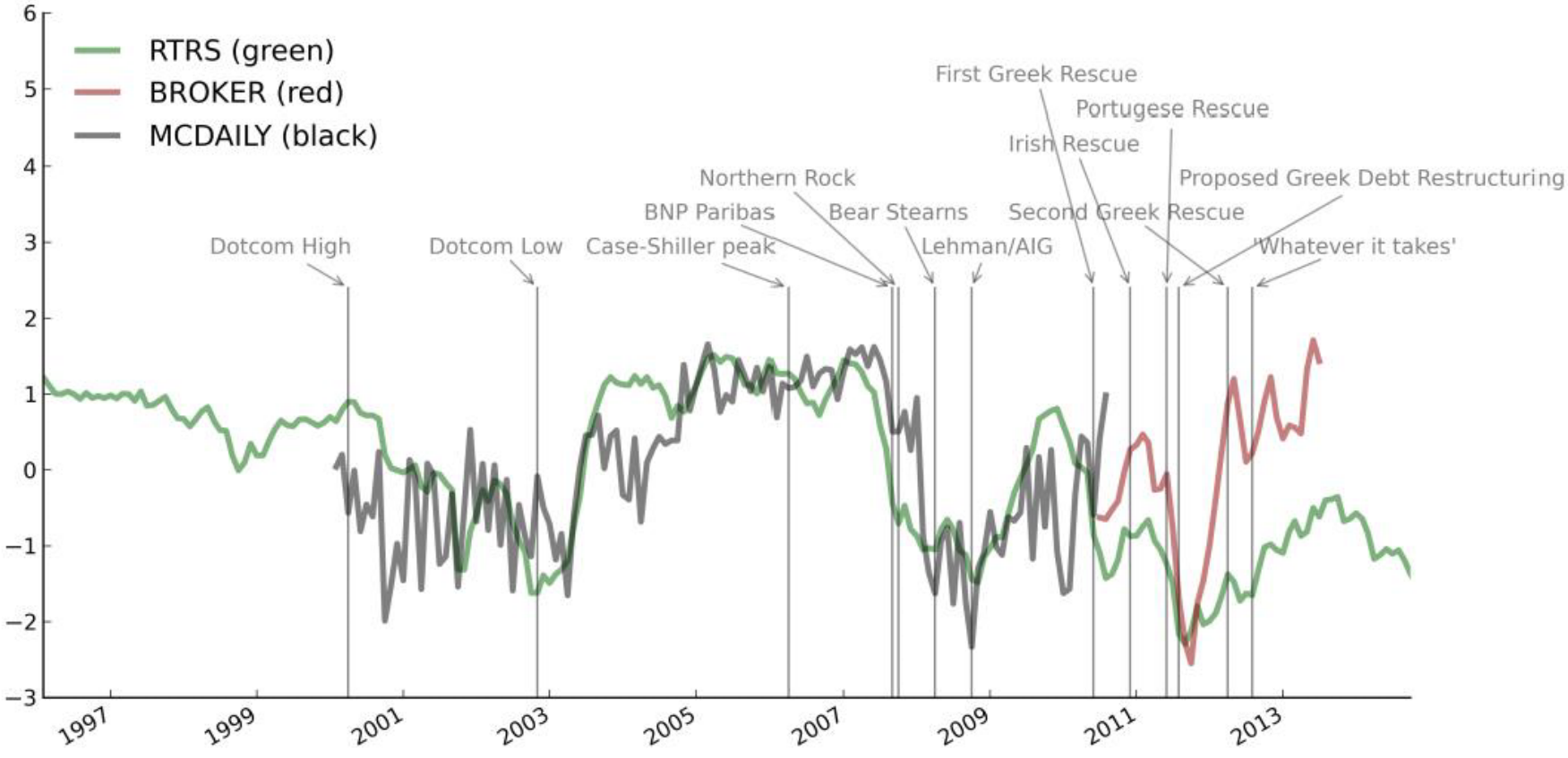
1. Bank of England market commentary.
2. Broker reports.
3. Reuters news archive.

# Dictionary

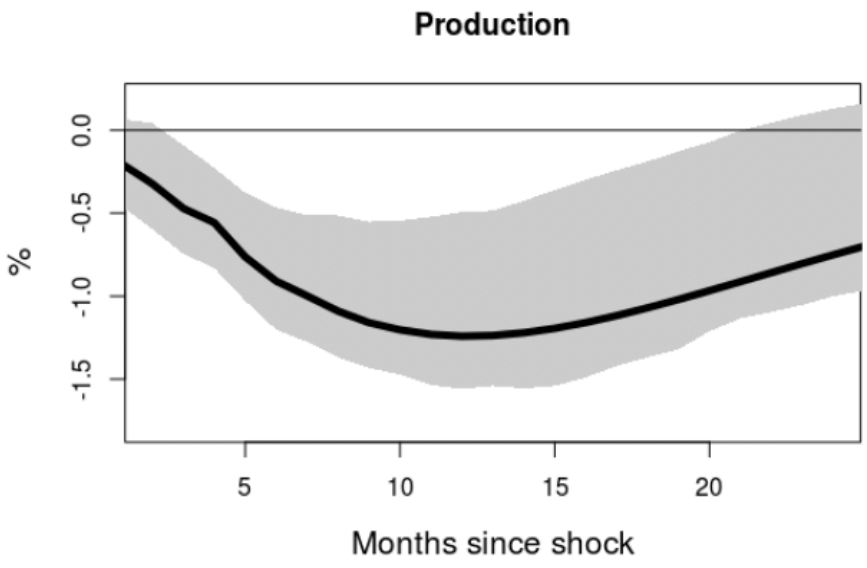
**Table 1: Emotion dictionary samples**

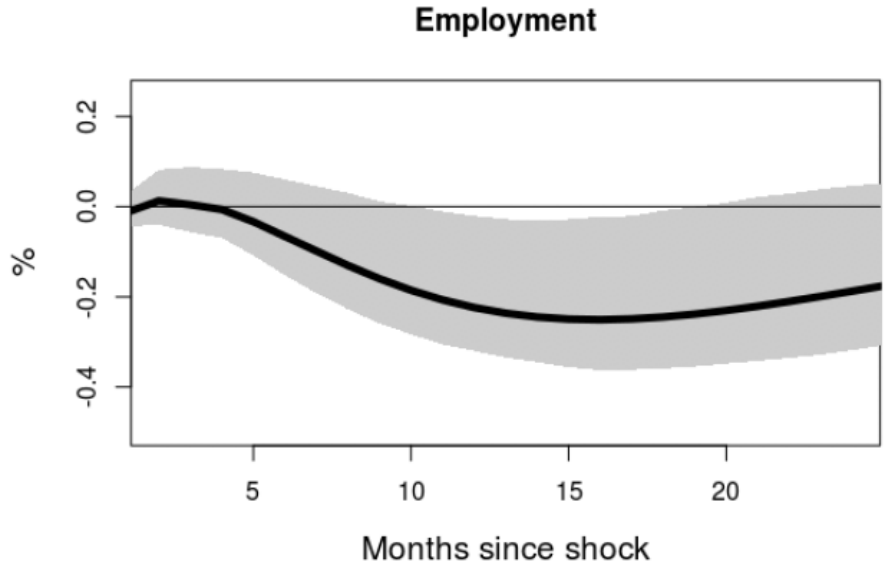
| Anxiety | | Excitement | |
|---|---|---|---|
| Jitter | Terrors | Excited | Excels |
| Threatening | Worries | Incredible | Impressively |
| Distrusted | Panics | Ideal | Encouraging |
| Jeopardized | Eroding | Attract | Impress |

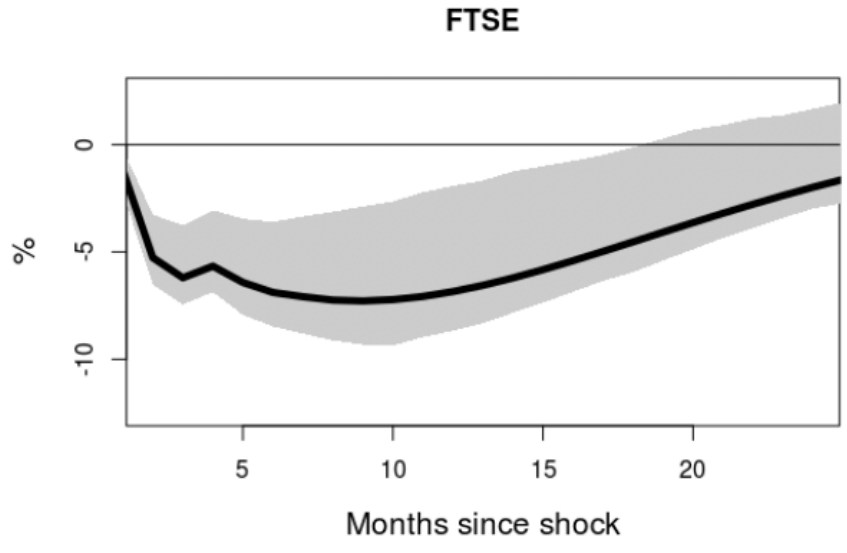
# Index


RTRS (green)
BROKER (red)
MCDAILY (black)
Dotcom High
Dotcom Low
Case-Shiller peak
BNP Paribas
Northern Rock
Bear Stearns
Lehman/AIG
First Greek Rescue
Irish Rescue
Portugese Rescue
Second Greek Rescue
Proposed Greek Debt Restructuring
'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

# Dictionary Methods + LDA

Sometimes the meaning of a dictionary can vary depending on the topic it discusses.

One can combine dictionary methods with the output of LDA to weight words counts by topic.

Recent application to minutes of the Federal Reserve to extract index of economic situation and forward guidance.

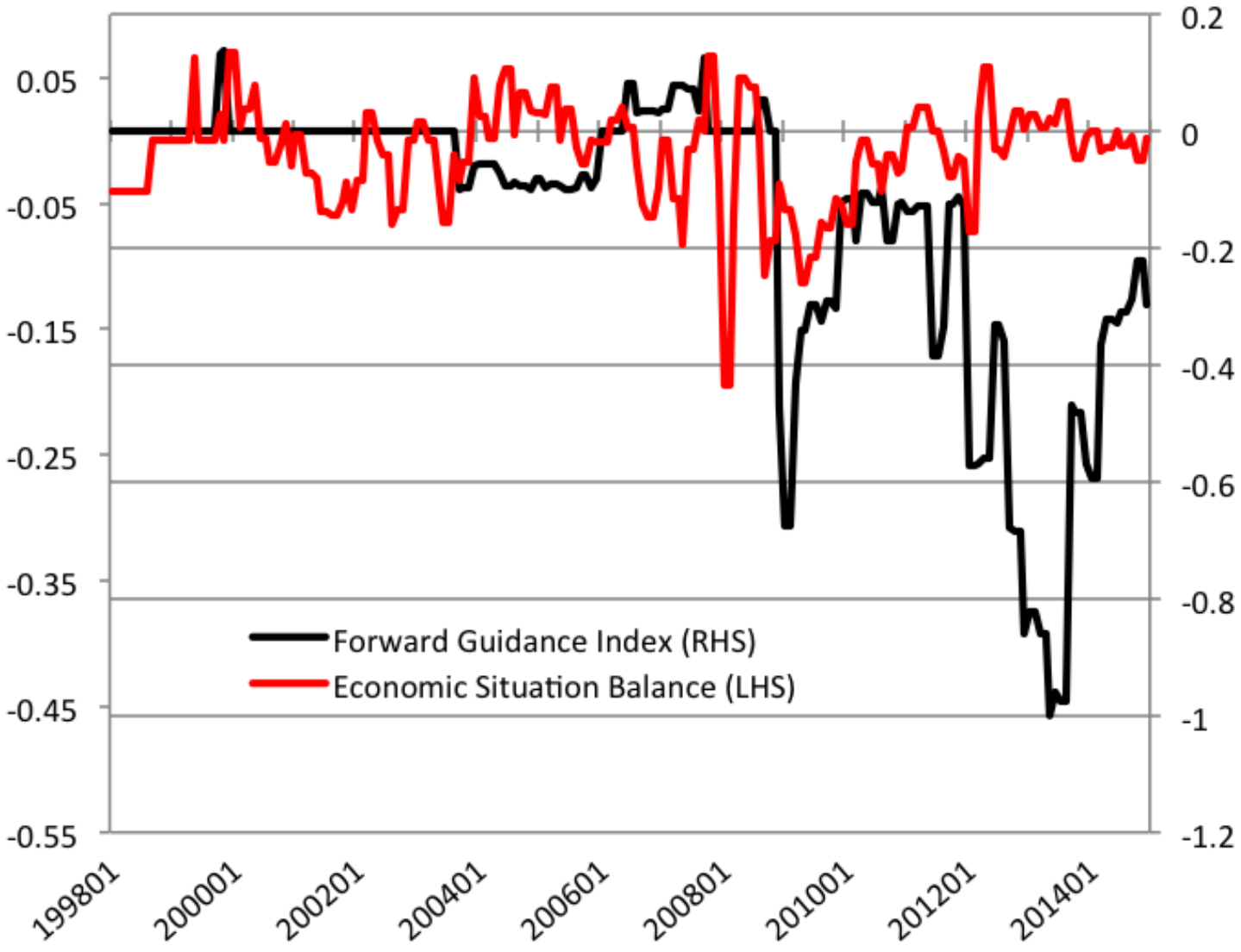
We run 15-topic model and identify two separate kinds of topic.

# Monetary Measures of Tone from Apel/Blix-Grimaldi

| **Contraction** | **Expansion** |
|---|---|
| decreas* | increas* |
| decelerat* | accelerat* |
| slow* | fast* |
| weak* | strong* |
| low* | high* |
| loss* | gain* |
| contract* | expand* |

# Example Topic

# Indices


0.05
-0.05
-0.15
-0.25
-0.35
-0.45
-0.55
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
199801
200001
200201
200401
200601
200801
201001
201201
201401
Forward Guidance Index (RHS)
Economic Situation Balance (LHS)

# Supervised Learning

One advantage of supervised learning over dictionary methods is that they are targeted directly at maximizing predictive accuracy, which is often what we care most about.

Suppose we have text features for document $d$ represented as $\mathbf{x}_d$ along with an associated sentiment variable $y_d$.

Two options for supervised machine learning:

1. Discriminative classifier that models $p(\, y_d \mid \mathbf{x}_d \,)$: e.g. LASSO, ridge regression.
2. Generative classifier that models the full joint distribution $p(\, y_d, \mathbf{x}_d \,)$: e.g. Naive Bayes, supervised LDA.

Generative classifiers have a higher asymptotic error than discriminative (Efron 1975) but can achieve their error faster (Ng and Jordan 2001).

# Feature Selection for Discriminative Classifier

One question is how to represent text: can use unigram, bigram, trigrams counts (and even more complex structures as in Shapiro et. al. 2018).

One can also apply a dimensionality-reduction algorithm to map $\mathbf{x}_d$ into a $K$-dimensional latent space, and use these as the features.

This technique is related to principal components regression, and is particularly appropriate when terms are highly correlated.

Can also use non-labeled texts along with labeled texts in topic modeling, since LDA uses no information from labels in estimation of topic shares.

Blei et. al. (2003) show that topic share representation is competitive with raw counts in classification.

# Example

In recent work with Michael McMahon and Matthew Tong, we study the impact of the release of the Bank of England's Inflation Report on bond price changes at different maturities.

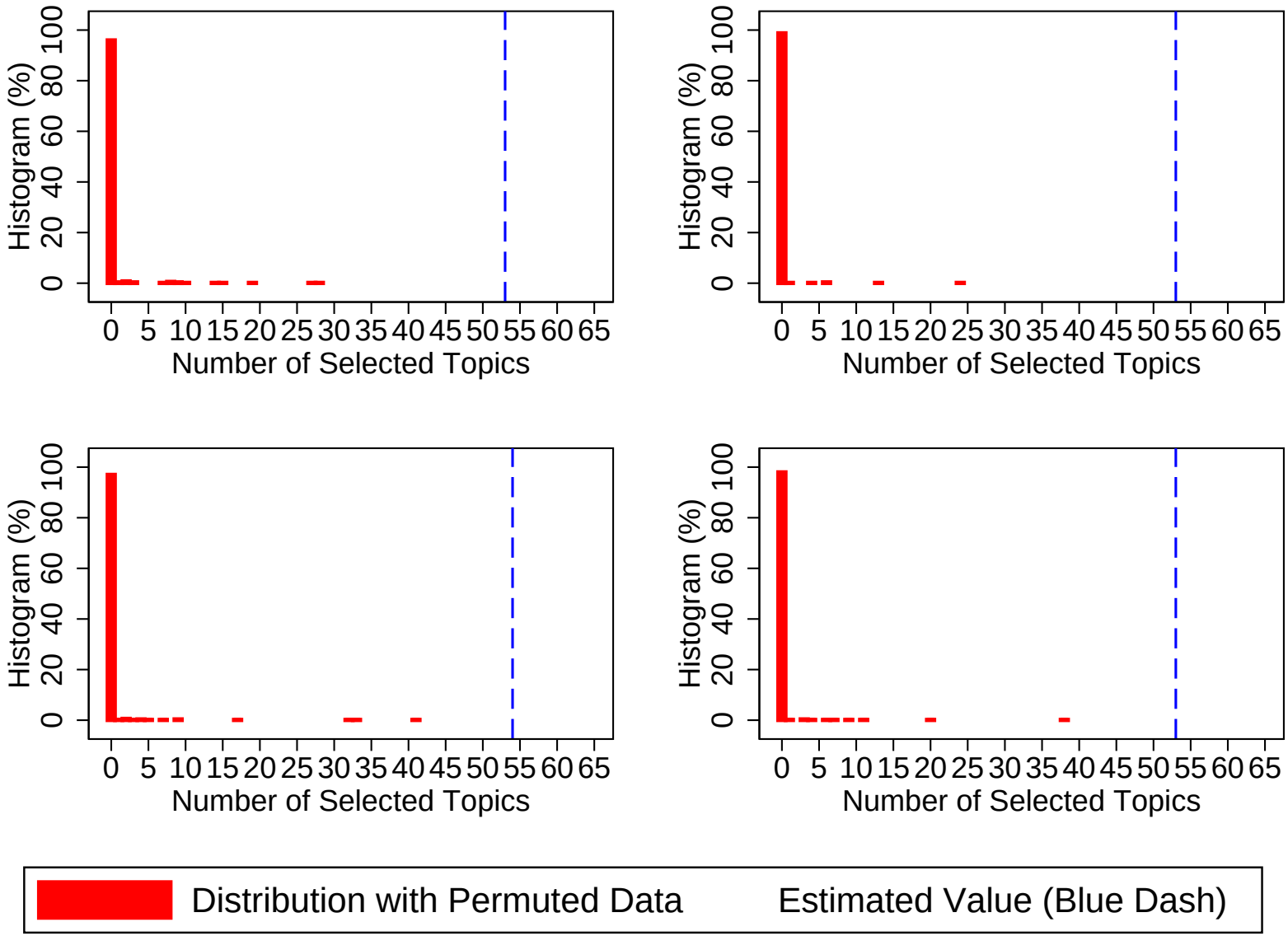
IR contains forecast variables we use as controls: (i) mode, variance, and skewness of inflation and GDP forecasts; (ii) their difference from the previous forecast.

To represent text, we estimate a 30-topic model and represent each IR in terms of (i) topic shares and (ii) evolution of topic shares from previous IR.

First step in the analysis is to partial out the forecast variables from bond price moves and topic shares by constructing residuals.

We are then left with 69 bond price moves (number of IRs in the data) and 60 text features.

# LASSO-based Test of Information Content of Narrative



# Which Information Matters?

LASSO selects dozens of features at all maturities: standard over-selection problem (Meinshausen and Bühlmann 2006, Annals).

How to identify key topics?

We apply a non-parametric bootstrap to simulate the "inclusion probabilities" of topic features at different maturities.

Draw with replacement from our 69 observations to obtain new sample, perform LASSO, and record whether each feature is included.

Repeat 500 times, and rank topics according to the fraction of bootstrap draws in which they appear.

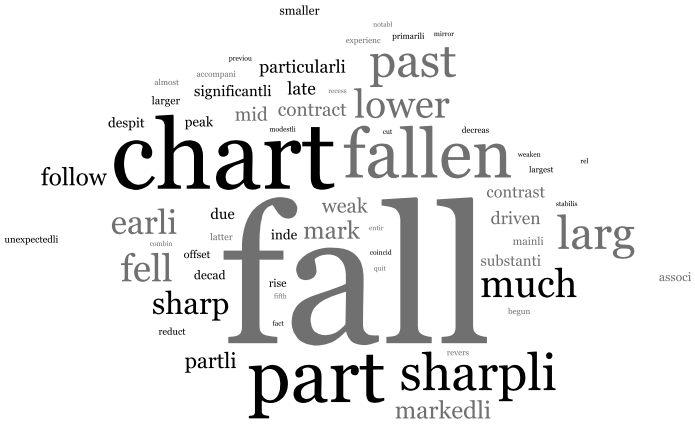
# Results: Top Topic

Top Topics for Different Yields (L=Level; D=Change)

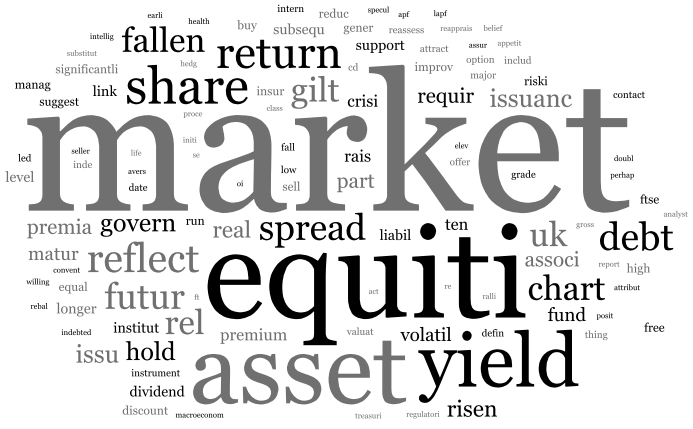
| $\lvert\Delta i_{0:12,t}\rvert$ | | $\lvert\Delta f_{36,t}\rvert$ | | $\lvert\Delta f_{60,t}\rvert$ | | $\lvert\Delta f_{60:120,t}\rvert$ | |
|---|---|---|---|---|---|---|---|
| Var | Selection % | Var | Selection % | Var | Selection % | Var | Selection % |
| L25 | 0.958 | D24 | 0.858 | L28 | 0.876 | D17 | 0.91 |
| D24 | 0.954 | D25 | 0.844 | D17 | 0.784 | D18 | 0.896 |
| L5 | 0.932 | L28 | 0.826 | D18 | 0.772 | L20 | 0.836 |
| L26 | 0.91 | D14 | 0.76 | L20 | 0.722 | D13 | 0.808 |

# Results: Top Topics

## 1-Year Spot Rate




(a) L25



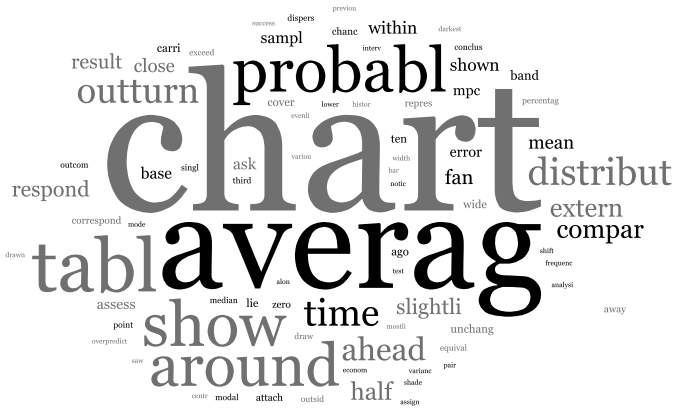

(b) D24


(c) L5


(d) L26

# Results: Top Topics

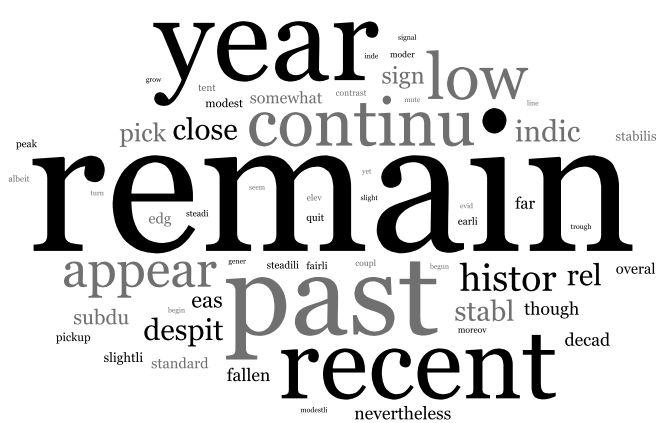
## 5-Year, 5-Year Forward Rate




(a) D17


(b) D18


(c) L20




(d) D13

# Monetary Shocks & Fed Statements

We examine the relationship between Fed statements and the direction of the Romer and Romer (2004) shocks.

To do so, we compute all unique two- and three-word phrases in Fed statements (bigrams/trigrams), and count their frequency in each documents.

Let $x_v^-$ ($x_v^+$) be the count of term $v$ among statements associate with negative (positive) shocks.

We rank terms according to their informativeness by $\log(x_v^-) - \log(x_v^+)$, and select the top 1,000.

# Most Informative Terms—Statements

| **negative shock** | **positive shock** |
|---|---|
| lower.target.feder | rais.target.feder |
| lower.target | rais.target |
| committe.continu | increas.discount.rate |
| continu.believ | increas.discount |
| committe.continu.believ | rise.energi |
| basi.point.reduct | point.increas.discount |
| point.reduct | point.increas |
| committe.decid | basi.point.increas |
| today.lower.target | action.stanc.monetari |
| today.lower | growth.price |

# Naive Bayes for Text

We can represent each statement as a length-1000 vector $\mathbf{x}_t$, where $x_{t,v}$ is the count of term $v$ in the time $t$ statement.

Let $RR_t \in \{-, +\}$ represent the direction of the shock in period $t$.

Suppose that term $v$ appears with probability $\beta_v^y$ when the shock is $y$, and that shock $y$ occurs with probability $\rho_y$. The log-likelihood of observing the data is then

$$\sum_t \mathbb{1}(RR_t = y) \log(\rho_y) + \sum_t \sum_v \mathbb{1}(RR_t = y) x_{t,v} \log\left(\beta_v^y\right).$$

Maximum likelihood estimation gives

$$\widehat{\rho}_y = \frac{N_y}{N} \text{ and } \widehat{\beta}_v^y = \frac{x_v^y}{\sum_v x_v^y}.$$

# Classification

One can use the MLE estimates to associate out-of-sample document $\mathbf{x}_d$ with label $y_d$. By Bayes' Rule we have

$$\Pr[\, y_d = y \mid \mathbf{x}_d \,] \propto \Pr[\, \mathbf{x}_d \mid y_d = y \,] \Pr[\, y_d = y \,]$$

and we can select

$$y_d = \arg\max_y \; \log(\widehat{\rho}_y) + \sum_v x_{d,v} \log\left(\widehat{\beta}_v^y\right).$$

# Classification

One can use the MLE estimates to associate out-of-sample document $\mathbf{x}_d$ with label $y_d$. By Bayes' Rule we have

$$\Pr[\, y_d = y \mid \mathbf{x}_d \,] \propto \Pr[\, \mathbf{x}_d \mid y_d = y \,] \Pr[\, y_d = y \,]$$

and we can select

$$y_d = \arg\max_y \; \log(\widehat{\rho}_y) + \sum_v x_{d,v} \log\left(\widehat{\beta}_v^y\right).$$

To evaluate the quality of the classification, a standard exercise is to:

1. Draw some fraction (one half in our case) of the data, and estimate parameters on it.
2. Use the estimates to classify the held-out documents.
3. Compare the predicted and actual labels.

We perform this exercise using 1,000 random draws for the training set.

# Classification Results—Statements

| | predicted | |
|---|---|---|
| actual | 0 | 1 |
| 0 | 17.706 | 2.658 |
| 1 | 6.911 | 13.725 |

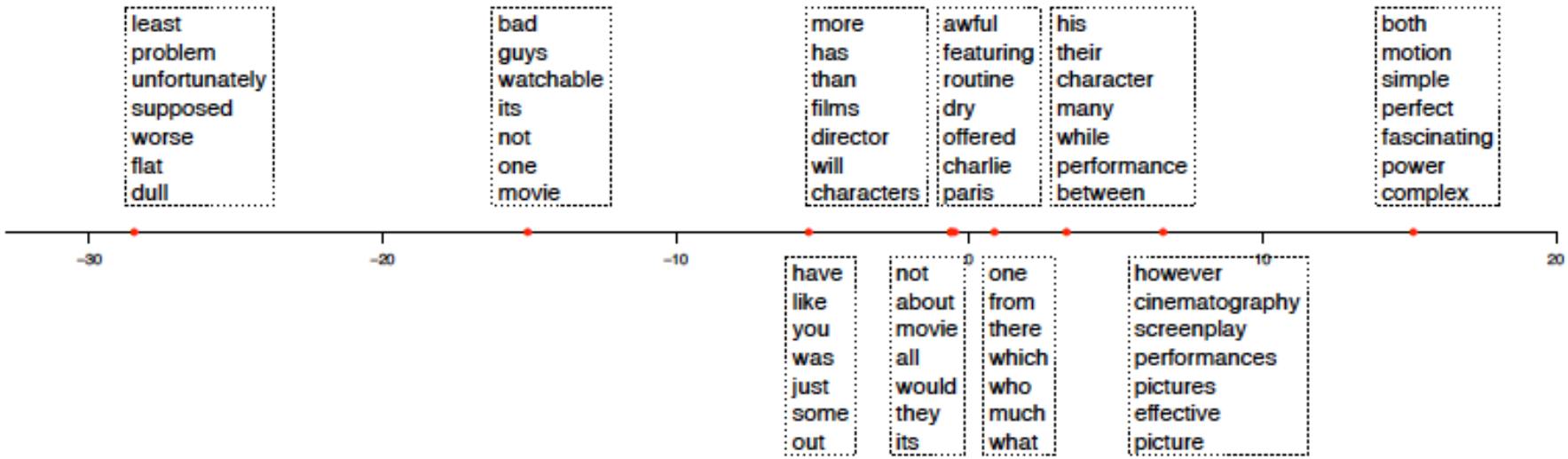
76% average classification accuracy, but asymmetry across shock values.

# Supervised LDA (Blei and McAuliffe)

1. Draw $\boldsymbol{\theta}_d$ independently for $d = 1, \ldots, D$ from Dirichlet($\boldsymbol{\alpha}$).
2. Each word $w_{d,n}$ in document $d$ is generated from a two-step process:
   - 2.1 Draw topic assignment $z_{d,n}$ from $\boldsymbol{\theta}_d$.
   - 2.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{z_{d,n}}$.
3. Draw $y_d$ from $\mathcal{N}(\boldsymbol{\phi}^T \bar{\mathbf{z}}_d, \sigma^2)$ where $\bar{\mathbf{z}}_d = (n_{d,1}/N_d, \ldots, n_{d,K}/N_d)$ and $z_{d,k}$ is the number of allocations to topic $k$ in document $d$.

Essentially plain LDA with a linear regression linking topic allocations with observed variables.

# Example of Supervised LDA with Movie Review Data



# Conclusion

Sentiment analysis is an example of a broader problem with big data: associate a label to a document based on its content.

Dictionaries allow the researcher to control the content that guides the classification, but supervised learning should generally perform better for classification accuracy.

When there are relatively few documents, generative models can perform very well even if they are more complex to estimate.